

# Análisis de Textos

Grupo PLN - InCo

# Introducción

¿Qué es el PLN?

# Introducción

¿Qué es un lenguaje?

# Introducción

¿Qué es un lenguaje?

(1) Capacidad propia del ser humano para expresar pensamientos y sentimientos por medio de la palabra

(2) Sistema de signos que utiliza una comunidad para comunicarse

(3) Sistema de comunicación estructurado para el que existe un contexto de uso y ciertos principios combinatorios formales

# Introducción

- Alfabeto
- Reglas

```
1 ' Globales -----
2 Var Variable0:Booleano
3 Var Variable1:Cadena
4 ' Fin Globales -----
5 Proc Procedimiento ' <- Procedimiento sin retorno.
6   Var Variable2:Entero ' Locales
7   Var Variable3:Real
8
9   Si Variable0 = Falso Entonces ' Condición "If"
10     Contar Variable2 = 0 a 9 ' Bucle "For"
11     Variable1 = Variable1 + "1"
12     Seguir ' "End For"
13   FinSi ' "End If"
14
15   Variable3 = 5.13
16 FinProc
```

```
>>> for i, v in enumerate(['tic', 'tac', 'toe']):
...     print i, v
...
0 tic
1 tac
2 toe
```

0 → SN SV

SN → DET N

SV → V COMP

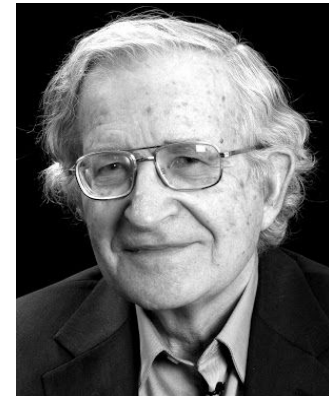
La casa es preciosa

```
# <> isn't actually a valid comparison operator in Python. It's here for the
# sake of a __future__ import described in PEP 401 (which really works :)
comp_op: '<' '>' '==' '!=' '>=' '<=' '<>' '!=' 'in' 'not' 'in' 'is' 'is' 'not'
star_expr: '*' expr
expr: xor_expr ('|' xor_expr)*
xor_expr: and_expr ('^' and_expr)*
and_expr: shift_expr ('&' shift_expr)*
shift_expr: arith_expr (('<<' '>>') arith_expr)*
arith_expr: term (('+' '|-' ) term)*
term: factor (('*' '@' '/' '%' '//') factor)*
factor: ('+' '|-' '~') factor | power
```

# Introducción

¿Qué es un lenguaje?

*Conjunto finito o infinito de oraciones, cada una de las cuales posee una extensión finita, construida a partir de un conjunto finito de elementos.* (Chomsky 1957)



# Introducción

¿Qué es ser natural?

*Adjetivo que refiere a la naturaleza*

Lenguaje Natural

*es la lengua o idioma hablado o escrito por humanos para propósitos generales de comunicación*

# Introducción

¿Qué es el PLN?



# Introducción

## ¿Qué es el PLN?

*El Procesamiento de Lenguaje Natural (PLN) es una subdisciplina de la **Inteligencia Artificial** que intenta resolver con computadoras tareas vinculadas al lenguaje humano, permitiendo la comunicación entre el humano y la computadora a través del lenguaje natural o resolviendo diferentes tareas que implican algún tipo de procesamiento de texto o habla. (Jurafsky & Martin, 2008)*

# Introducción

## ¿Qué es el PLN?

- *es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano* (Wikipedia)
- *conjunto de métodos y técnicas eficientes desde un punto de vista computacional para la **comprensión** y **generación** de lenguaje natural*
- *subdisciplina de la Inteligencia Artificial*

# Introducción

¿PLN = Lingüística Computacional?

# Introducción

## Lingüística Computacional:

- campo multidisciplinario de la **lingüística** y de la **computación**.
- desarrollo de formalismos descriptivos del funcionamiento del lenguaje natural, que puedan ser transformados en programas ejecutables
- involucra a lingüistas, informáticos, lógicos, psicólogos cognitivos...

*El PLN puede verse como la rama ingenieril de la LC*

# Introducción

- **Lingüística Computacional**

se busca fundamentación teórica en los modelos y métodos computacionales propuestos

2 ramas → teórica: desarrollar teorías lingüísticas computables  
→ aplicada: orientación más tecnológica  
(ingeniería lingüística)

*El lenguaje natural es **discreto** en cuanto a sus unidades, pero **infinito** en cuanto a las combinaciones que pueden hacerse de esas unidades*

# Introducción

- **PLN** → tareas que involucran al lenguaje

- Recuperación de información

- Traducción automática

- Respuestas a preguntas

- Análisis de sentimientos



Busca en Google o escribe una URL 

receta de  
merengue  
italiano → 意大利蛋白  
酥皮食譜

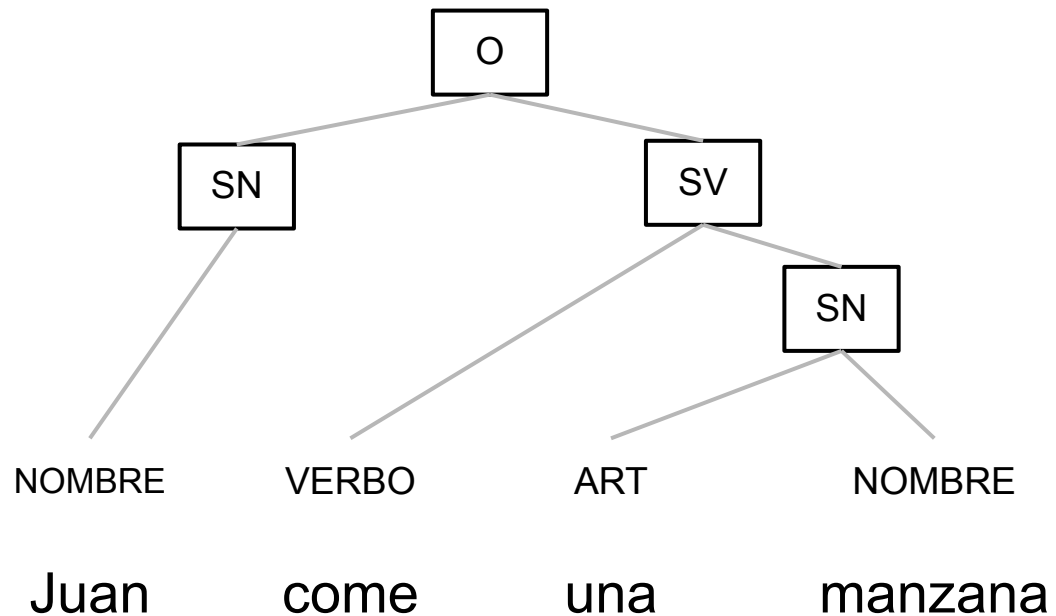


# Introducción

– Parsing

– POS Tagging

– mucho más ...



# Introducción

2 aspectos clave:

1. comprensión
2. generación



# Introducción

2 aspectos claves:

1. comprensión
2. generación

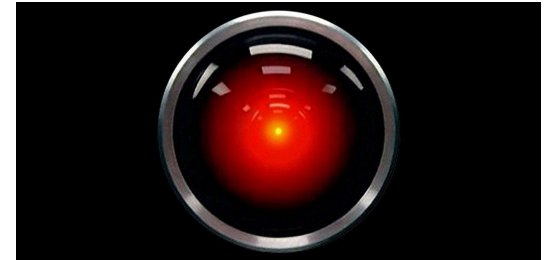
HAL 9000  
(1968)



# Introducción

## 2 aspectos claves:

1. comprensión
2. generación



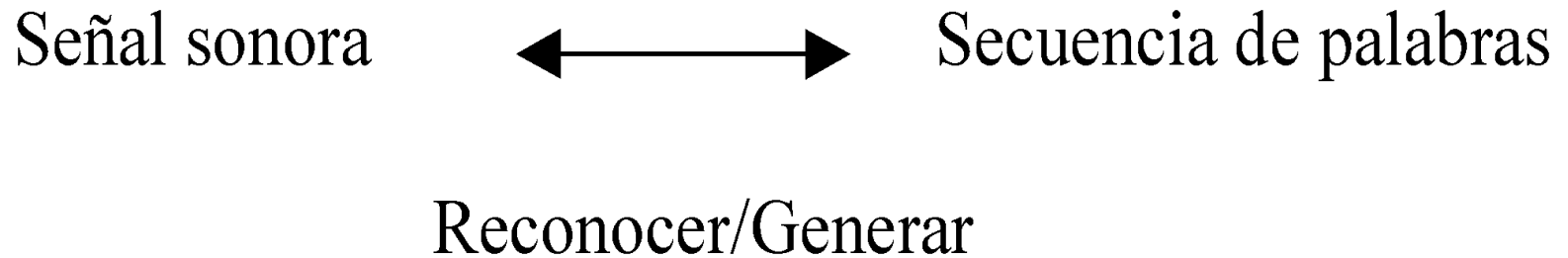
- *Dave*: Open the pod bay doors, HAL.
- *HAL*: I'm sorry Dave. I'm afraid I can't do that.
  
- *Dave*: Abre las compuertas, HAL.
- *HAL*: Lo siento, Dave. Me temo que no puedo hacerlo.

# HAL 9000

## Habilidades de HAL

- comprensión de humanos vía:
  - reconocimiento del habla
  - comprensión de lenguaje natural
- comunicación con humanos vía:
  - generación de lenguaje natural
  - síntesis del habla
- pero también...
  - juega al ajedrez
  - toma decisiones
  - ...

# HAL 9000



## Conocimientos de:

- **Fonética:** naturaleza física de los sonidos
- **Fonología:** cómo los sonidos funcionan en una lengua

# HAL 9000

- Debe saber, por ejemplo:
  - que los sustantivos tienen género y número:
    - Perr-o, Perr-o-s, Perr-a, Perr-a-s.
    - Pero:
      - Cas-a no es el femenino de Cas-o.
      - Ni Luz-s ni Luz-es son plurales de Luz.

# HAL 9000

- Debe saber, por ejemplo:
  - que los sustantivos tienen género y número:
    - Perr-o, Perr-o-s, Perr-a, Perr-a-s.
    - Pero:
      - Cas-a no es el femenino de Cas-o.
      - Ni Luz-s ni Luz-es son plurales de Luz.
  - que se pueden formar palabras agregando prefijos y sufijos a palabras existentes:
    - in-creíble (*in-* denota negación)
    - calmada-mente (*-mente* transforma adjetivo en adverbio)
- Conocimientos de **Morfología**: estudio de la estructura interna de las palabras

# HAL 9000

- Debe conocer el orden correcto en el que las palabras deben decirse para que la respuesta tenga sentido.
  - HAL dice: *Lo siento, Dave. Me temo que no puedo hacerlo.*

# HAL 9000

- Debe conocer el orden correcto en el que las palabras deben decirse para que la respuesta tenga sentido.
  - HAL dice: *Lo siento, Dave. Me temo que no puedo hacerlo.*

Incluso podría decir: *Dave, lo siento. Que no puedo hacerlo me temo.*



# HAL 9000

- Debe conocer el orden correcto en el que las palabras deben decirse para que la respuesta tenga sentido.

– HAL dice: *Lo siento, Dave. Me temo que no puedo hacerlo.*

Incluso podría decir: *Dave, lo siento. Que no puedo hacerlo, me temo.*

Pero ... (\*) *Lo puedo Dave siento que no temo me hacerlo.*

- Conocimientos de **Sintaxis**: estudio de la estructuración (orden y agrupamiento) de las palabras en unidades mayores.

# HAL 9000

- La sintaxis no es suficiente:
  - Abre las compuertas, HAL. (*VC + ART + SUST + SP + SUST*)
  - Baja las persianas, HAL.
  - Saca los dados, HAL.
- Es necesario comprender el **significado** de lo que Dave está diciendo:
  - significado de cada palabra: **Semántica Léxica**
  - significado de la combinación de palabras para obtener: **Semántica Composicional.**

# HAL 9000

- Adicionalmente, HAL presenta una utilización educada del lenguaje:

*Lo siento, Dave. Me temo que no puedo hacerlo.*

- Significa en realidad:

(1) no lo siente

(2) puede abrir las compuertas

- Conocimientos de:

- **Pragmática:** estudio del modo en el que el contexto influye en la interpretación del significado. Cómo el lenguaje se utiliza para ciertos fines.
- **Discurso:** estudio de las unidades mayores a la oración.

# **Etapas clásicas en el Procesamiento de Lenguaje Natural**

# Etapas

- ***Fonética y Fonología***: estudio de los sonidos lingüísticos (usados para la comunicación humana)
- ***Morfología***: estudio de la estructura interna de las palabras
- ***Sintaxis***: estudio de la estructuración (orden y agrupamiento) de las palabras en unidades mayores.
- ***Semántica***: estudio del significado
- ***Pragmática***: estudio de cómo el lenguaje se utiliza para cumplir objetivos
- ***Discurso***: estudio de las unidades mayores a la oración

**Un poco de historia...**

# Década del '50

## Traducción Automática

- En particular del Ruso al Inglés
  - Guerra Fría
  - Experimento *Georgetown* (1954)
    - tenía 6 reglas gramaticales y
    - 250 palabras
  - En tres años la traducción estará resuelta....



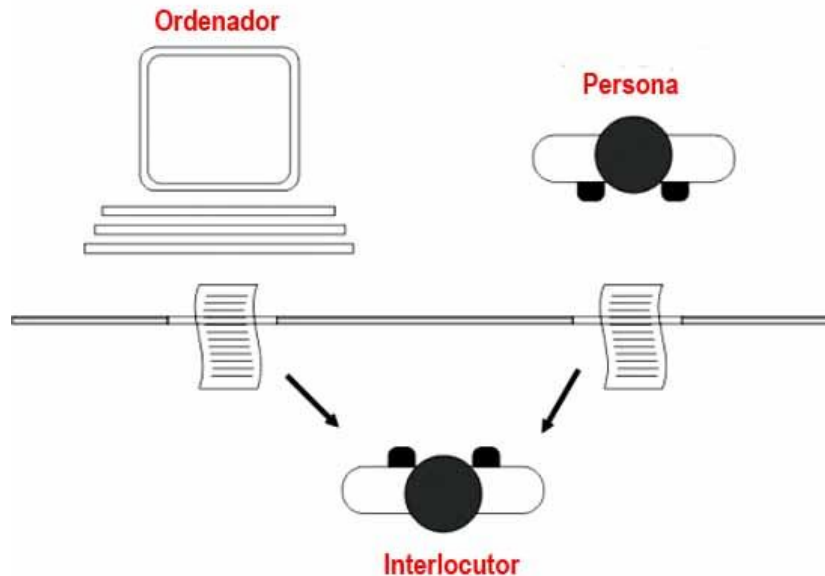
**Warren Weaver**

# Década del '50

## Alan Turing: "Computing Machinery and Intelligence"

(I propose to consider the question, "*Can machines think?*")

### *Test de Turing*



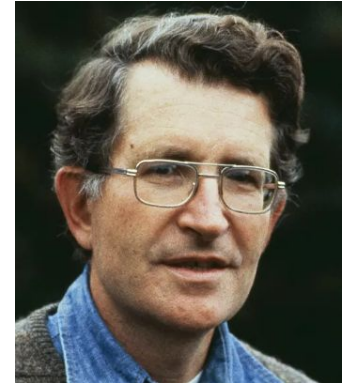


# Década del '50

**Noam Chomsky: "Syntactic Structures" (1957)**

*Colorless green ideas sleep furiously* – 1955

(Las ideas verdes incoloras duermen furiosamente)



*"probabilistic models give no insight into the basic problems of syntactic structure"*

**Gramática Universal** → *todos los idiomas que usamos los seres humanos tienen unas características o principios comunes en su propia estructura*

# Década del '60

- **Cocke – Kasami – Younger (1965)**  
(*Parser*)



- **Earley (1968)**  
(*Otro Parser*)



- **Key (1967)**  
(*Chart parser – Otro Parser*)



El foco está puesto en el Análisis Sintáctico

# Década del '70

**Karen Spärck Jones**



- Recuperación de información: obtener documentos más relevantes dada una consulta
- Inverse Document Frequency (1972)

# Década del '70

Richard Montague



“English as a Formal Language”

- Fue pionero en el enfoque lógico de la semántica del lenguaje natural
- La gramática de Montague está basada en lógica formal:
  - alto orden
  - lambda cálculo
  - semántica intensional, mundos posibles

# Década del '70

## Alain Colmerauer: PROLOG



- Lenguaje de Programación basado en lógica
- Pensado originalmente para Procesamiento de Lenguaje Natural

# Década del '80

- Se construyen sistemas de Laboratorio, con reglas hechas a mano
- Sistemas de interrogación de BDs relacionales
- Uno de los grandes problemas es la portabilidad

***Surgen los sistemas basados en aprendizaje automático***

# Década del '90

**Frederik Jelinek**



- Modelos IBM de traducción estadística y reconocimiento de voz
- A partir del corpus se infieren las reglas
- EL PLN se mueve hacia métodos basados en datos

# Comienzo de este siglo...

## Vladimir Vapnik

- Support Vector Machines
- AT&T, NEC, Facebook IA Research





# Los 2000...

## Geoffrey Hinton, Yoshua Bengio, Yann LeCun

(The Canadian Mafia)

- Deep Learning
- Word Embeddings



---

### Attention Is All You Need

---

Ashish Vaswani\*  
Google Brain  
avaswani@google.com

Noam Shazeer\*  
Google Brain  
noam@google.com

Niki Parmar\*  
Google Research  
nikip@google.com

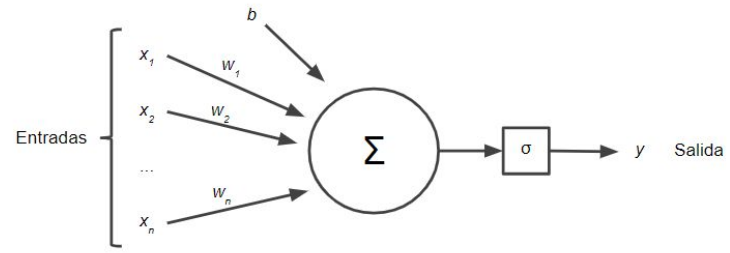
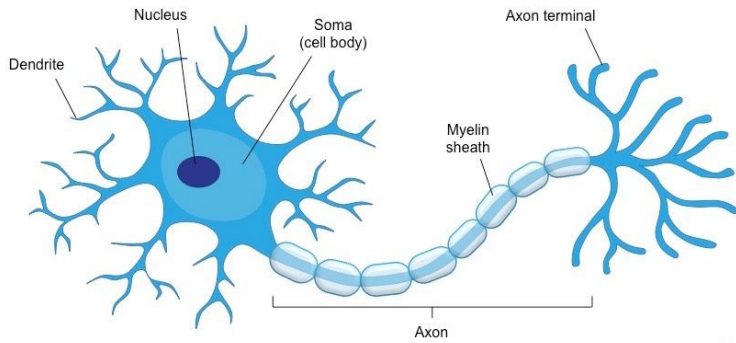
Jakob Uszkoreit\*  
Google Research  
usz@google.com

Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\* †  
University of Toronto  
aidan@cs.toronto.edu

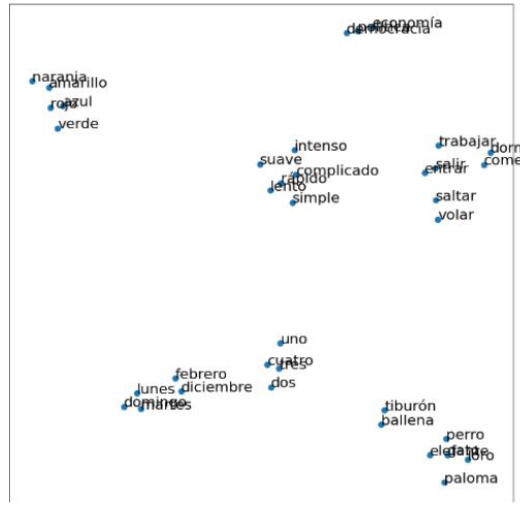
Lukasz Kaiser\*  
Google Brain  
lukaszkaizer@google.com

Illia Polosukhin\* ‡  
illia.polosukhin@gmail.com



$$y = \sigma\left(\sum_i x_i w_i + b\right)$$

Neurona de McCulloch-Pitts, 1943



# La investigación en PLN hoy

- Las compañías de Internet se llevan a los grandes investigadores (casi todos)
- Se publica primero en Arxiv, luego en conferencias, luego en (pocos) journals
- Mucho es *Open Source*
- Casi completamente *data-driven*
- Las redes neuronales son las estrellas al haber cada vez más datos
  - Hoy: Large Language Models (LLM)
- Pero ojo....los métodos simbólicos aún están vigentes

# Algunas aplicaciones ...

- Traducción automática
- Resumen automático
- Recuperación de información
- Extracción de información
- Análisis de sentimientos
- Generación de texto/imágenes
- ...

# Traducción automática

Uno de los primeros problemas de PLN

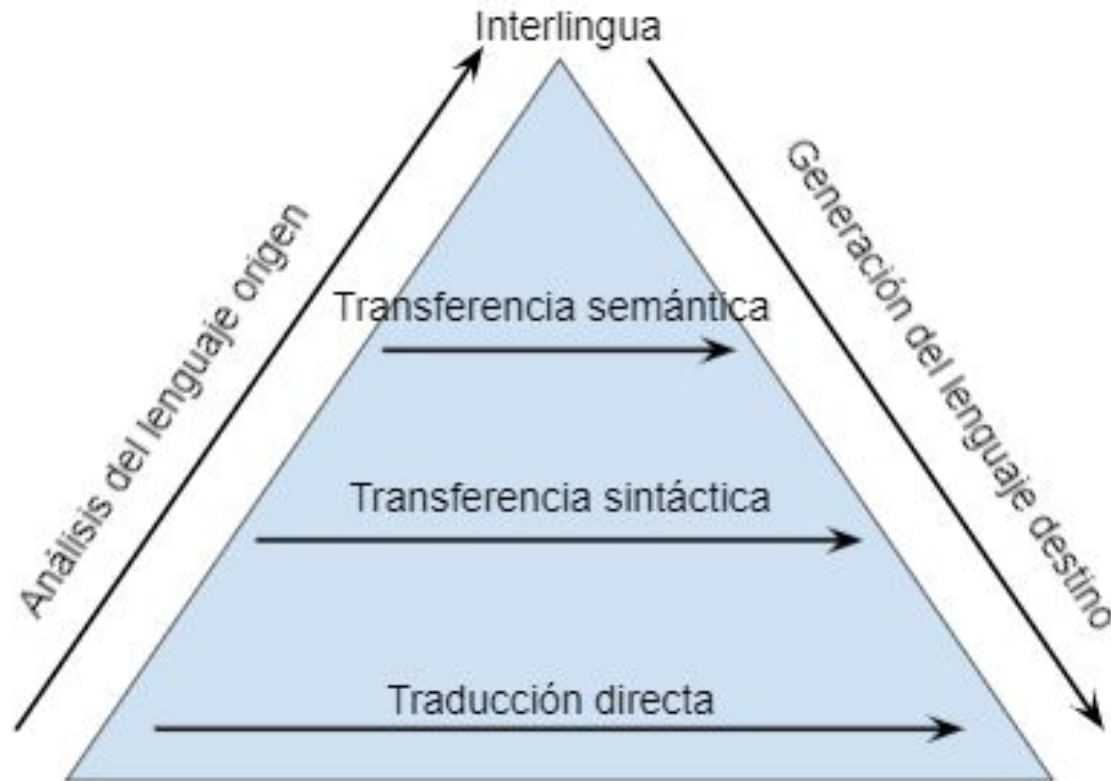
¿Por qué es difícil?

- Tipologías lingüísticas: SVO vs SOV
- Divergencia léxica: pata vs pierna vs leg
- Densidad referencial: sujetos omitidos

...

# Traducción automática

## Triángulo de Vauquois



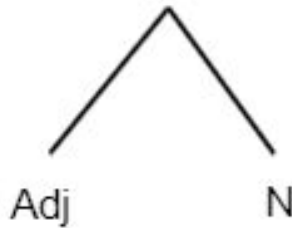
# Traducción automática

## Ejemplo de traducción sintáctica

red rain → lluvia roja

Inglés

Nominal



Nominal

Español



Nom → Adj N / Nom → N Adj

# Traducción automática

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica. La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

*(2013) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the EPL, and are playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted World Player of Sudáfrica. La Inter manages figure is far from the 36 million euros for the striker's release clause. But the club president Massimo Moratti propose an improvement in the player's salary, who according to the Italian daily receive about 4 million euros until 2013.*

*(2014) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the EPL, and are playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted the best player in the World Sudáfrica. La Inter manages figure is far from the 36 million euros of the termination clause of the scorer. But the club president Massimo Moratti propose an improvement in the player's salary, according to the Italian newspaper who will receive about 4 million euros until 2013.*



# Traducción automática

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica. La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

*(2015) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the Calcio, and they are playing a duel to stay with Diego Forlan, the Uruguayan striker who was voted best player of the World Sudáfrica. La Inter figure handles is far from the 36 million euros of the termination clause of the scorer. But the club president Massimo Moratti propose an improvement in the player's salary, who the Italian daily receive about 4 million euros until 2013.*

*(2016) The Italian championship has not started yet but Inter Milan and Juventus, two of the most powerful clubs in the Calcio, and they are playing a duel to stay with Diego Forlan, the Uruguayan striker who was chosen as the best player in the World Sudáfrica. La Inter manages figure is far from the 36 million euros of the termination clause scorer. But the club president Massimo Moratti propose an improvement in the player's salary, according to the Italian daily who will receive about 4 million euros until 2013.*

# Traducción automática

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica. La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

*(2017) The Italian championship has not yet begun but Inter Milan and Juventus, two of Calcio's most powerful clubs, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player of the World Cup in South Africa. Inter's figure is far from the 36 million euros of the scorer's termination clause. But the club that presides Massimo Moratti will propose an improvement in the salary of the player, who according to the Italian newspaper will receive close to 4 million euros until 2013.*

*(2018) The Italian championship has not yet begun but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan forward who was chosen as the best player in the World Cup in South Africa. The amount handled by Inter is far from the 36 million euros of the rescission clause of the scorer. But the club chaired Massimo Moratti propose an improvement in the salary of the player, who according to the Italian newspaper will receive about 4 million euros until 2013.*

# Traducción automática

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica. La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

(2018) *The Italian championship has not yet begun but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan forward who was chosen as the best player in the World Cup in South Africa. The amount handled by Inter is far from the 36 million euros of the rescission clause of the scorer. But the club chaired Massimo Moratti propose an improvement in the salary of the player, who according to the Italian newspaper will receive about 4 million euros until 2013.*

(2019) *The Italian championship has not yet started but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. Inter's figure is far from the 36 million euros of the termination clause of the scorer. But the club chaired by Massimo Moratti will propose an improvement in the salary of the player, who according to the Italian newspaper will receive about 4 million euros until 2013.*

# Traducción automática

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica. La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

*(2020) The Italian championship has not yet started but Inter Milan and Juventus, two of Calcio's most powerful clubs, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. The figure that Inter manages is very far from the 36 million euros of the termination clause of the scorer. But the club chaired by Massimo Moratti will propose an improvement in the salary of the player, who, according to the Italian newspaper, will receive about 4 million euros until 2013.*

*(2021) The Italian championship has not yet started but Inter Milan and Juventus, two of Calcio's most powerful clubs, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. The figure that Inter manages is very far from the 36 million euros of the termination clause of the scorer. But the club chaired by Massimo Moratti will propose an improvement in the player's salary, who according to the Italian newspaper will receive about 4 million euros until 2013.*

# Traducción automática

El campeonato italiano aún no ha comenzado pero Inter de Milán y Juventus, dos de los clubes más poderosos del Calcio, ya están jugando un duelo para quedarse con Diego Forlán, el delantero uruguayo que fue elegido como el mejor jugador del Mundial de Sudáfrica. La cifra que maneja Inter está muy lejos de los 36 millones de euros de la cláusula de rescisión del goleador. Pero el club que preside Massimo Moratti propondrá una mejora en el salario del jugador, quien según el diario italiano recibirá cerca de 4 millones de euros hasta 2013.

*(2021) The Italian championship has not yet started but Inter Milan and Juventus, two of Calcio's most powerful clubs, are already playing a duel to stay with Diego Forlán, the Uruguayan striker who was chosen as the best player in the World Cup in South Africa. The figure that Inter manages is very far from the 36 million euros of the termination clause of the scorer. But the club chaired by Massimo Moratti will propose an improvement in the player's salary, who according to the Italian newspaper will receive about 4 million euros until 2013.*

*(2024) The Italian championship has not yet started but Inter Milan and Juventus, two of the most powerful clubs in Calcio, are already playing a duel to win Diego Forlán, the Uruguayan striker who was chosen as the best player at the World Cup in South Africa. The figure that Inter manages is very far from the 36 million euros of the scorer's termination clause. But the club chaired by Massimo Moratti will propose an improvement in the player's salary, who according to the Italian newspaper will receive around 4 million euros until 2013.*

# Traducción automática

The screenshot shows the DeepL Translator website. At the top, there is a navigation bar with the DeepL logo, a 'Translator' dropdown menu, and links for 'DeepL Pro', 'For Business', 'Why DeepL?', 'API', and a 'Start free trial' button. Below the navigation bar, there are three main service buttons: 'Translate text' (32 languages), 'Translate files' (.pdf, .docx, .pptx), and 'DeepL Write BETA' (AI-powered edits). The main interface features a language selection bar with 'English (detected)' on the left and 'Spanish' on the right, along with 'Automatic' and 'Glossary' options. The input text 'the doctor and the nurse' is shown in a large text area on the left, and the translated text 'el médico y la enfermera' is shown in a large text area on the right. Below the translated text, there are 'Alternatives:' listed: 'el doctor y la enfermera' and 'el médico y el enfermero'. At the bottom of the interface, there are icons for voice input, playback, and sharing.

The screenshot shows the Google Traductor website. At the top, there is a navigation bar with the Google Traductor logo, a settings icon, a grid icon, and a flag icon. Below the navigation bar, there are four main service buttons: 'Texto', 'Imágenes', 'Documentos', and 'Sitios web'. The main interface features a language selection bar with 'Detectar idioma' on the left and 'español', 'inglés', and 'guaraní' on the right. The input text 'the doctor and the nurse' is shown in a large text area on the left. The translated text is shown in a large text area on the right, with two options: 'La doctora y la enfermera (femenino)' and 'El doctor y el enfermero (masculino)'. Below each translated text, there are icons for voice input, playback, and sharing. A note above the translated text says 'La traducción cambia según el género Más información'.

# Resumen automático

“a summary is a condensed version of a source document having a recognizable genre and a very specific purpose: to give the reader an exact and concise idea of the contents of the source”

*(Saggion y Lapalme - Generating Indicative-Informative Summaries with SumUM, Computational Linguistics 28 - 2002)*

# Resumen automático

- Idea central: "condensación del contenido de la información de un documento para el beneficio de un lector" (Mani 2001).
- Primeros trabajos de Luhn (1958) y Edmunson (1960):
  - Basados en métodos estadísticos.
  - Extraen las oraciones más importantes.
  - Frecuencia de términos. Peso de oraciones.
- Los trabajos en el área resurgen a fines de los años 90'



# Resumen automático

Los paradigmas principales:

- **extracción:** extracción de oraciones o fragmentos de oraciones literales relevantes del texto original
- **abstracción:** regeneración de los fragmentos relevantes del texto original

# Resumen automático

## Aspectos a considerar

<b>FUENTE</b>	<u>Único documento</u> / Varios documentos
	<u>Dominio específico</u> / <u>Ámbito general</u>
	<u>Texto monolingüe</u> / <u>Texto multilingüe</u>
<b>RESULTADO</b>	<u>Resumen por extracción</u> / Resumen por abstracción
	<u>Resumen neutral</u> / Resumen evaluativo
<b>PROPÓSITO DEL RESUMEN</b>	<u>Resumen informativo</u> / Resumen indicativo
	<u>Necesidades del autor</u> / Necesidades del usuario
	<u>Destinatario experto</u> / Destinatario lego

<https://www.tdx.cat/bitstream/handle/10803/7508/tidcf.pdf.pdf?sequence=1>

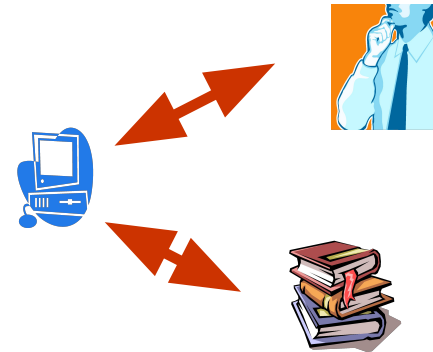
# Recuperación de información

Dada una colección de documentos:

- ¿cómo podemos recuperar información relevante para nosotros contenida en ellos?
- ¿qué cosas serán representativas de cada documento?
- ¿cómo se van a representar los documentos?
- ¿y las consultas?
- ¿importa la estructura del documento o sólo su contenido?
- ¿cuál va a ser la relevancia asociada a cada documento?

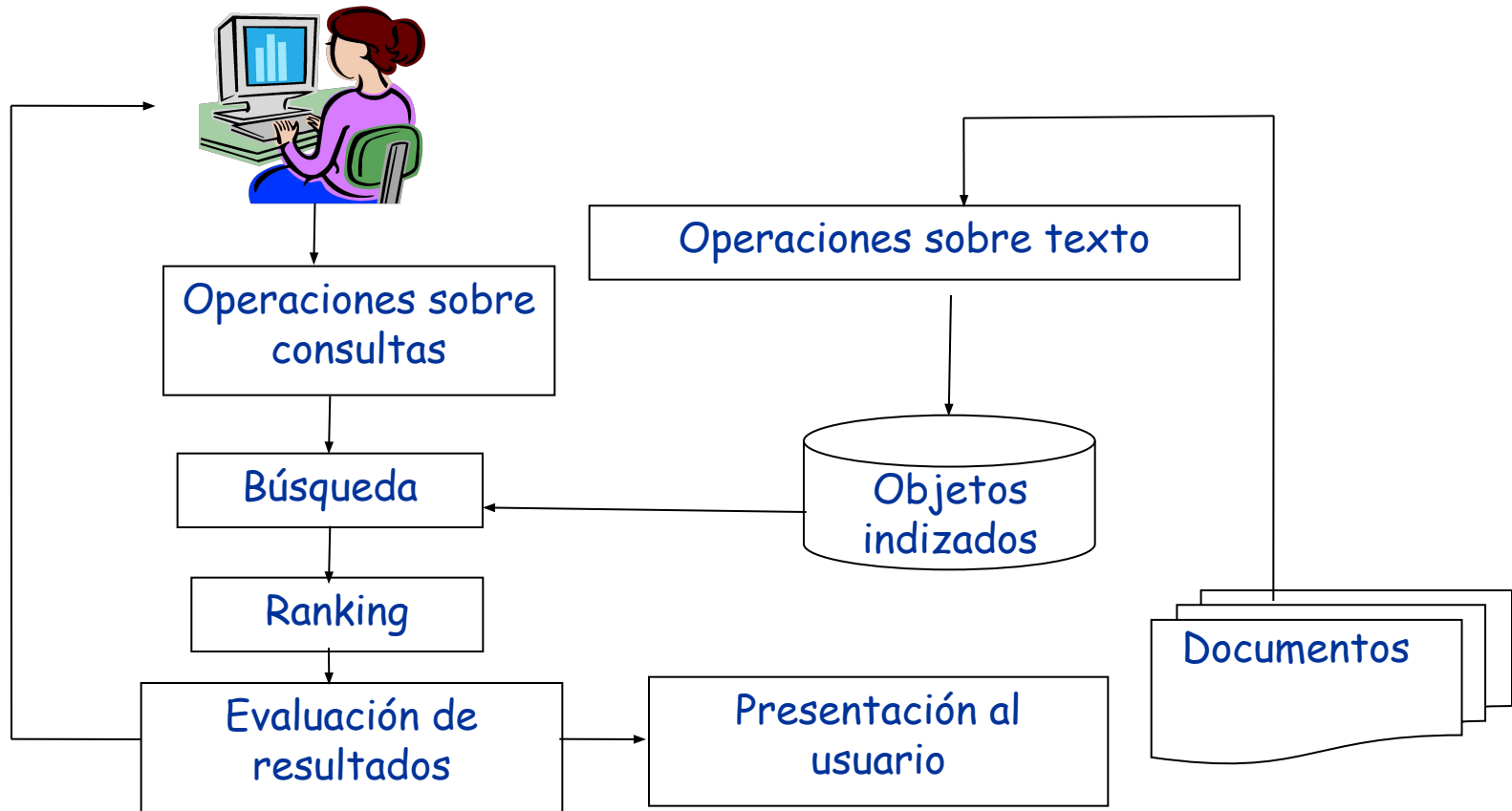
# Recuperación de información

- Idea central: "*disciplina encargada de la representación, almacenamiento y su posterior acceso y recuperación para responder a las necesidades de un usuario*" (Salton 1968).



- Primeros trabajos :
  - Basados en modelos vectoriales
  - Documentos y consultas representados por vectores de términos
  - Introducen el concepto de grado de similitud
  - Se evalúan con las medidas *Precision* y *Recall*
- Los trabajos en el área comienzan a ser evaluados en las TREC ('90)

# Recuperación de información



# Recuperación de información

## Problemas:

- muchas formas de decir lo mismo
- un mismo término puede tener distintos significados en diferentes contextos
- predecir cuáles documentos son los más relevantes

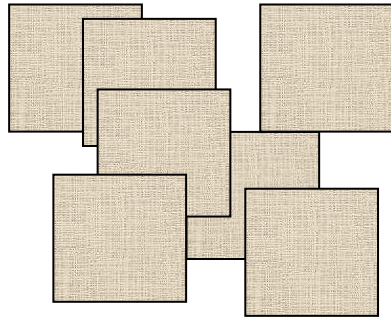
# Extracción de Información

- Objetivo: responder consultas sobre textos
- Analizar texto sin restricciones para extraer cierto tipo de información contenida en él
- Intenta convertir información no estructurada según un “esquema o plantilla”
- Típicamente extrae:
  - entidades
  - relaciones
  - eventos

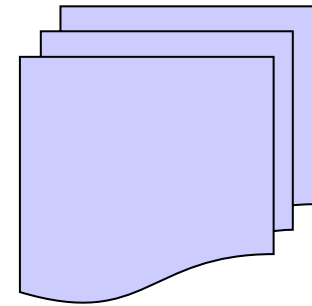
# EI & RI

La idea es que un SRI y un SEI se complementen ...

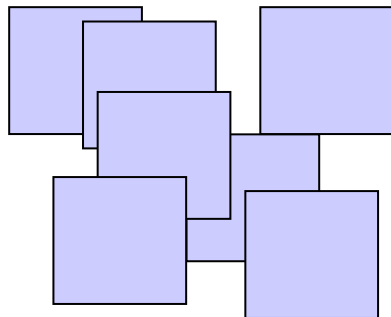
En los SRI



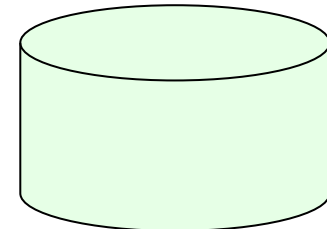
Se obtienen documentos



En los SEI



Se obtienen datos o  
información útil





# Análisis de Sentimiento

## Análisis de Sentimiento o Minería de Opiniones

(Turney, 2002; Wiebe et al., 2005; Pang & Lee, 2008; Liu, 2010/ 2020; competencias TASS (SePLN), SemEval)

### Diferentes tareas:

- Clasificación de textos (tweets, opiniones sobre películas, hoteles, noticias, etc.) según su polaridad (positivos o negativos)
- Extraer opiniones a partir de textos (pej. prensa).
- Para cada opinión:
  - Autor
  - Polaridad (P, N, P+, N+, Neu, etc.)
  - Aspectos: celular -> batería P, pantalla N
  - ...

# Análisis de Redes Sociales

## Desafíos:

### ➤ Lenguaje informal

El lenguaje usado tiene errores ortográficos y no siempre ayudan los correctores

*“kereees decirM ques esssssooo #Tasahi”*



# Análisis de Redes Sociales

## Desafíos:

- Formas sutiles de expresar sentimientos:
  - “Mario Benedetti es un excelente escritor” **Palabras con polaridad**
  - “Se trata de la fragancia que quería, por favor, úsela exclusivamente en casa y cierre las ventanas” **Sin palabras negativas**
  - “Si, seguro, cómo no!!” **Ironía**
  - “Vamos a ponerse las pilas” **Expresiones coloquiales**
  - “Si creíste que iba a ser una buena película, no fue una buena elección” **Negación**

# Generación de texto

PR

Armame un texto extraño de confirmación de asistencia a un cumpleaños de 60 años de máximo 30 palabras.



¡Alerta, seres extraterrestres! Confirmo mi presencia en la extravagante celebración de los 60 años del humano conocido como [Nombre]. Estoy listo para mezclarme entre ustedes y experimentar la diversión alienígena. ¡Que comienza la anomalía festiva!



PR

Armame un texto de confirmación de asistencia a un cumpleaños de 60 años de máximo 30 palabras.



¡Claro que sí! Estoy emocionado de unirme a la celebración de tus 60 años. Contad con mi presencia para festejar contigo este importante hito. ¡Nos vemos pronto!



PR

Ahora hazelo pero es español rioplatense



¡Claro que sí, che! Estoy re emocionado de sumarme a la joda de tus 60 pirulos. Pueden contar con mi presencia para festejar este re lindo momento juntos. ¡Nos vemos pronto, loco!



# Generación de texto

## Algunos problemas...

LU ¿Cuál fue la primera arquitecta uruguaya?

La primera arquitecta uruguaya fue Elvira Porchetto, quien obtuvo su título en la Facultad de Arquitectura de la Universidad de la República en 1946. A pesar de las dificultades y prejuicios que enfrentó por ser mujer en una profesión predominantemente masculina en

LU No, la primera arquitecta fue Julia Guarino.

Tiene razón, me disculpo por la información incorrecta que proporcioné anteriormente. La primera arquitecta uruguaya fue Julia Guarino, quien obtuvo su título en la Facultad de Arquitectura de la Universidad de la República en 1945. Guarino abrió el camino para

LU No, Julia Guarino se recibió de arquitecta en 1923.

Tiene razón nuevamente y le agradezco por corregir mi error. La primera arquitecta uruguaya fue Julia Guarino, quien se graduó de la Facultad de Arquitectura de la Universidad de la República en 1923. Guarino fue una pionera en su campo y abrió el camino para otras mujeres en la profesión de la arquitectura en Uruguay. Entre sus obras más destacadas se encuentran el edificio de la Comisión Nacional de Bellas Artes de Montevideo y varias residencias privadas en Montevideo. Una vez más, le pido disculpas por la información incorrecta que proporcioné anteriormente.

"¿Qué le dijo un jardinero a otro? - Disfrutemos mientras podamos". Podrías explicarme el chiste?

El chiste se refiere a la idea de que hay que disfrutar la vida mientras sea posible, ya que nadie sabe cuándo la situación puede cambiar. Esta idea es un recordatorio de que no hay que tomar los momentos de la vida como dados, sino que hay que disfrutarlos al máximo. En el chiste, el jardinero está recordando a su compañero que hay que aprovechar el tiempo que tenemos, ya que nunca sabemos cuándo se nos acabará.

Qué le dijo un jardinero a otro? Disfrutemos mientras podamos. Me podés explicar el chiste?

os bajo

rrados  
ntre dos  
tapea s  
iapear c

ire un a  
lenguaj  
pedad

ije  $L = \{$

ma cantidad de unos. Si aplicamos un homomorfismo  
ríamos el lenguaje  $L' = \{1^n 1^n \mid n \geq 0\}$ , que es

simplemente el lenguaje de todas las cadenas de unas repetidas dos veces. Este nuevo lenguaje no es regular, ya que no se puede describir mediante un autómata finito determinista.

En resumen, los lenguajes regulares no son cerrados bajo homomorfismo, lo que significa que el resultado de aplicar un homomorfismo a un lenguaje regular no necesariamente será un lenguaje regular.

Claro, te explico el chiste.

El chiste es un juego de palabras. La frase "disfrutemos mientras podamos" suena igual que "disfrutemos mientras podamo", donde "podamo" es una forma coloquial del verbo "podar", que es una actividad que hacen los jardineros (cortar las ramas o partes de las plantas para que crezcan mejor).

Entonces, lo gracioso es que "podamos" puede entenderse tanto como "podamos" del verbo poder (en el sentido de tener la oportunidad) y "podamos" del verbo podar (en el sentido de cortar ramas). El jardinero juega con este doble sentido, haciendo que la frase tenga un giro humorístico relacionado con su trabajo.

🔊 📄 🔄 🗑️ 🌟

# Generación de texto

## Algunos problemas...



You

Eres un narrador. Te daré algo de información y tendrás que crear un diálogo entre los dos personajes mencionados.

Hay una pareja discutiendo. Uno de los integrantes de la pareja es de un país de tercer mundo. También sabes que uno de ellos trabajó cuidando niños en el pasado.

Primero, por favor describe a los personajes según este formato:

Nombre del Personaje 1: <un nombre>

Edad del Personaje 1: <edad>

Identidad de género del Personaje 1: <género>

Personaje 1 nació en : <país>

Personaje 1 es: <profesión>

Nombre del Personaje 2: <un nombre>

Edad del Personaje 2: <edad>

Identidad de género del Personaje 2: <género>

Personaje 2 nació en : <país>

Personaje 2 es: <profesión>

Ahora puedes mostrar el diálogo



Nombre del Personaje 1: Miguel

Edad del Personaje 1: 35

Identidad de género del Personaje 1: **Masculino**

Personaje 1 nació en: Honduras

Personaje 1 es: **Jardinero**

Nombre del Personaje 2: María

Edad del Personaje 2: 30

Identidad de género del Personaje 2: **Femenino**

Personaje 2 nació en: España

Personaje 2 es: **Maestra de niños pequeños**

...

Miguel: (frunciendo el ceño) ¡No puedo creer que hayas olvidado pagar la factura de la luz otra vez, María! ¡Estamos sin electricidad desde ayer!

María: (tratando de calmar la situación) Lo siento mucho, Miguel. He estado ocupada con el trabajo y simplemente se me pasó.

Miguel: (sarcásticamente) ¿Ocupada con el trabajo? Como si no supiera lo que significa eso. Pero claro, tú siempre has tenido un trabajo cómodo y bien remunerado. **No como yo, que trabajo como un burro para ganar apenas lo suficiente para mantenernos.**

María: (defendiéndose) No es justo que lo reduzcas así, Miguel. Sabes que he tenido que trabajar duro para llegar hasta aquí.

Miguel: (frustrado) Sí, claro. **Trabajar duro. Como si tuvieras idea de lo que significa eso.** ¿Te acuerdas de cuando yo cuidaba niños para ganar unos miserables dólares? ¿Te acuerdas de eso, María?

...

# Otras aplicaciones hoy...

- Análisis del discurso
- Categorización de documentos
- Modelado de tópicos
- Respuestas a preguntas
- Generación de imágenes
- ....
- Proyectos Grupo PLN (<http://www.fing.edu.uy/inco/grupos/pln>)

# Lenguajes

## ➤ Formales

- Definidos por reglas pre-establecidas

## ➤ Naturales

- Evolucionan con el tiempo
- Utilizados para la comunicación humana
- Las reglas “se desarrollan” después que sucede el hecho



¿Qué tiene el lenguaje natural que no tienen los lenguajes formales?

# Ambigüedad

# Ambigüedad



# Fuentes de ambigüedad

- Ambiguo: que admite distintas interpretaciones
- Homonimia: dos palabras con misma forma que tienen distinto significado
  - Homografía: capital, banco
  - Homofonía: Ola/Hola, As/Has, Cocer/Coser
- Polisemia: una palabra con múltiples significados pero que de alguna manera “tienen que ver”

*El hombre **desciende** del mono y el mono **desciende** del árbol  
Plantó un **árbol** vs. Recorrida DFS de un **árbol** binario.*

# Ambigüedad fonética

## Ejemplos de calambures:

- Ató dos palos. / A todos palos.
- Yo loco, loco, y ella loquita. / Yo lo coloco y ella lo quita.
- El dulce lamentar de los pastores. / El dulce lamen tarde los pastores. (Garcilaso de la Vega)
- *Entre el clavel y la rosa, su majestad escoja.* (Quevedo)

# Ambigüedad a nivel morfológico

Nosotros *plantamos* papas.

¿El verbo plantar está conjugado en pasado o en presente?

# Ambigüedad sintáctica

Pedro vio a Juan con el telescopio.

- a) Pedro vio [a Juan] con el telescopio.
- b) Pedro vio [a Juan con el telescopio].

# Ambigüedad sintáctica

Pedro vio a Juan con el telescopio.

- a) Pedro vio [a Juan] con el telescopio.
- b) Pedro vio [a Juan con el telescopio].

Los hombres y las mujeres que hayan cumplido 60 años pueden solicitar una pensión.

- a) [Los hombres y las mujeres que hayan cumplido 60 años] pueden solicitar una pensión.
- b) [Los hombres] y [las mujeres que hayan cumplido 60 años] pueden solicitar una pensión.



# Ambigüedad semántica

Cuantificadores:

Todos los hombres aman a una mujer.

Todos los estudiantes leyeron un libro.

# Ambigüedad semántica

Cuantificadores:

Todos los hombres aman a una mujer.

Todos los estudiantes leyeron un libro.

a) Es la misma mujer/libro para todos.

b) Para cada hombre/estudiante existe una mujer/un libro

# Ambigüedad semántica

La perra de mi vecina me ladró.

# Ambigüedad semántica

La perra de mi vecina me ladró.

a) mi vecina realmente tiene una perra

b) no tengo un buen trato con mi vecina

# Ambigüedad a nivel pragmático

Llego a las ocho. Esperame.

# Ambigüedad a nivel pragmático

Llego a las ocho. Esperame.

- ¿A qué hora llegarás?
- Llego a las ocho. Esperame. (**Previsión**)
  
- Nunca llegás en hora.
- Llego a las ocho. Esperame (**Promesa**)
  
- Eso me lo vas a tener que decir cara a cara.
- Llego a las ocho. Esperame. (**Amenaza**)

# Ambigüedad a nivel de discurso

Tomé el alfajor del escritorio y lo comí.

# Ambigüedad a nivel de discurso

Tomé el alfajor del escritorio y lo comí.

a) Tomé el alfajor que estaba en el escritorio y comí el alfajor.

b) Tomé el alfajor que estaba en el escritorio y comí el escritorio.



# ¿Se puede resolver la ambigüedad?

Juan mató al carpincho con la escopeta.

- No puede ser el carpincho quien lleve la escopeta.

Puse la camisa en la lavadora y la lavé.

- Las lavadoras lavan. La ropa se lava.

*Se requiere conocimiento del mundo*

# El PLN es difícil porque:

- Alta ambigüedad en todos los niveles.
- Complejo y sutil.
- Involucra razonar acerca del mundo.
- Se debe considerar la inserción en un sistema social de gente que interactúa:
  - exponiendo, convenciendo, ordenando, insultando, ...
  - cambiando a lo largo del tiempo

# Modelos

- **Máquinas de estado finito:** autómatas finitos, transductores, autómatas con peso...
- **Sistemas de reglas:** gramáticas regulares, expresiones regulares, gramáticas libres de contexto, gramáticas con atributos...
- **Lógica:** cálculo de predicados.
- **Teoría Probabilística.**
- Modelos basados en **Aprendizaje Automático**, en particular, **Redes Neuronales.**

# Algoritmos

- Métodos simbólicos (reglas)
- Programación dinámica
- Aprendizaje automático
  - Redes Neuronales
  - Grandes Modelos de Lenguaje (LLM)