



Curso de  
**Fundamentos de  
Procesamiento de  
Lenguaje Natural  
con Python y NLTK**

Francisco Camacho  
@el\_pachocamacho



---

# [C1] Introducción al NLP: Perspectivas y estado del arte

NLP como el camino hacia el ideal  
de IA

¿Qué significa NLP?  
**Natural Language  
Processing**

---

# ¿Qué significa NLU? Natural Language Understanding

---

“

... Si un humano no puede distinguir entre una máquina y otra persona en **una conversación**, entonces esa máquina ha alcanzado un nivel de inteligencia comparable al de un humano ...

”

*Test de Turing*

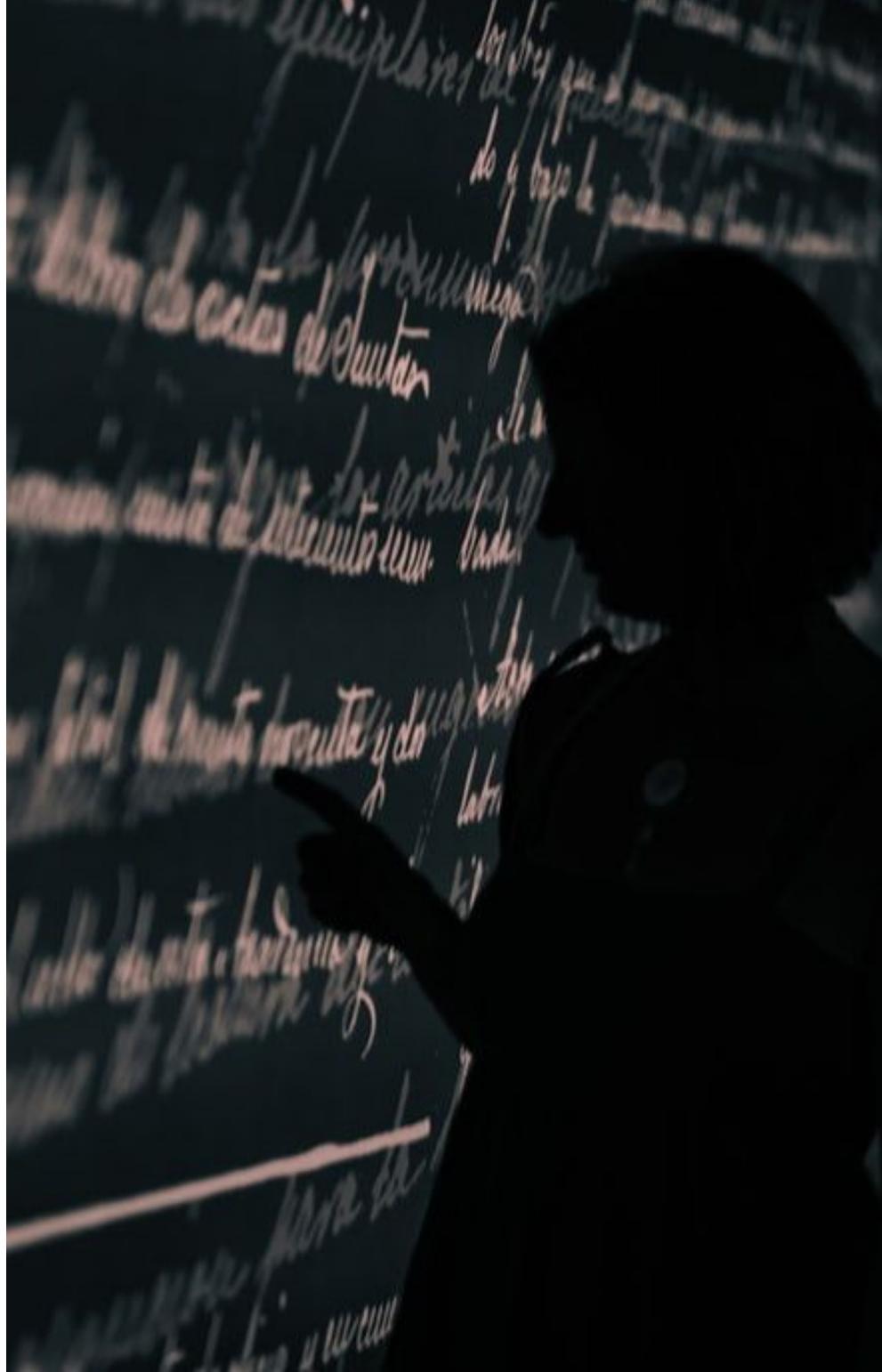
# En la Cultura Sci-Fi

Test de Voight-Kampff (Blade Runner, 1982)



# Usos actuales del NLP

- Máquinas de Búsqueda
- Traducción de texto
- Chatbots
- Análisis de discurso
- Reconocimiento del habla
- Etc ...



# ¿Por qué es tan difícil ?

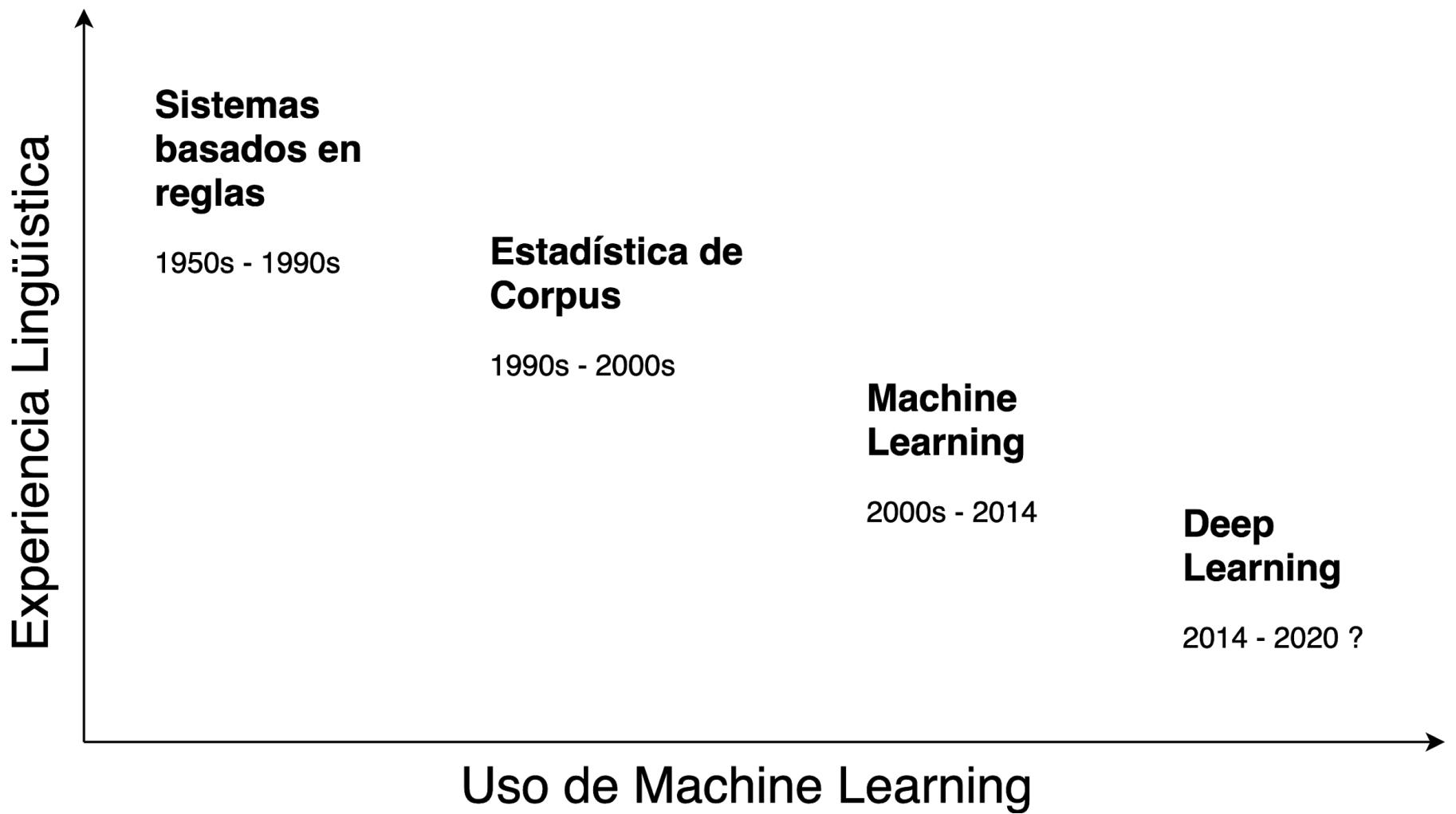
BUSCO GENTE  
PARA TRABAJAR  
ENTRE 18 Y 30 AÑOS

# ¿Por qué es tan difícil ?

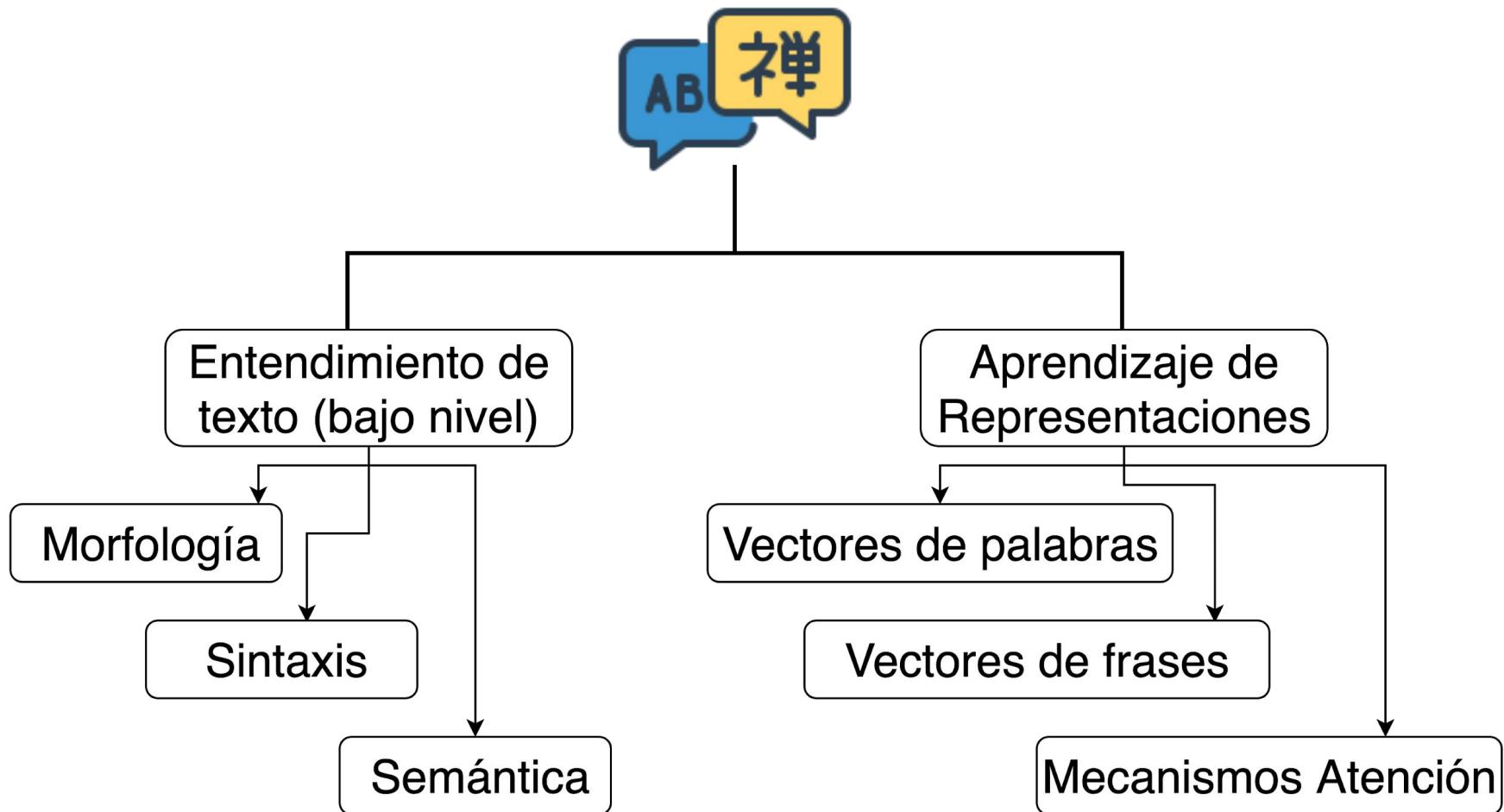


El lenguaje humano es **difuso, ambiguo** y requiere mucho **contexto**

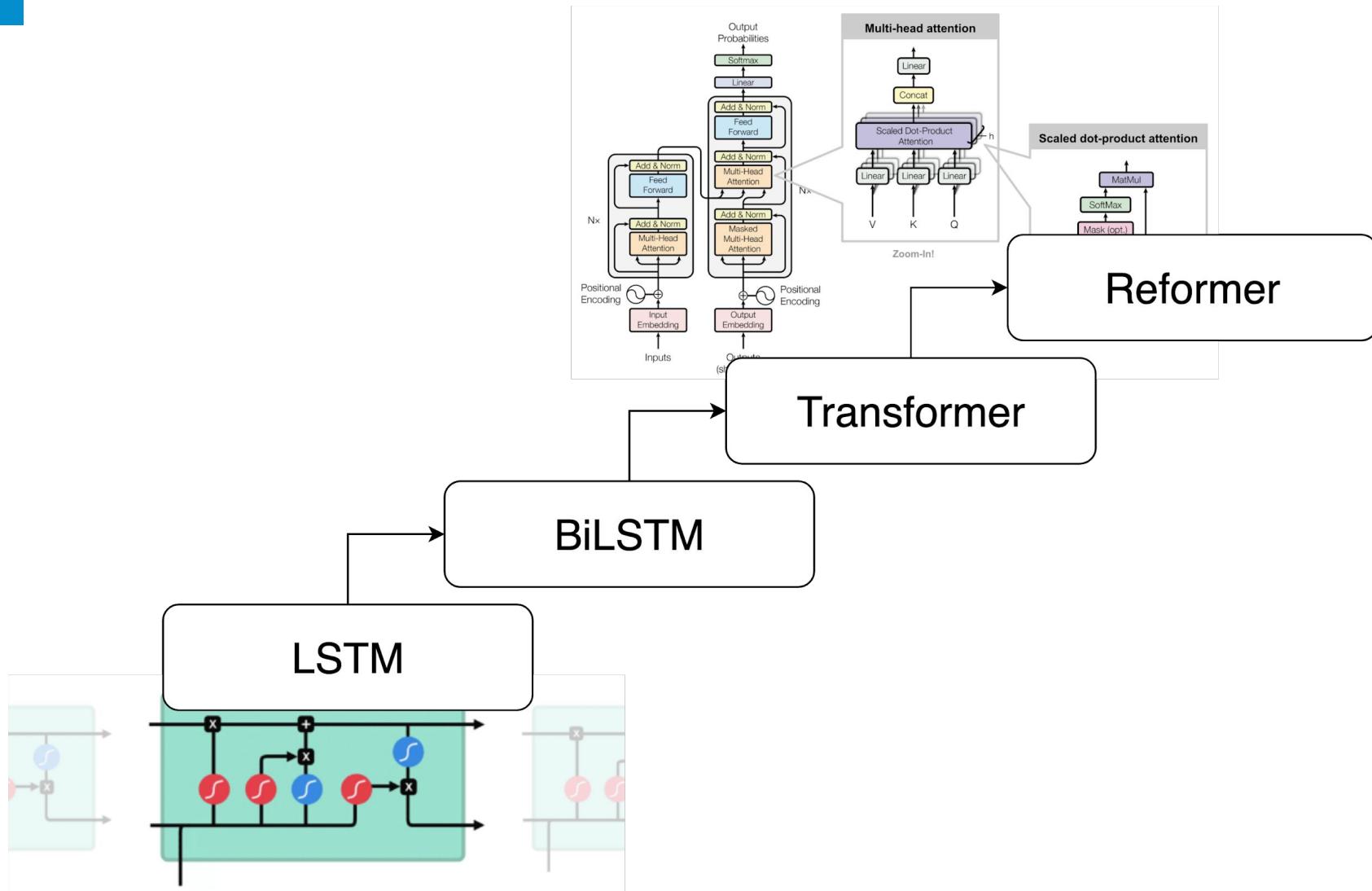
# Evolución del NLP



# Avances del NLP



# Avances del NLP



# Pero ... no tan rápido

Artificial Intelligence / Machine Learning

**AI still doesn't have the common sense to understand human language**

Natural-language processing has taken great strides recently—but how much does AI really understand of what it reads? Less than we thought.

MIT  
Technology  
Review

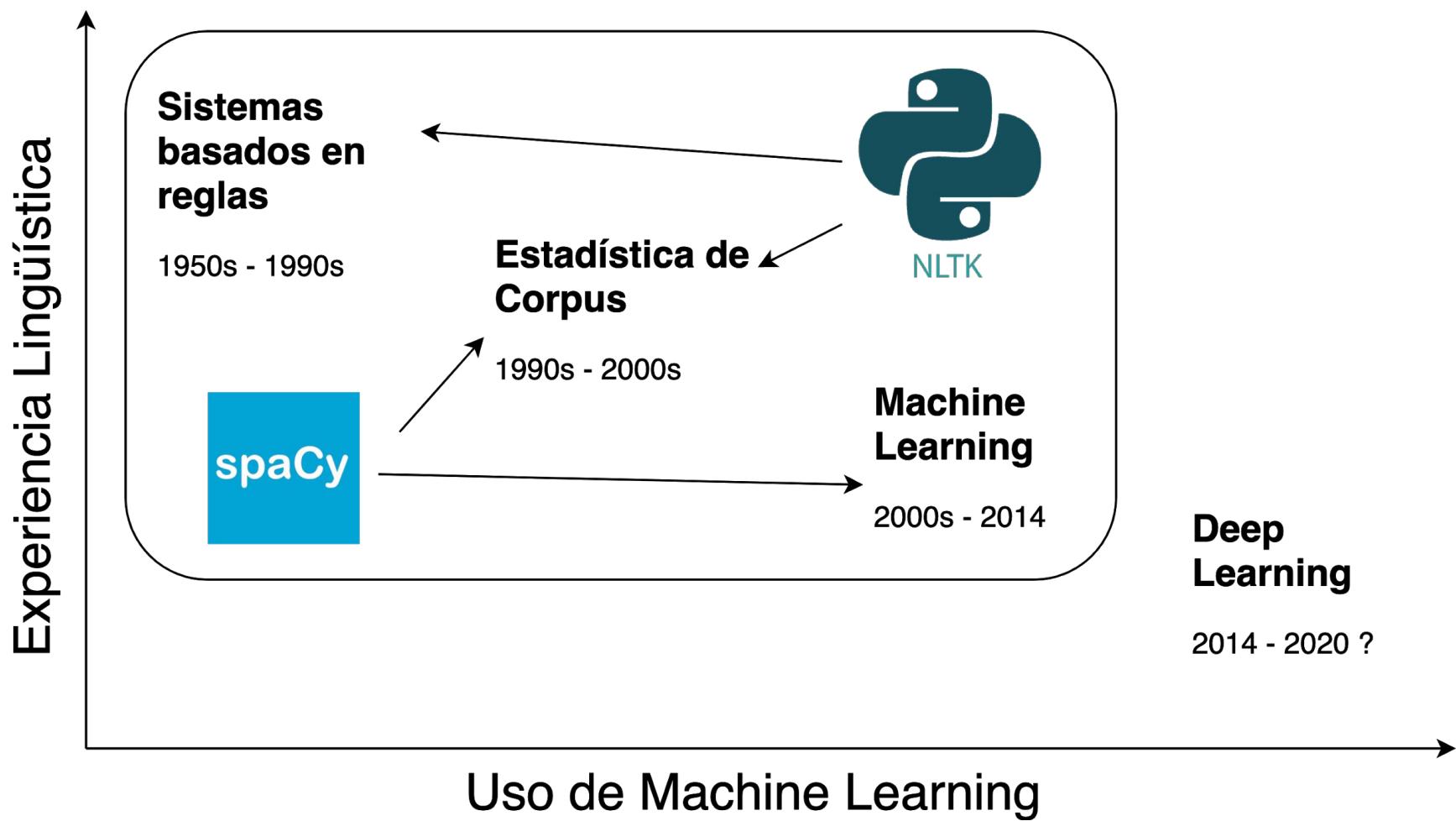
El trofeo no cabe en la caja  
porque es muy grande

El trofeo no cabe en la caja  
porque es muy pequeña

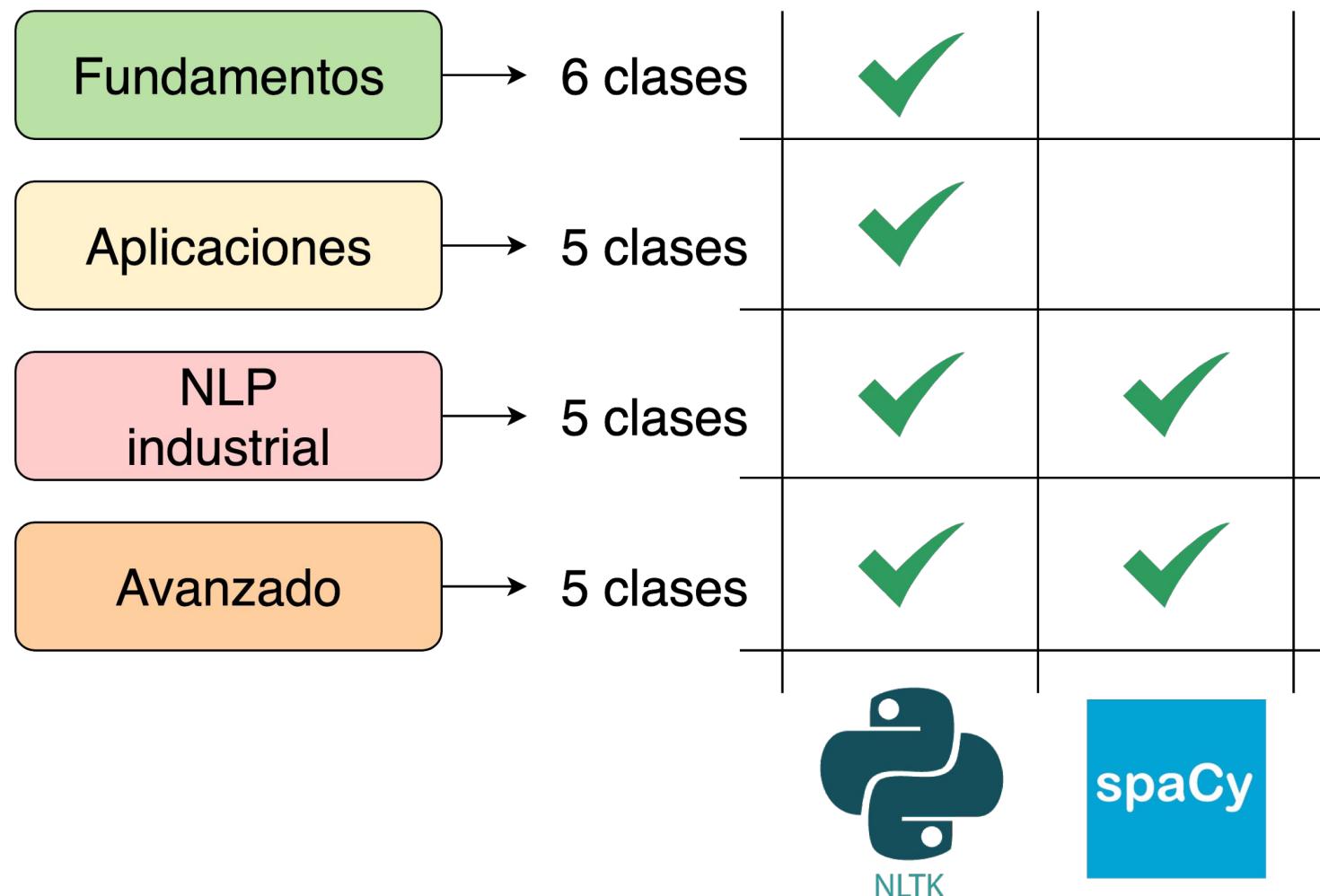
---

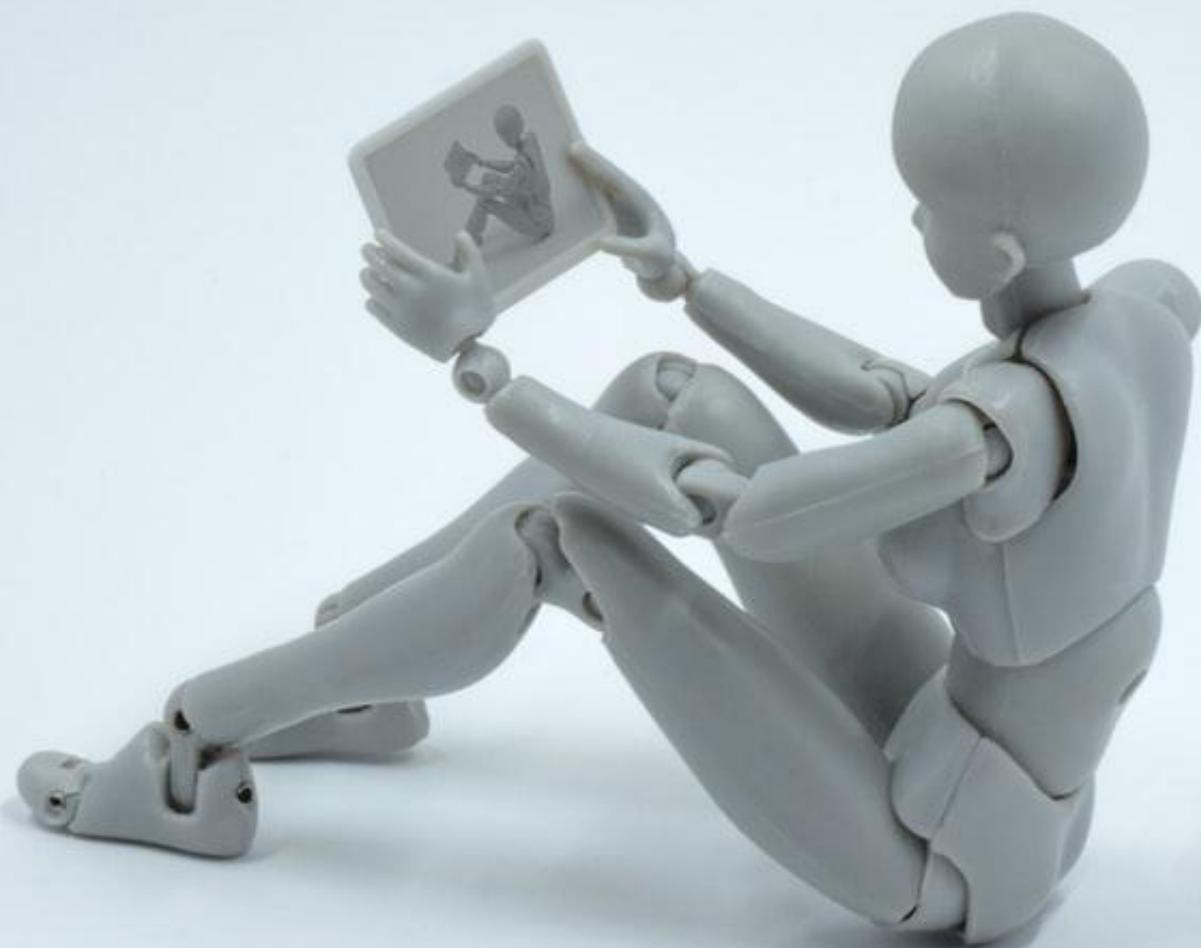
ALDE

# ¿Qué vamos a estudiar?



# Roadmap del contenido





---

# [C2] Introducción al NLP: conceptos básicos

Estructuras básicas del lenguaje  
humano

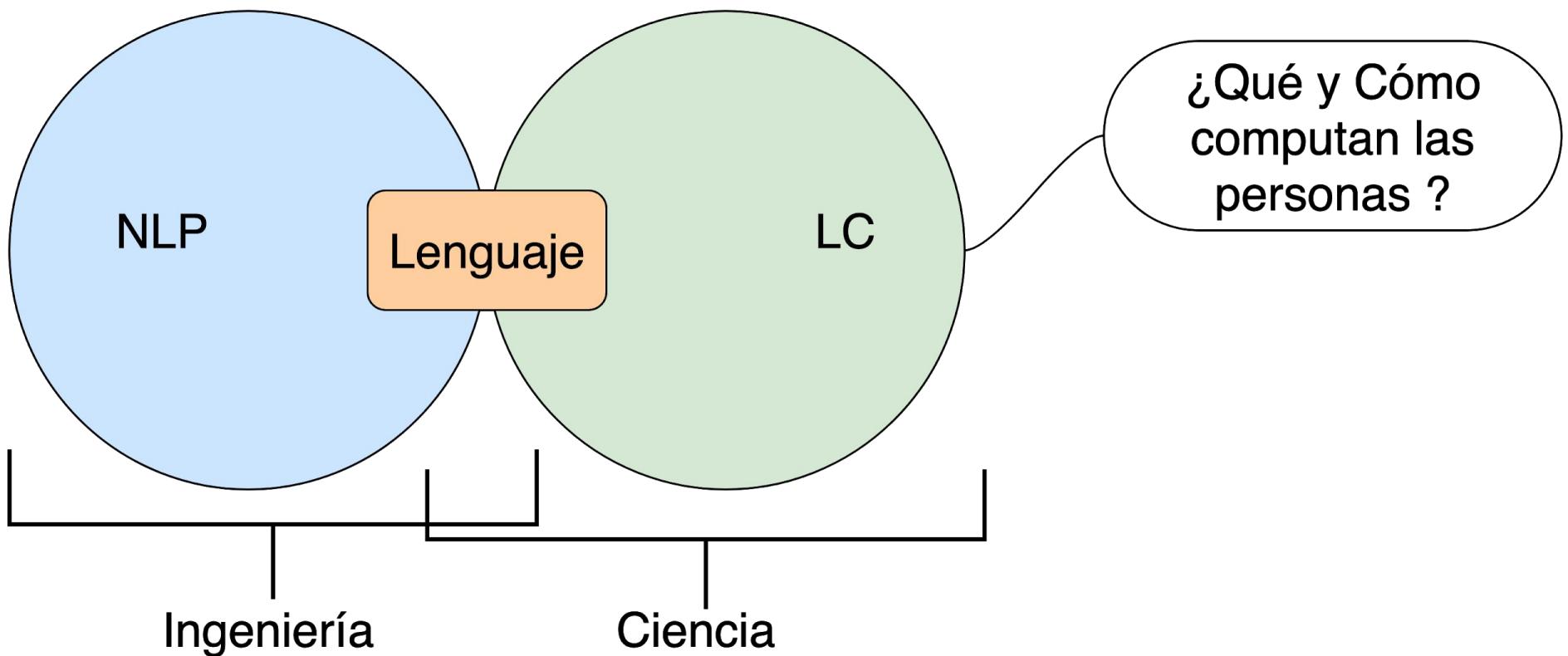
“

**Entender y caracterizar las  
reglas que determinan cómo  
estructurar expresiones  
lingüísticas ...**

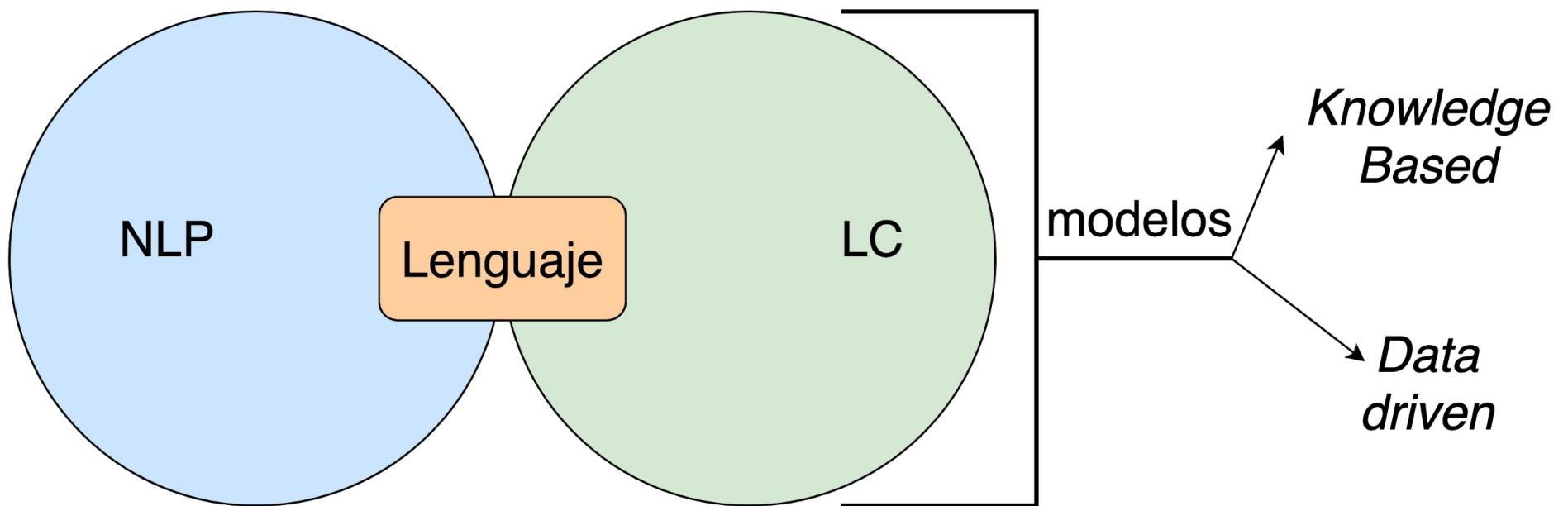
”

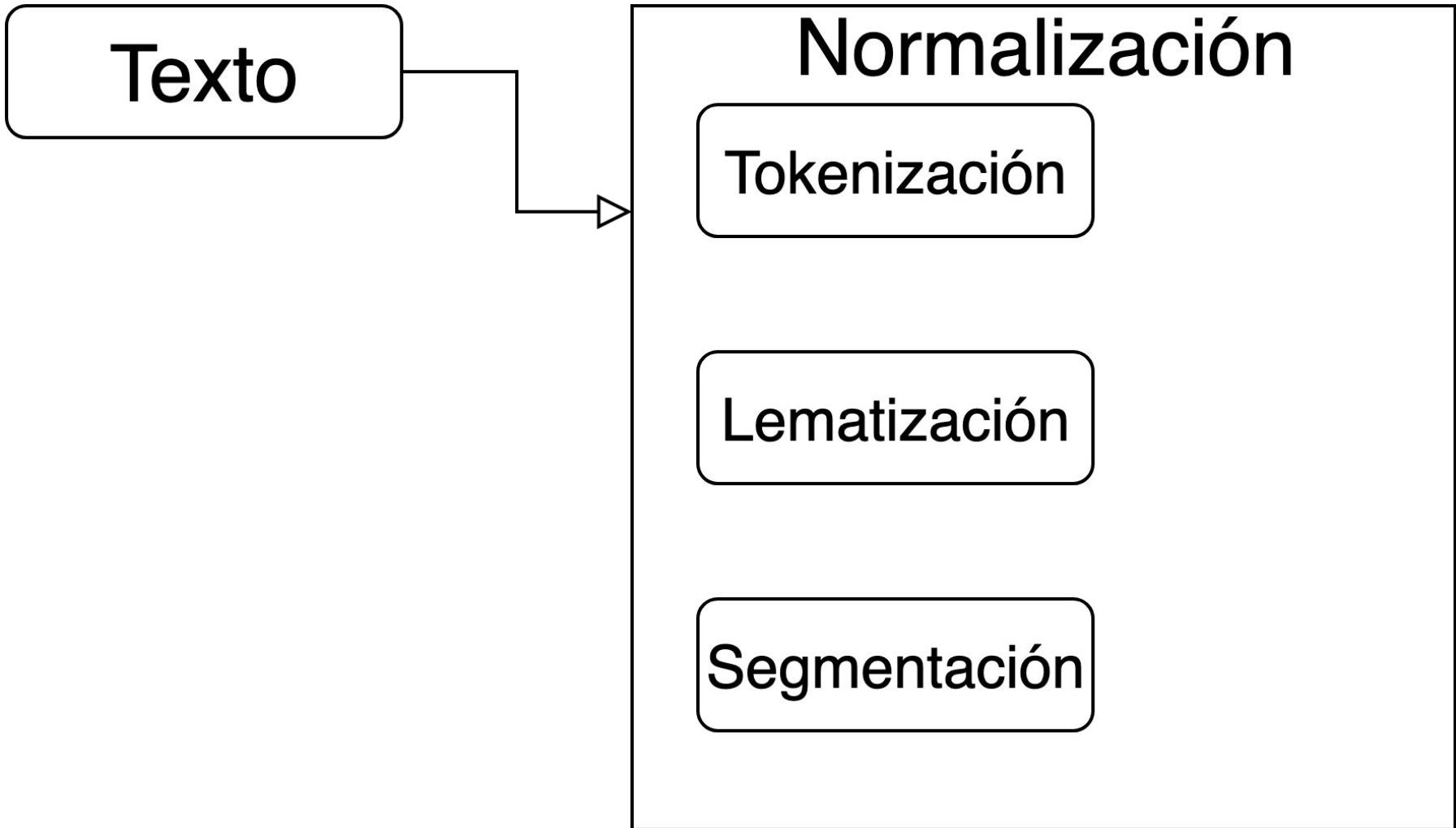
*Manning & Schütze (1999), Foundations of Statistical  
Natural Language Processing*

# Lingüística Computacional (LC)



# Lingüística Computacional (LC)





Texto

## Normalización

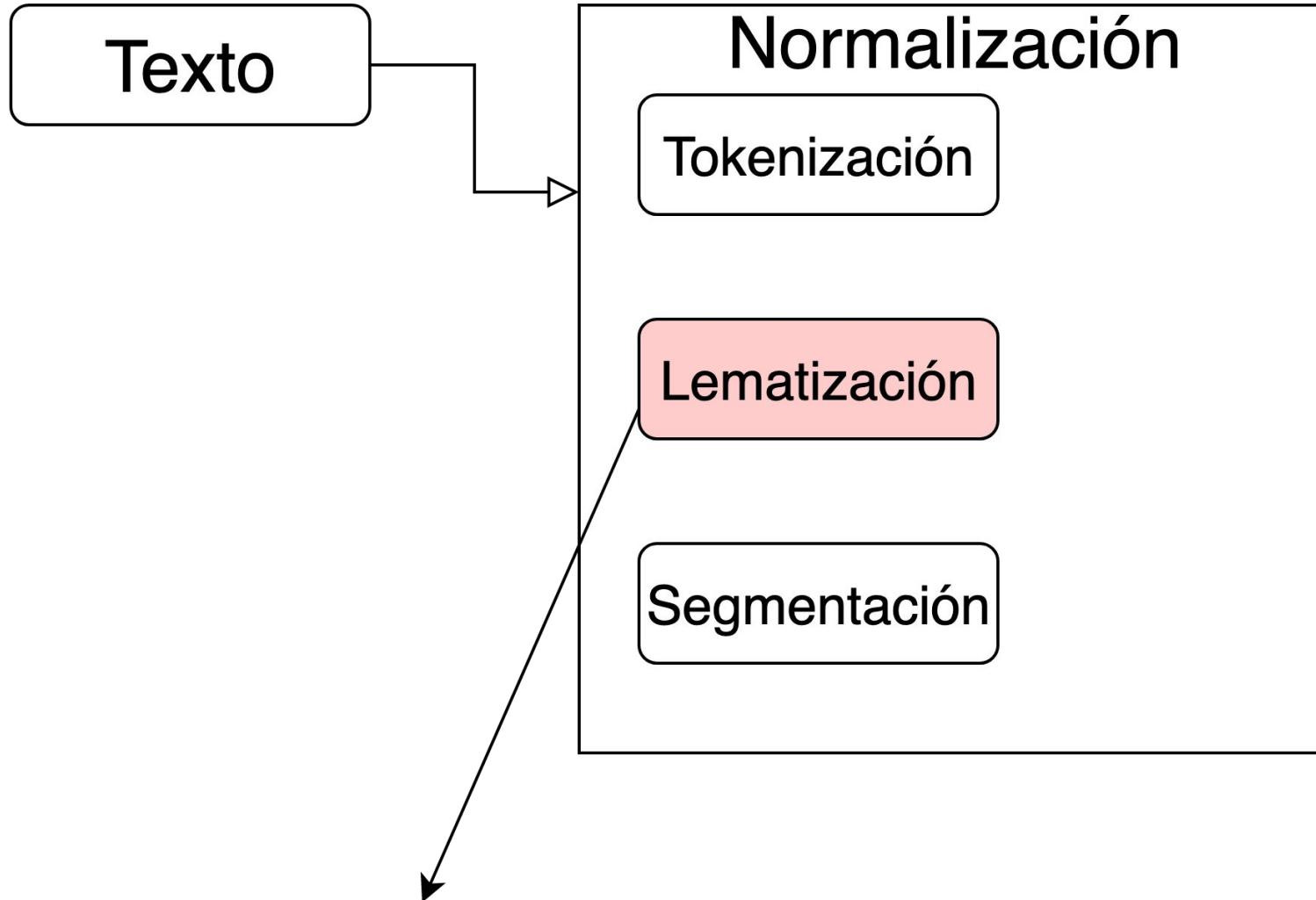
Tokenización

Lematización

Segmentación

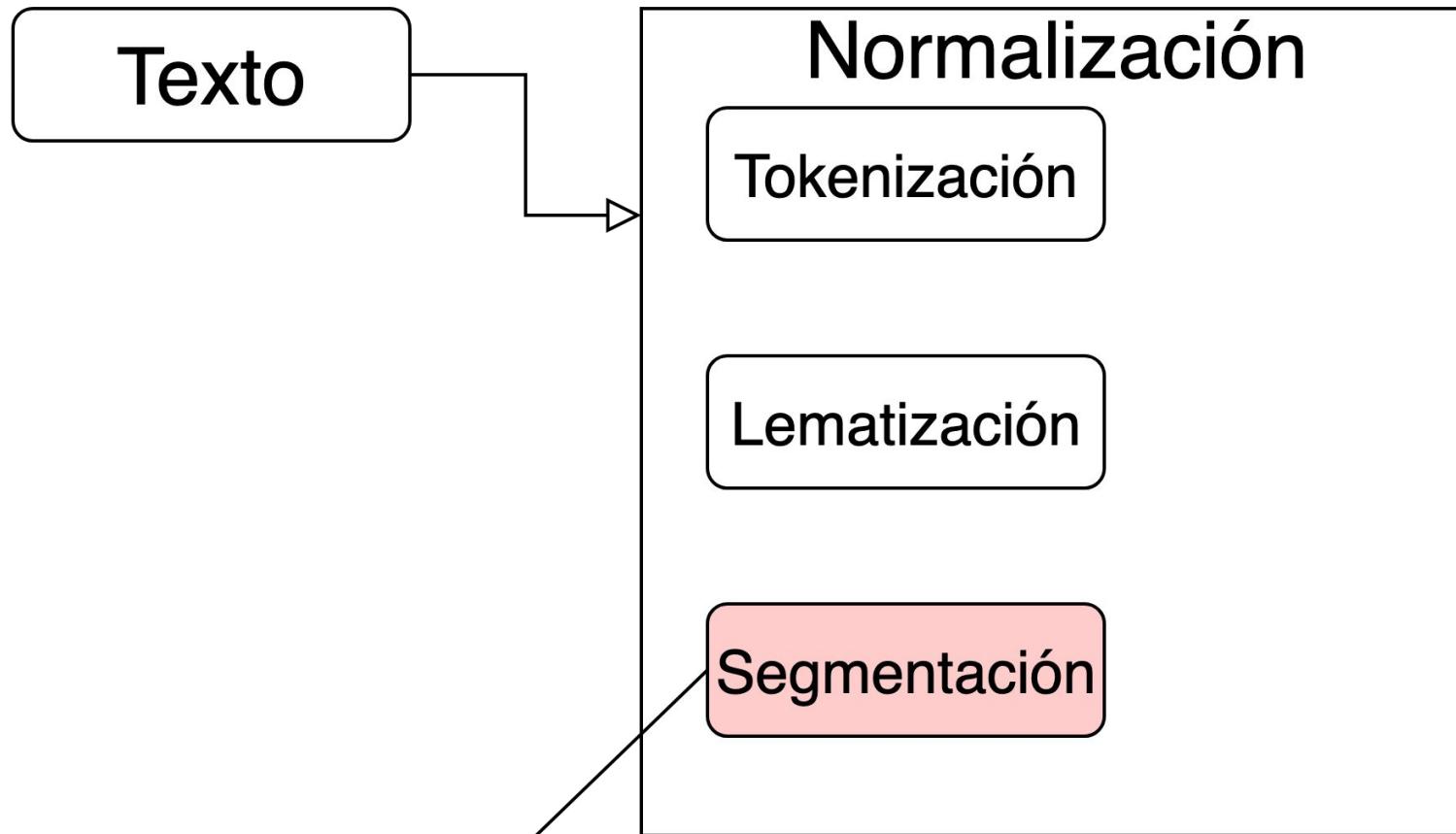
Mi hermano dejó de comer

Mi hermano dejó de comer



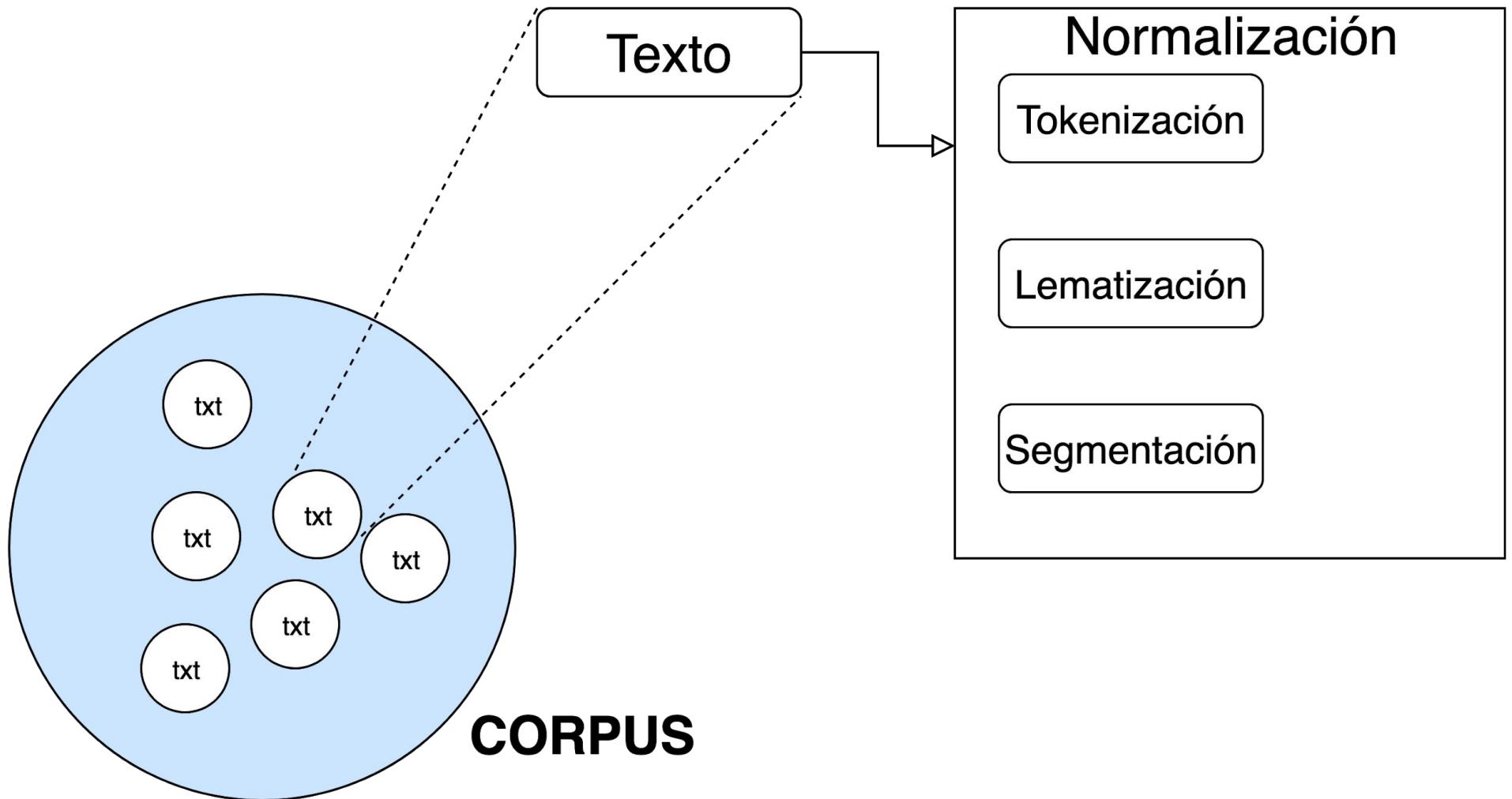
Mi hermano **dejó** de comer

Mi	hermano	<b>dejar</b>	de	comer
----	---------	--------------	----	-------

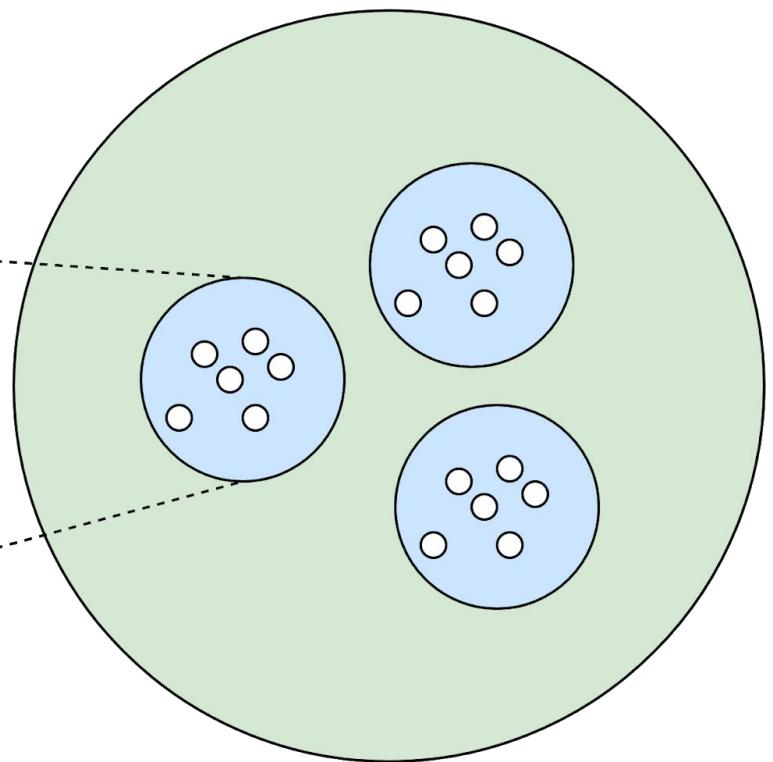
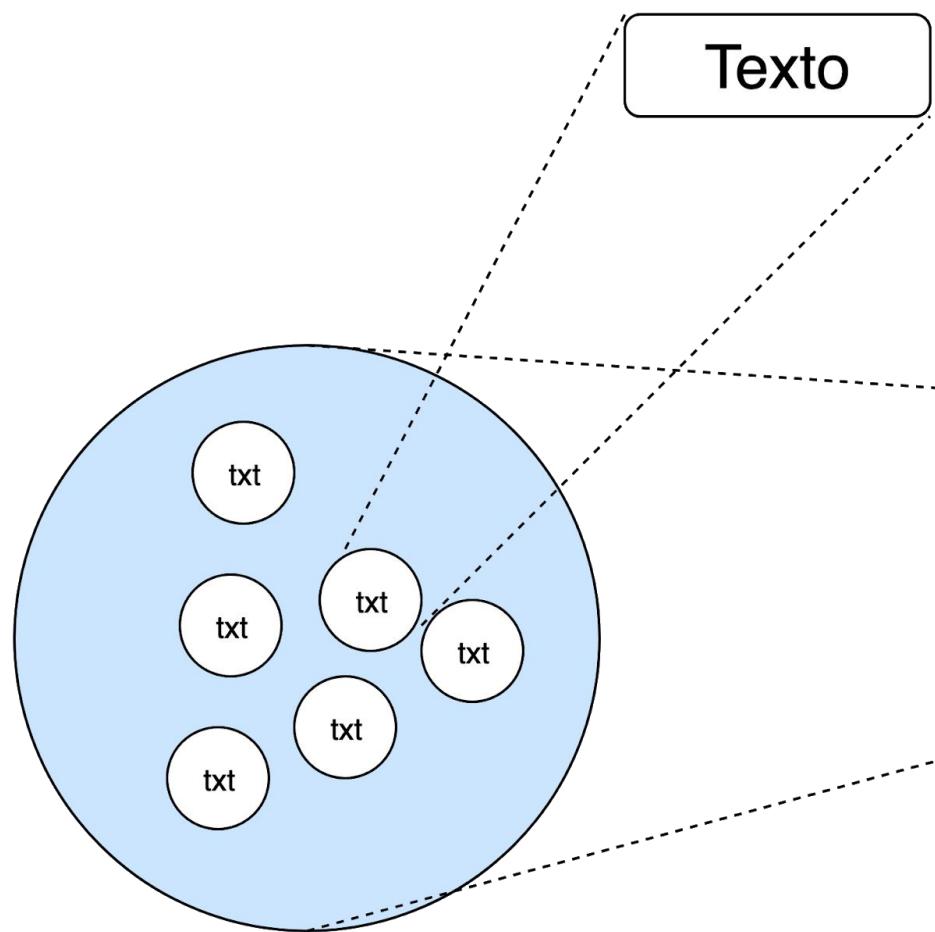


Mi hermano dejó de comer,  
no se sentía muy bien.

Mi hermano dejó de comer,  
no se sentía muy bien.



**Texto**



---

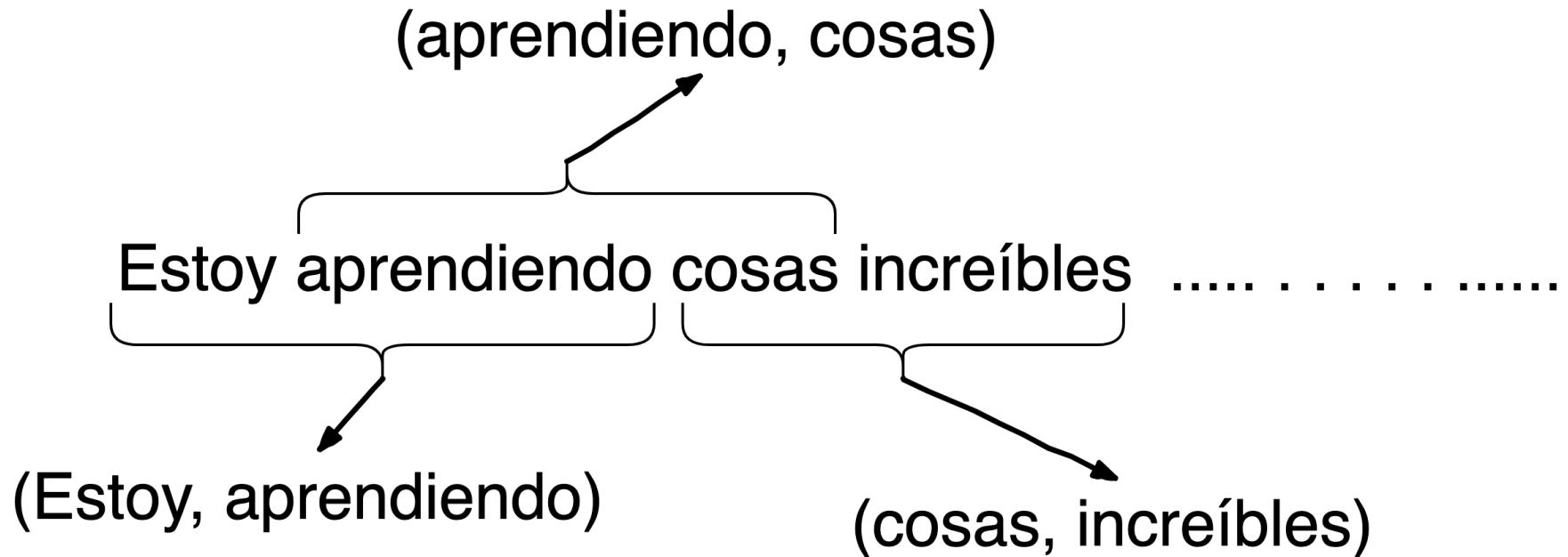
# [C5] Fundamentos con NLTK: estadística del lenguaje

Frecuencias de Ocurrencia para  
palabras, n-gramas y colocaciones

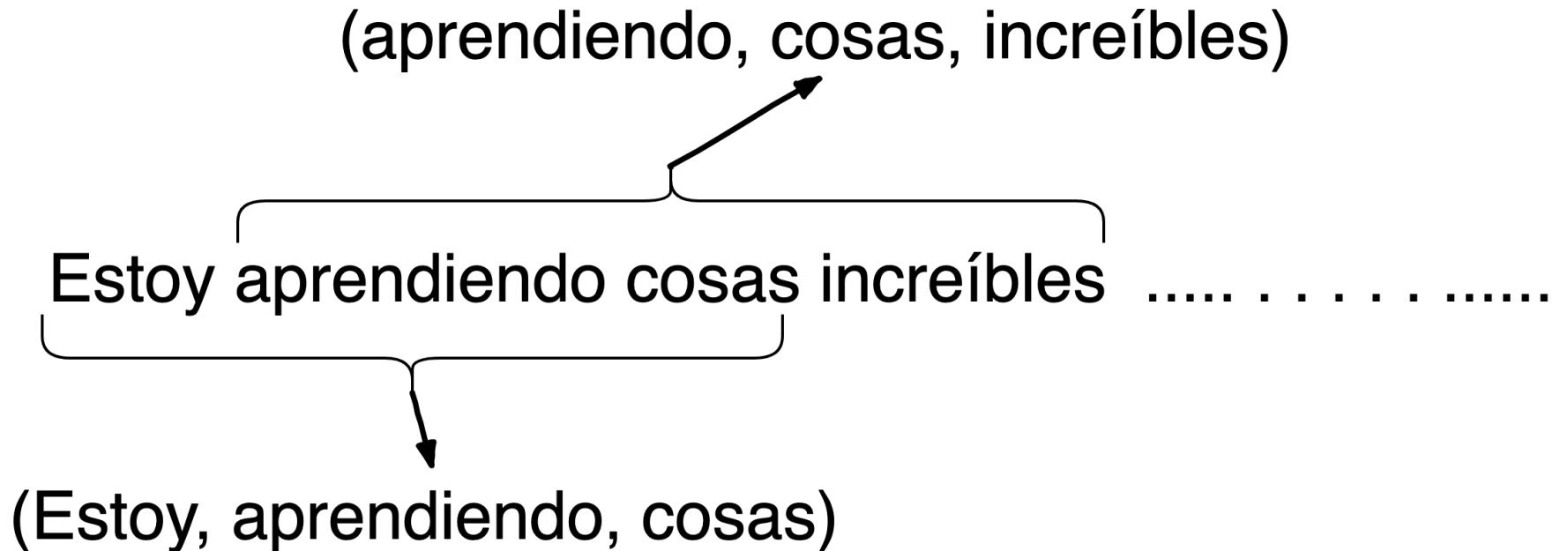
¿Qué es un N-grama?  
Es una secuencia de n  
palabras

---

# Bi-gramas



# Tri-gramas



“

**Las colocaciones de una palabra  
son sentencias que indican los  
lugares que acostumbra a tomar  
esa palabra en el lenguaje (sin  
seguir las reglas del lenguaje) ....**

”

*Firth (1957), Modes in Meaning - Paper in Linguistics*

# Colocaciones

Le dieron ganas de dormir

Le introdujeron ganas de dormir

ventilar secretos !

“

Vamos a un notebook en  
Google Colab ...



”

---

# [C6] Fundamentos con NLTK: recursos léxicos

¿Qué son los recursos léxicos y  
cómo podemos usarlos?

**¿Qué es un recurso léxico?**  
Colecciones de palabras o  
frases con meta-datos

# ¿Cómo es?

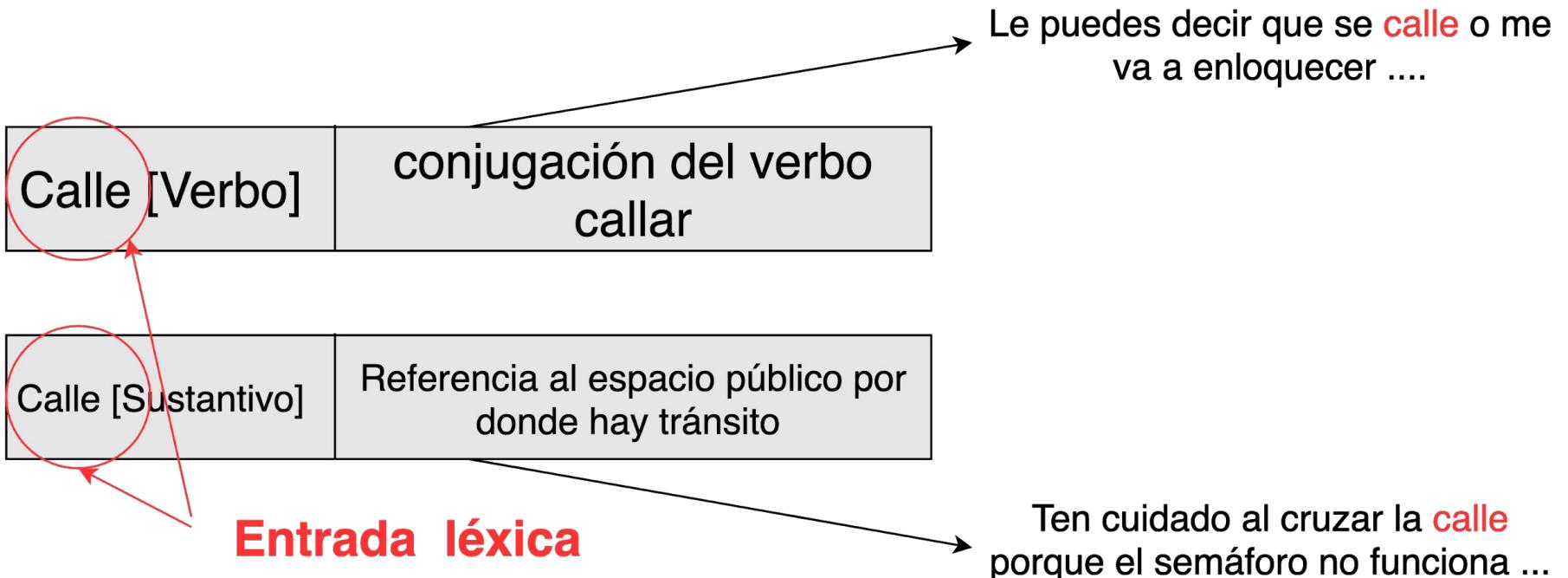
Calle [Verbo]	conjugación del verbo callar
---------------	---------------------------------

Calle [Sustantivo]	Referencia al espacio público por donde hay tránsito
--------------------	---

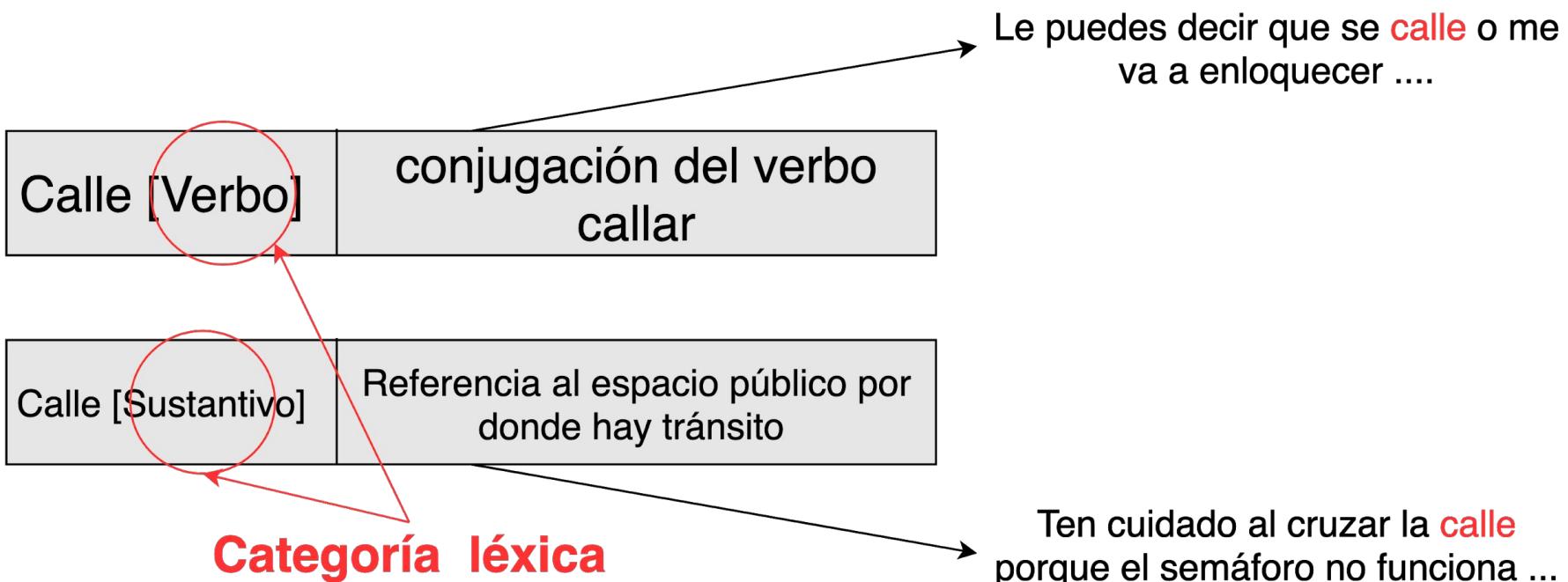
Le puedes decir que se **calle** o me va a enloquecer ....

Ten cuidado al cruzar la **calle** porque el semáforo no funciona ...

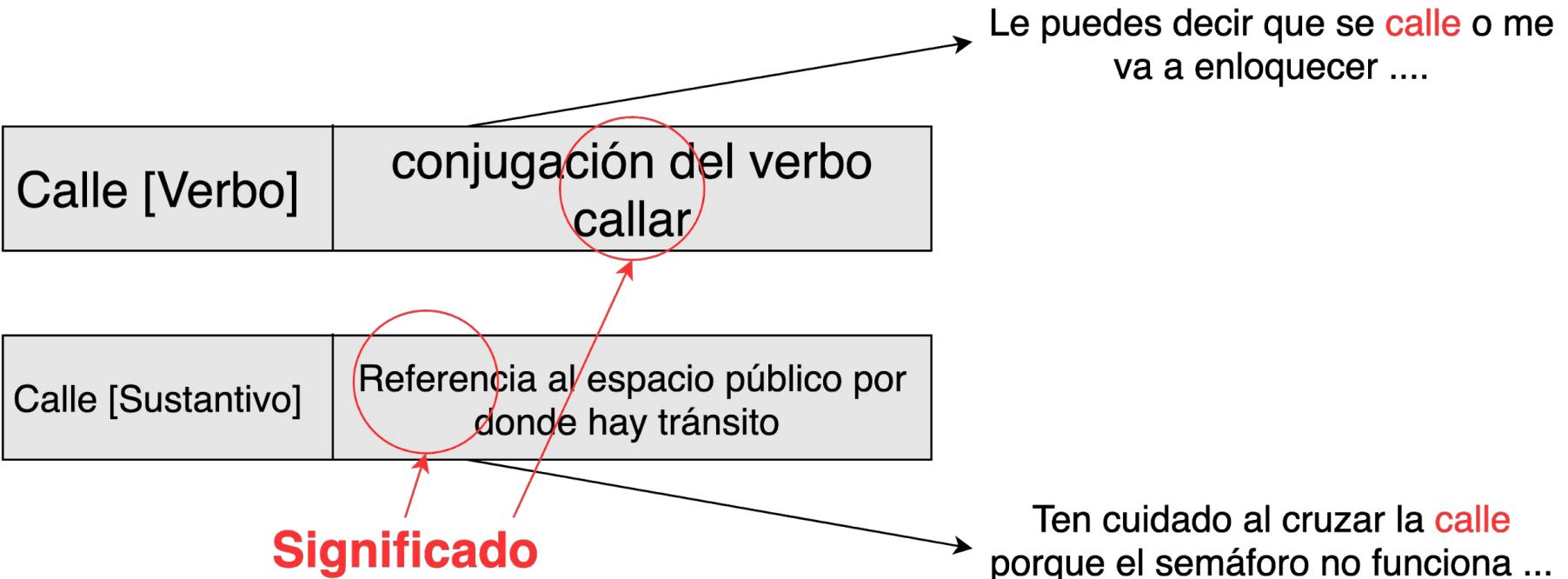
# ¿Cómo es?



# ¿Cómo es?



# ¿Cómo es?

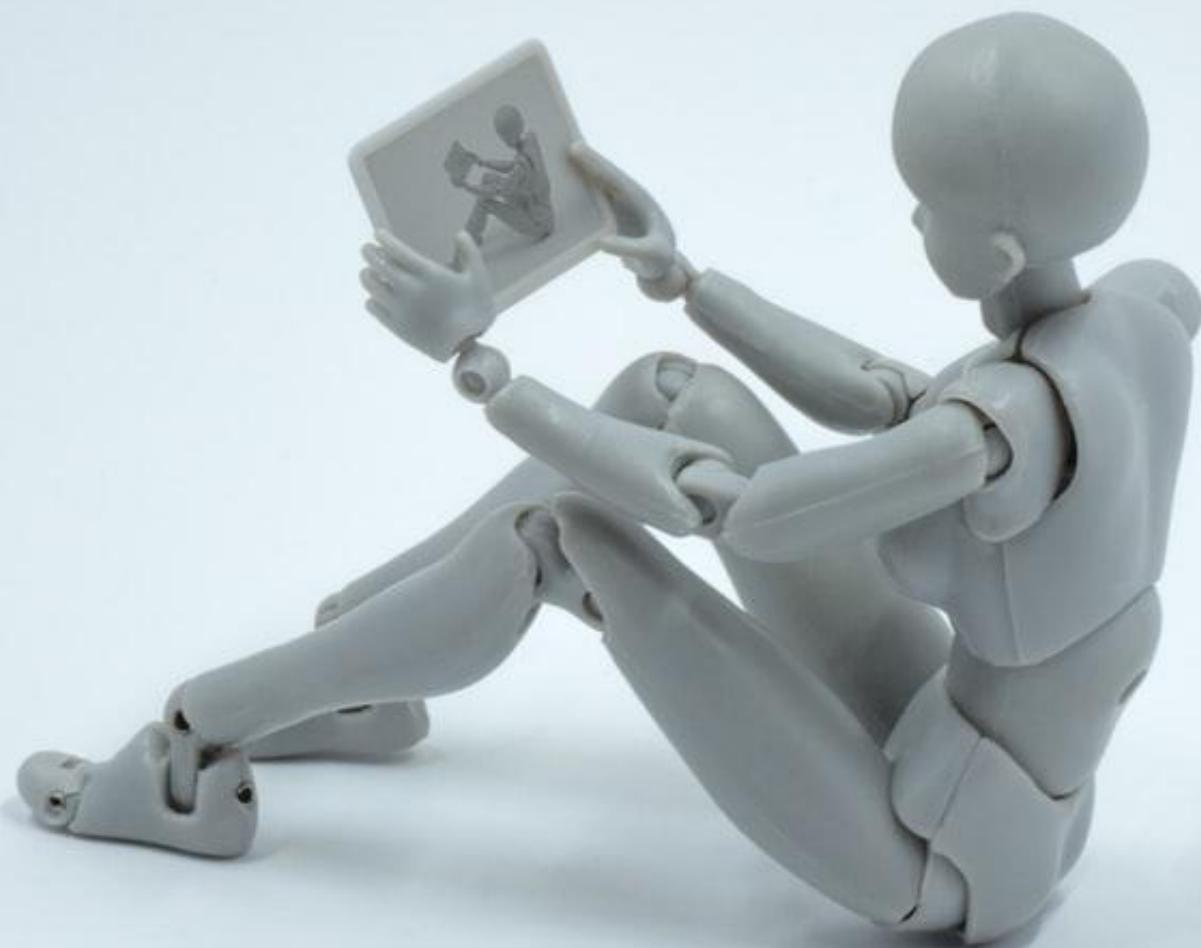


“

Vamos a un notebook en  
Google Colab ...



”



---

[C6-2]

# Fundamentos con NLTK: WordNet

... una base de datos léxica ...

“

Es una base de datos con carácter léxico para el idioma inglés. Se compone por conjuntos de sinónimos (**synsets**), cada uno expresando un concepto diferente. Diferentes **synsets** se relacionan por su relación conceptual semántica ...

”

[WordNet reference, Princeton University](#)

# ¿cómo es un synset?

## Synset

Carro

Automóvil

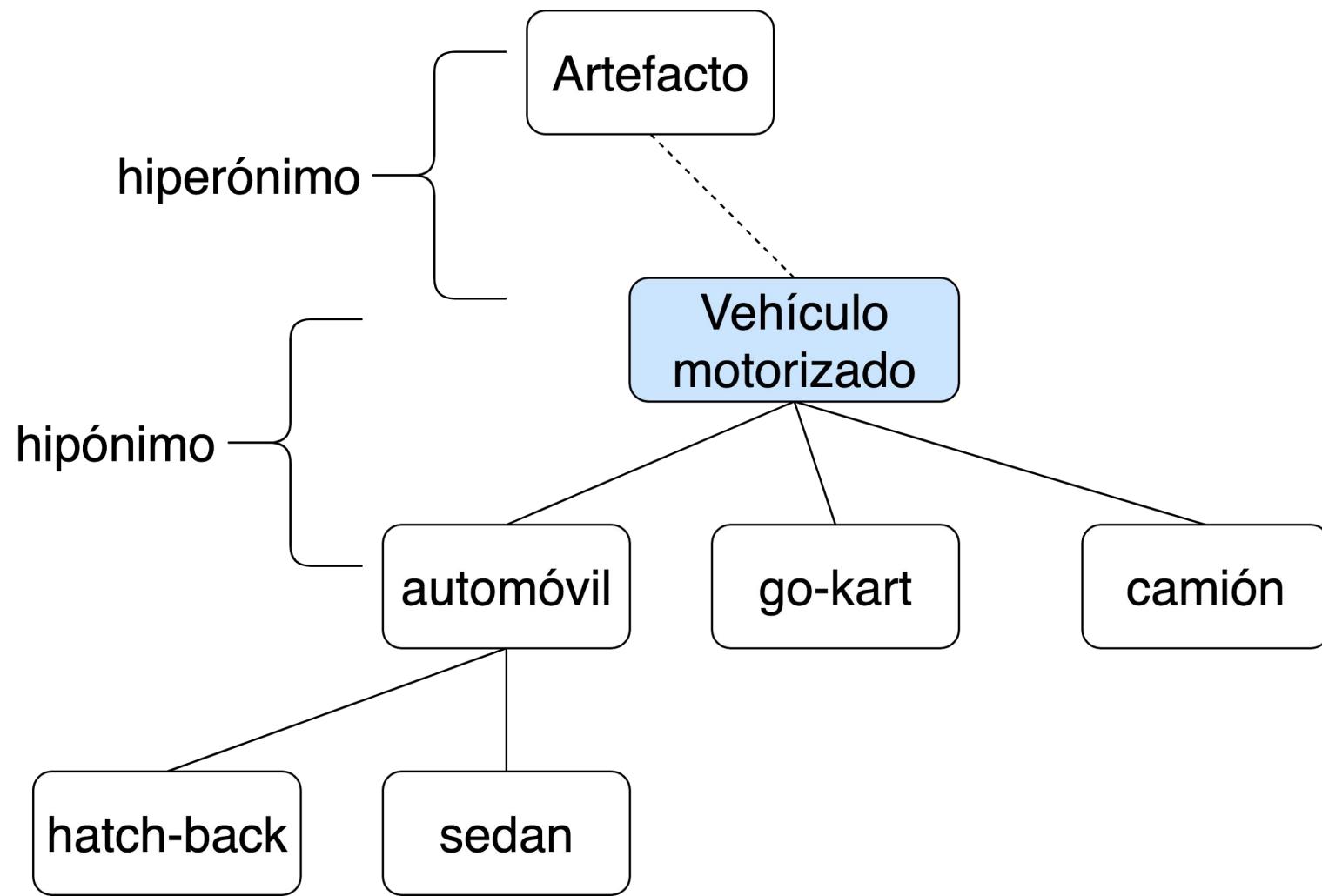
Auto

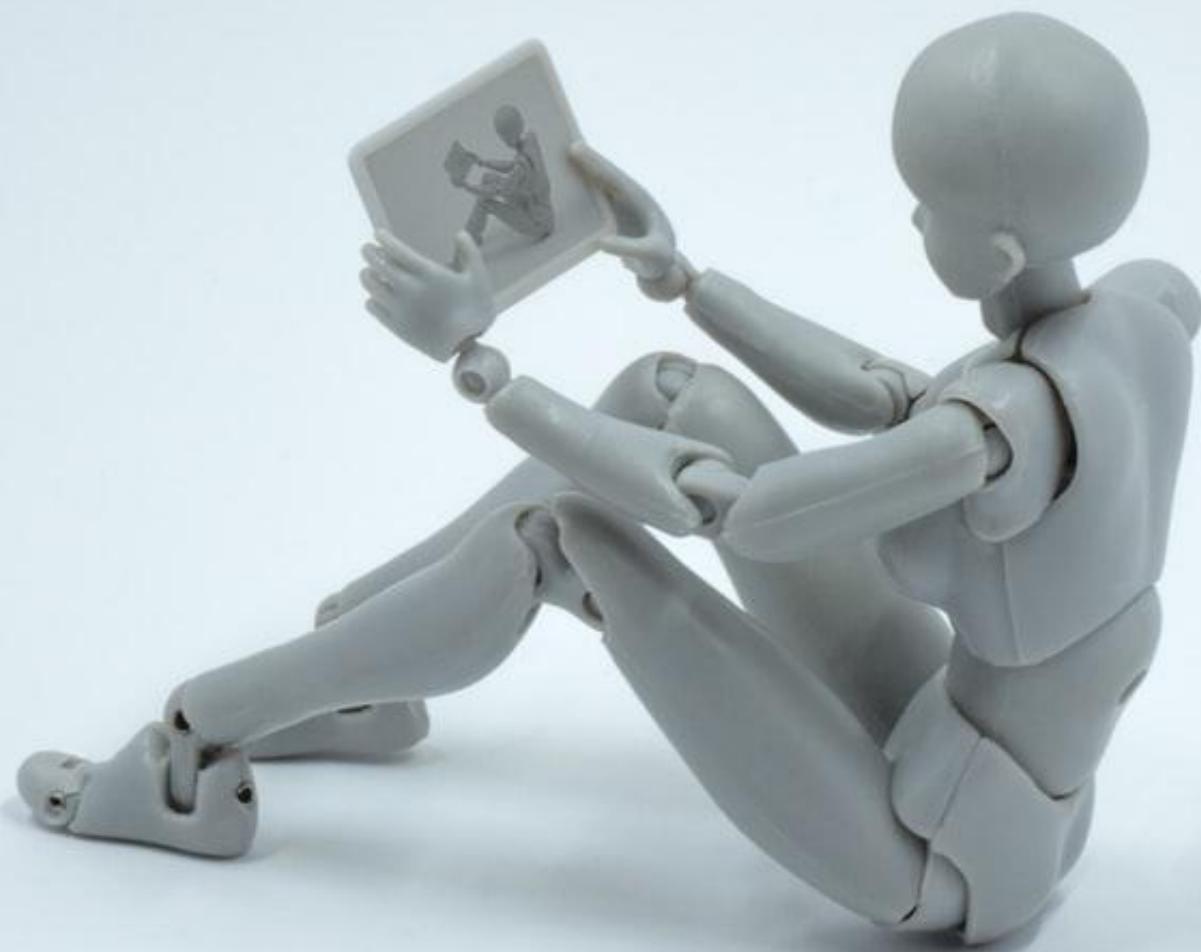
Coche

## Definición:

Vehículo motorizado de cuatro ruedas, impulsado por un motor de combustión interna

# Jerarquía WordNet





---

# [C8] Fundamentos con NLTK: uso de código estructurado

Prácticas avanzadas de escritura de  
código para NLP

```
import re
def get_text(file):
    '''Read text from file ...'''
    text = open(file).read()
    text = re.sub(r'<.*?>', ' ', text)
    text = re.sub(r'\s+', ' ', text)
    return text
```

```
from urllib import request
from bs4 import BeautifulSoup

def freq_words(url, n):
    req = request.urlopen(url)
    html = req.read().decode('utf8')
    raw = BeautifulSoup(html, 'html.parser')
    text = raw.get_text()
    tokens = word_tokenize(text)
    tokens = [t.lower() for t in tokens]
    fd = nltk.FreqDist(tokens)
    return [t for (t, _) in fd.most_common(n)]
```

“

Vamos a un notebook en  
Google Colab ...



”