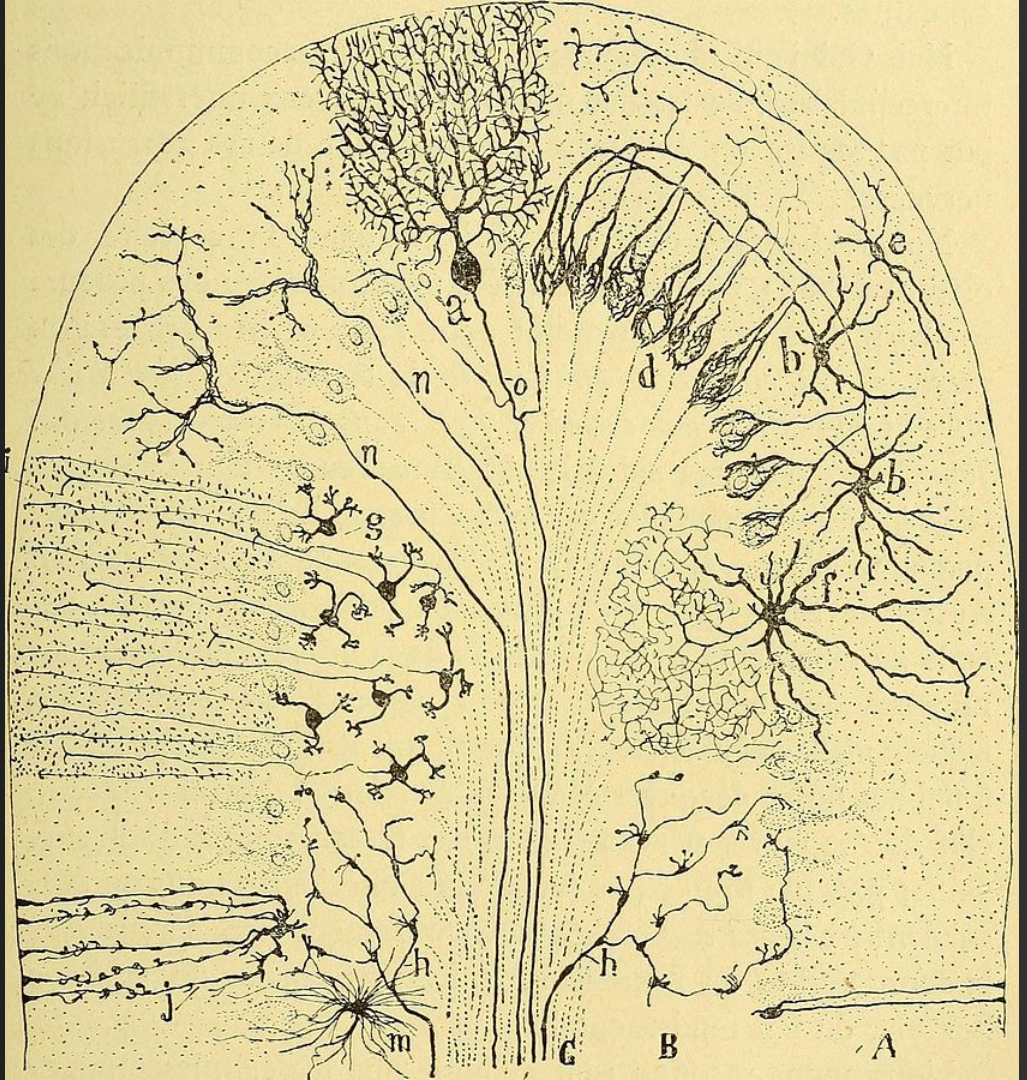


La lengua al microscopio

Introducción básica al PLN desde Humanidades

Elena Álvarez Mellado

¿Lingüística computacional? ¿Procesamiento de Lenguaje Natural?


GACAGACATGACTTTGGATTTCCCCAGGAGGAGTTTGGCAACCAGTTCCAAAAGGCT
GAAACCATCCCTGTCCTCCATGAGATGATCCAGCAGATCTTCAATCTCTTCAGCACA
AAGGACTCATCTGCTGCTTGGGATGAGACCCTCCTAGACAAATTCTACACTGAACTC
TACCAGCAGCTGAATGACCTGGAAGCCTGTGTGATACAGGGGGTGGGGGTGACAGAG
ACTCCCCTGATGAAGGAGGACTCCATTCTGGCTGTGAGGAAATACTTCCAAAGAATC
ACTCTCTATCTGAAAGAGAAGAAATACAGCCCTTGTGCCTGGGAGGTTGTCAGAGCA
GAAATCATGAGATCTTTTTCTTTGTCAACAAACTTGCAAGAAAGTTTAAGAAGTAAG
¿Qué ve un ordenador cuando se enfrenta a un texto?
ATGACCTGGAAGCCTGTGTGATACAGGGGGTGGGGGTGACAGAGACTCCCCTGATGA
AGGAGGACTCCATTCTGGCTGTGAGGAAATACTTCCAAAGAATCACTCTCTATCTGA

Tokenización

Imagen de nicobou

```
>>> import nltk
>>> texto= "¿No es verdad, ángel de amor, que en esta apartada orilla más pura la luna brilla y se respira mejor?"
>>> nltk.wordpunct_tokenize(texto)
['\xbf', 'No', 'es', 'verdad', ',', '\xe1ngel', 'de', 'amor', ',', 'que', 'en', 'esta', 'apartada', 'orilla', 'm\xe1s', 'pura', 'la', 'luna', 'brilla', 'y', 'se', 'respira', 'mejor', '?']
>>>
```

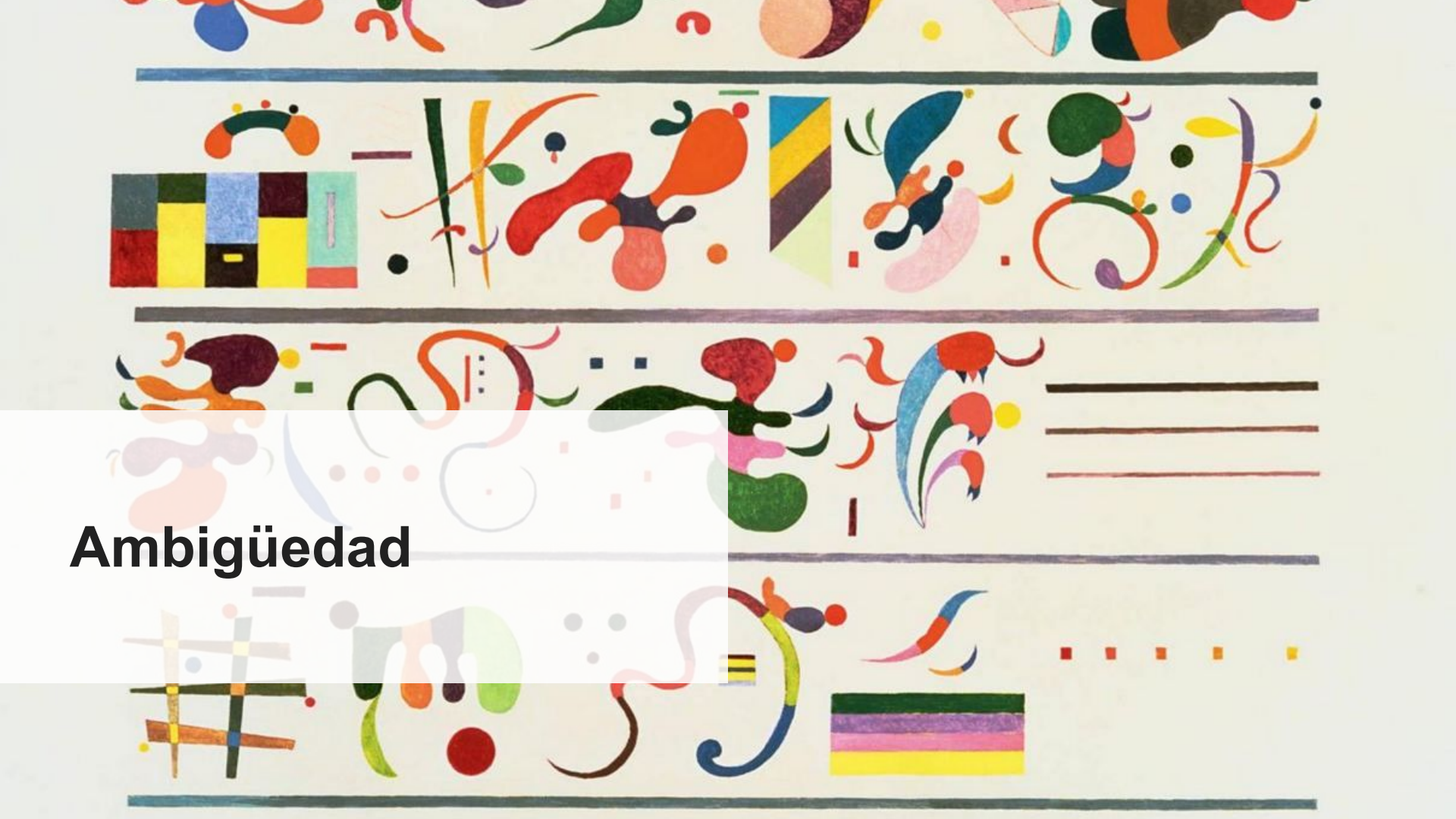
POS-tagging y lematización

Imagen de nicobou

Los ciudadanos reciben el euro con euforia


```
"palabra": "los"
"lema": "el"
"categoria": "articulo"
"analisis": "ADMP"

"palabra": "ciudadanos"
"lema": "ciudadano"
"categoria": "sustantivo"
"analisis": "NCMP"

"palabra": "reciben"
"lema": "recibir"
"categoria": "verbo"
"analisis": "VPI3P"


"palabra": "el"
"lema": "el"
"categoria": "articulo"
"analisis": "ADMP"
```

```
"palabra": "euro"
"lema": "euro"
"categoria": "sustantivo"
"analisis": "NCMP"

"palabra": "con"
"lema": "con"
"categoria": "preposición"
"analisis": "P000"

"palabra": "euforia"
"lema": "euforia"
"categoria": "sustantivo"
"analisis": "NCFS"
```



Ambigüedad

Scheherezade Surià
@Scheherezade_SL

Siguiendo

No deja de sorprenderme la pasmosa facilidad del inglés para convertirlo todo en verbo. Ahora vas y lo traduces. #LibertéFraternitéBeyoncé

to forget what he did. That's scary as fuck too. Someone I've only been with for a year means *that* much to me? But Chris . . . he's different.

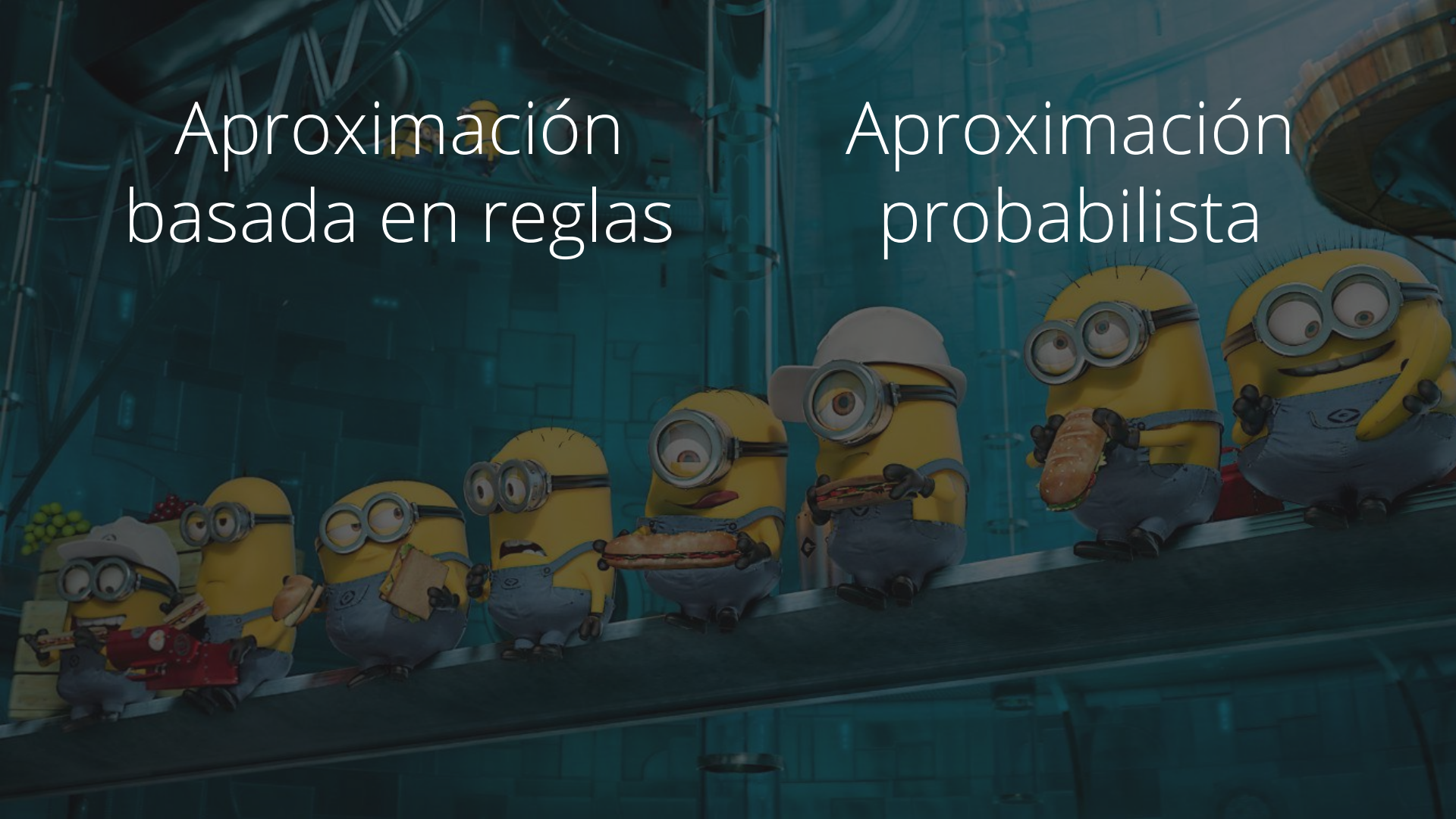
You know what? I'll Beyoncé him. Not as powerful as a nineties R&B breakup song, but stronger than a Taylor Swift. Yeah. That'll work. I tell Hailey and Maya, "I'll handle him."

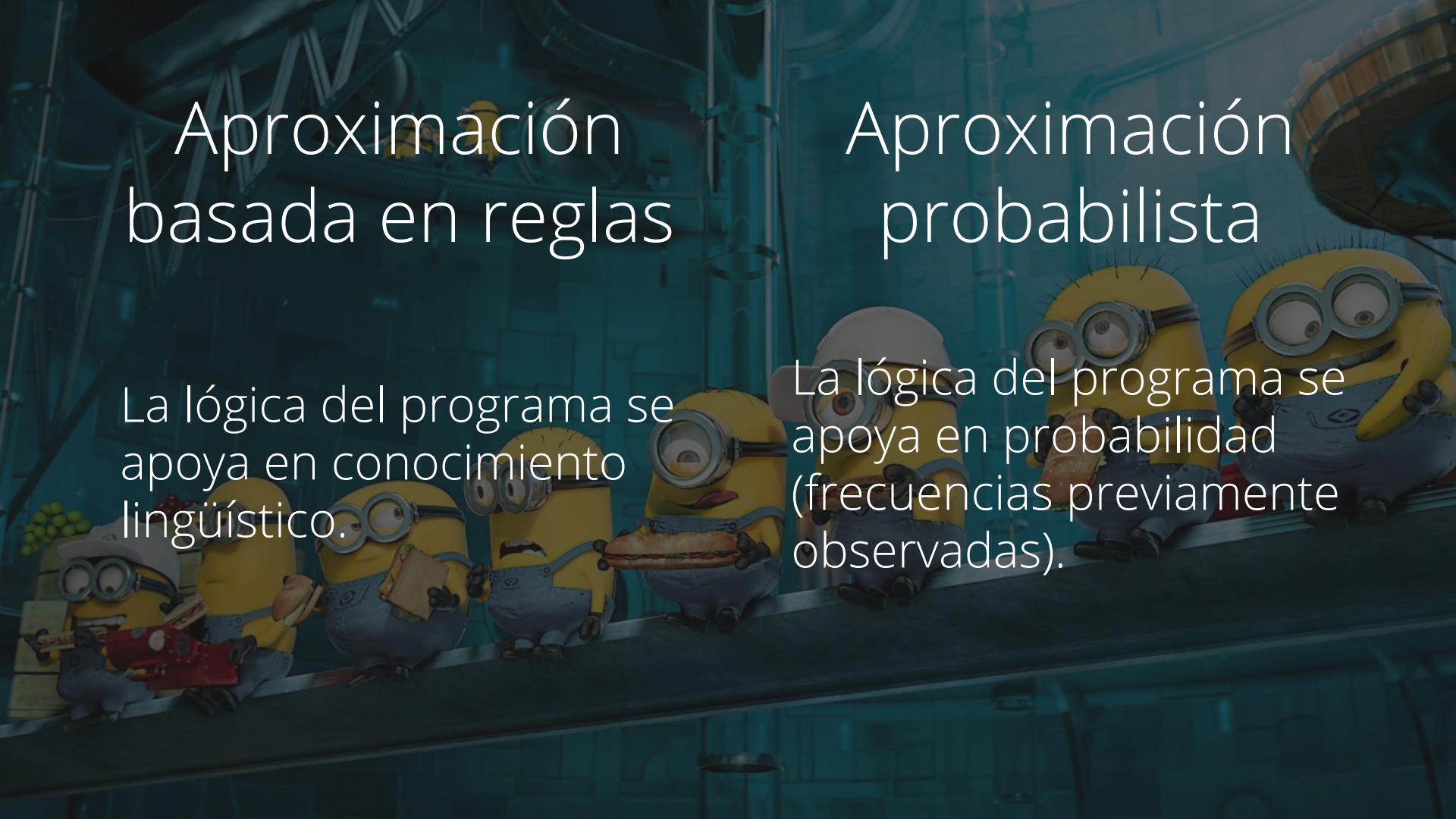
They move so I'm between them like they're my bodyguards, and we go to the door together.

Chris bows to us. "Ladies."

Aproximación basada en reglas

Aproximación probabilista





Aproximación basada en reglas

La lógica del programa se apoya en conocimiento lingüístico.

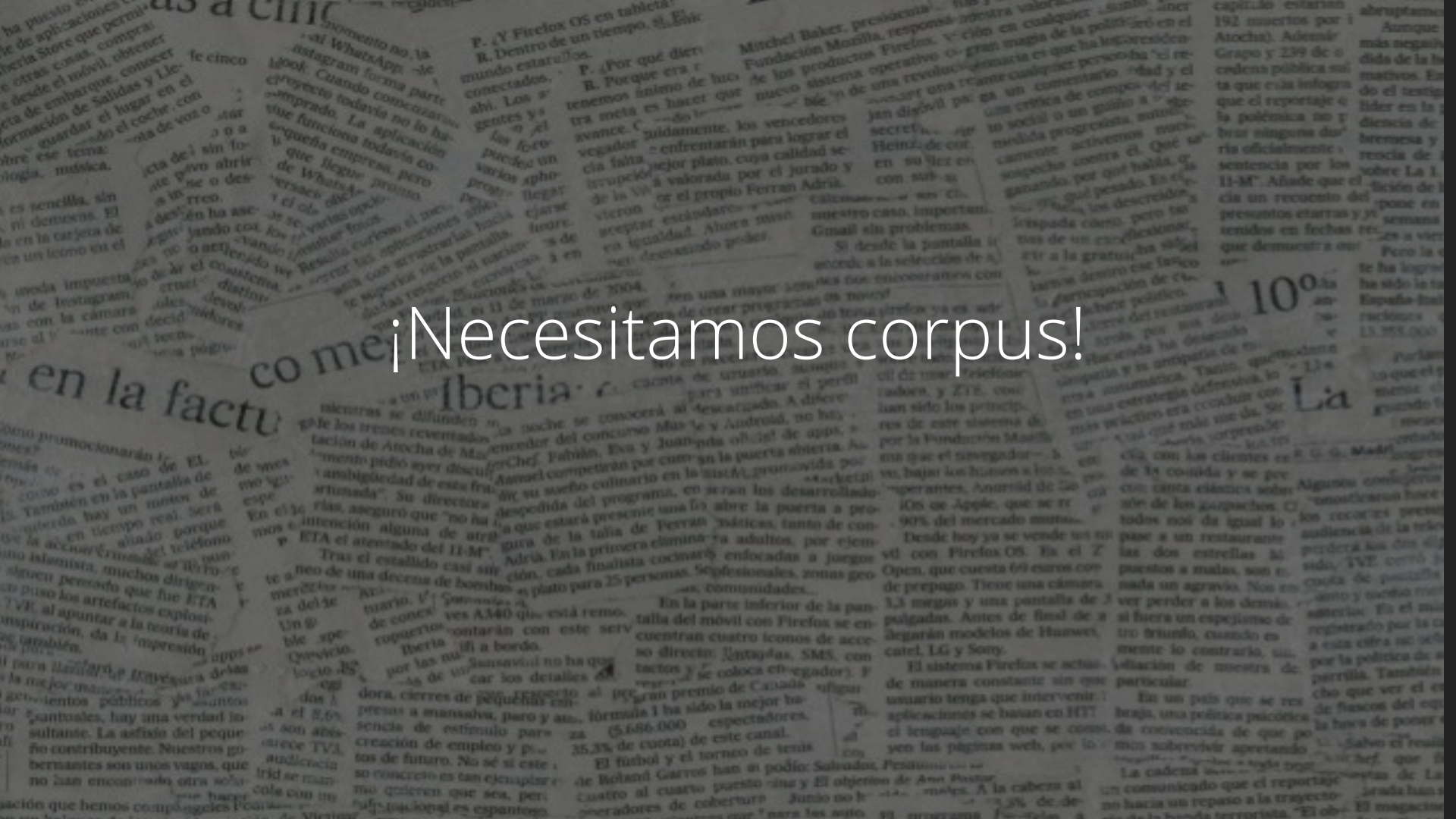
Aproximación probabilista

La lógica del programa se apoya en probabilidad (frecuencias previamente observadas).

Anytime a linguist leaves the group,

the recognition rate goes up.

Frederick Jelinek

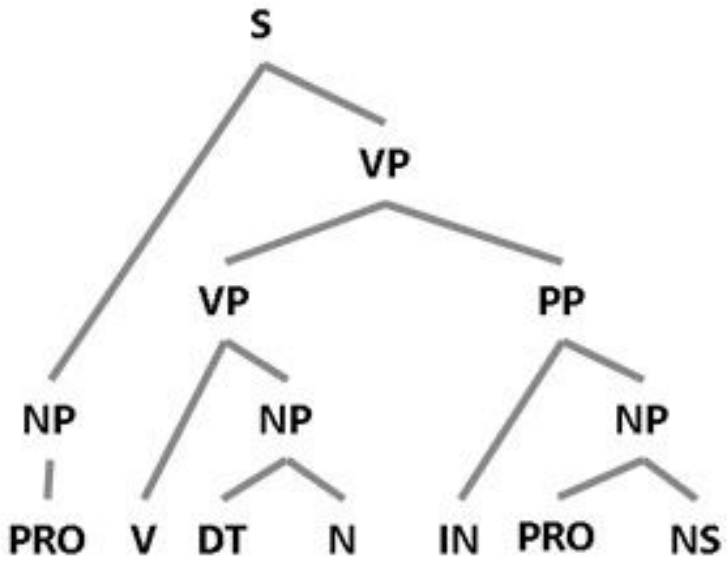

¡Necesitamos corpus!

Análisis sintáctico (parsing)

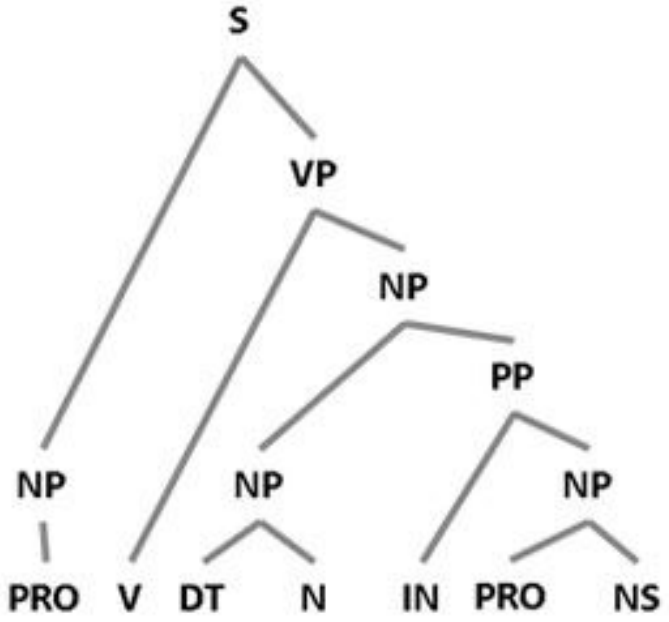
Imagen de nicobou

One morning I shot an elephant in my pajamas. How he got into my pajamas I'll never know.

How Parse Trees Work



I shot an elephant in my pajamas.



I shot an elephant in my pajamas.

Key: N = Noun | NS = Plural Noun | NP = Noun Phrase | PRO = Pronoun | V = Verb | VP = Verb Phrase | DT = Determiner | IN = preposition | PP = Prepositional Phrase

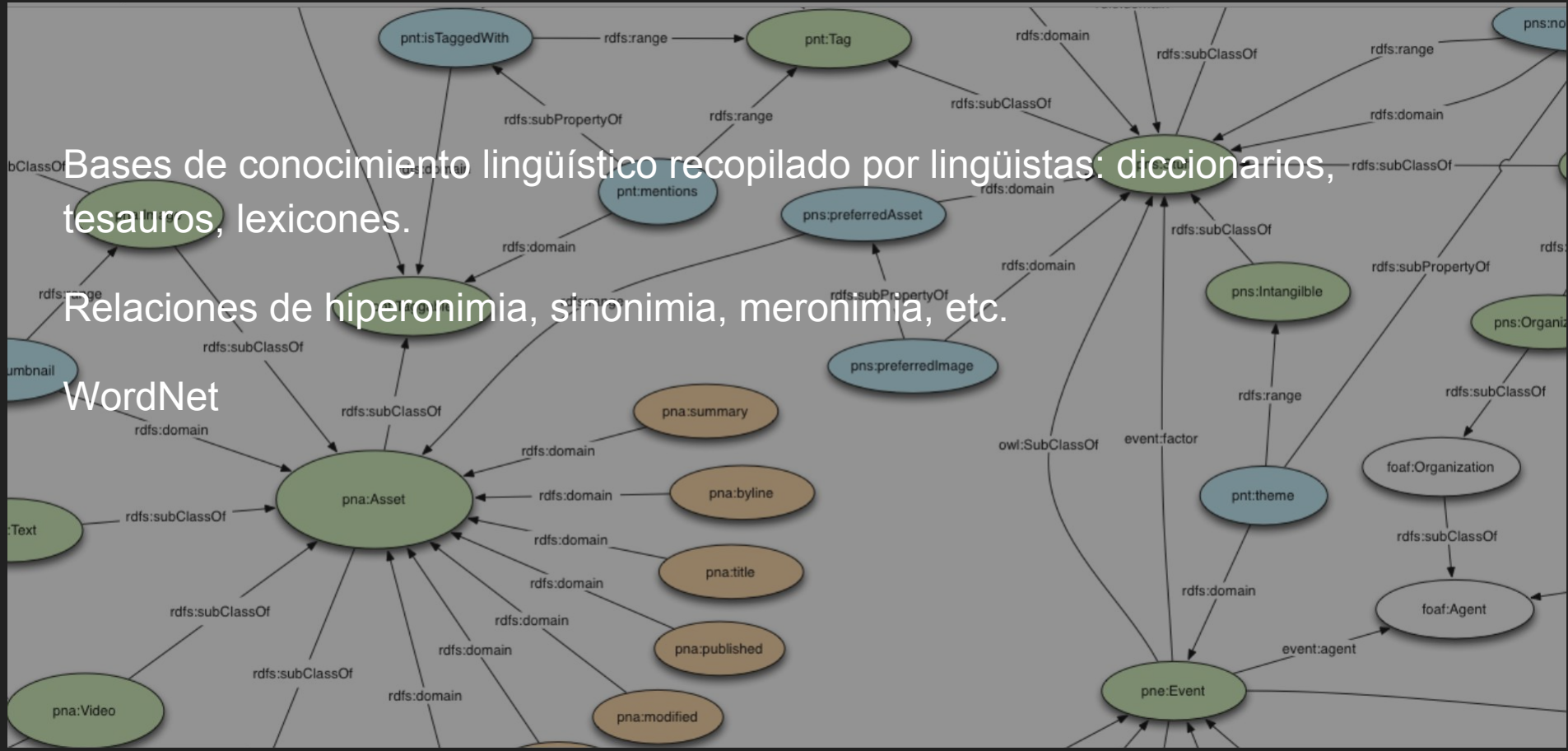
Imagen extraída de Elsevier

Análisis semántico

Imagen de nicobou


This is linguistics. You can study any aspect of language you want, and you will always find structure
But what is that shadowy place over there?
That is semantics. You must never go there

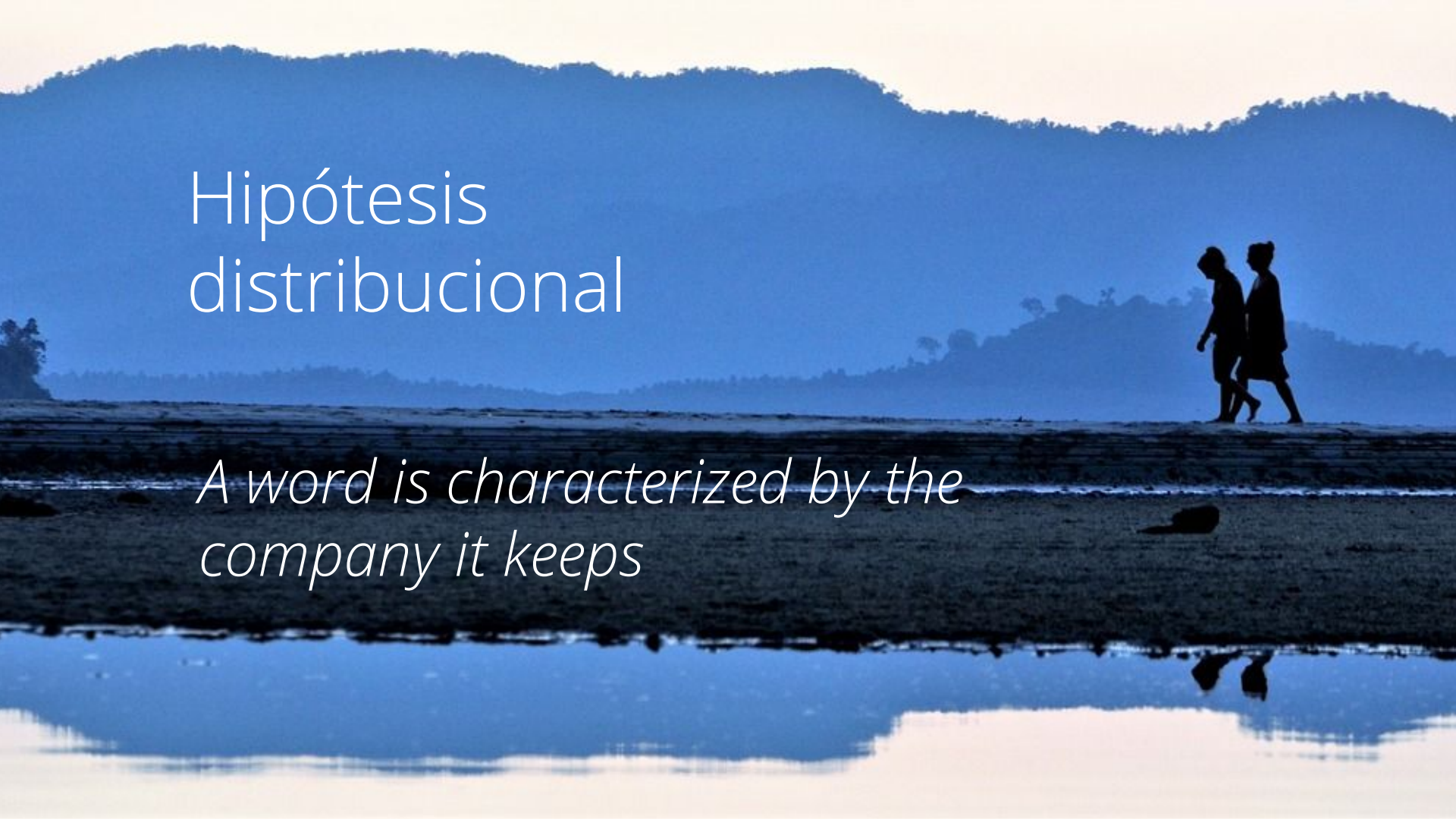
THE THESAURUS



Bases de conocimiento lingüístico recopilado por lingüistas: diccionarios, tesauros, lexicones.

Relaciones de hiperonimia, sinonimia, meronimia, etc.

WordNet



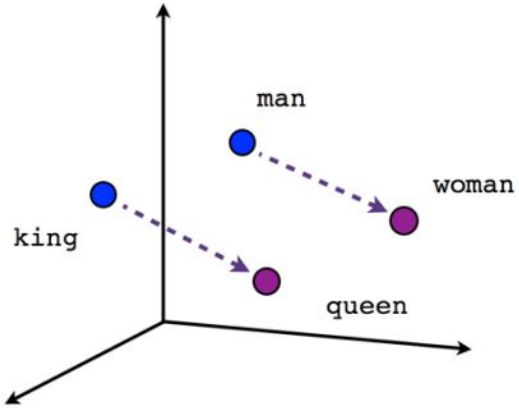
Hipótesis distribucional

A word is characterized by the company it keeps

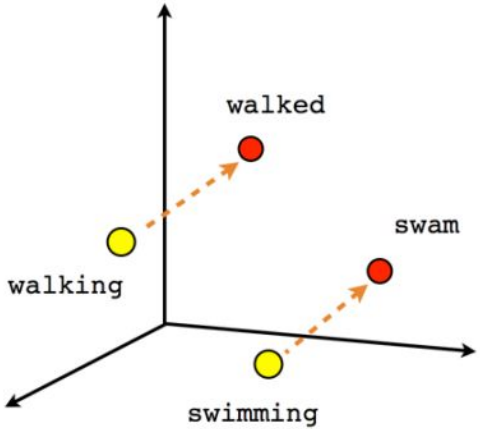
- I like deep learning.
- I like NLP.
- I enjoy flying.

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

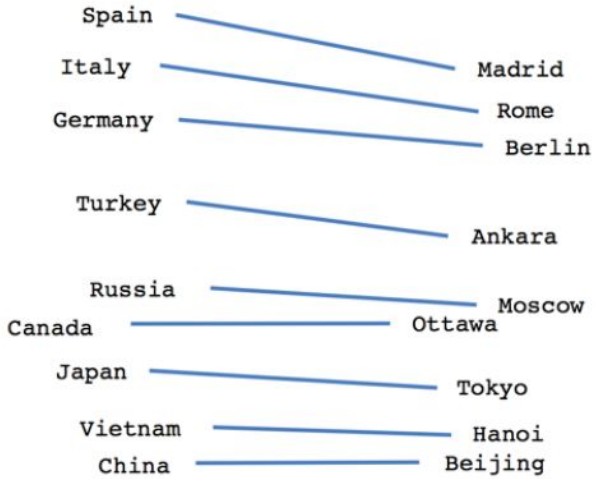
Imagen extraída de CS224 de Stanford



Male-Female



Verb tense



Country-Capital

Imagen extraída de TensorFlow

Todo esto, ¿para qué?


5