Gramáticas Formales para el Lenguaje Natural

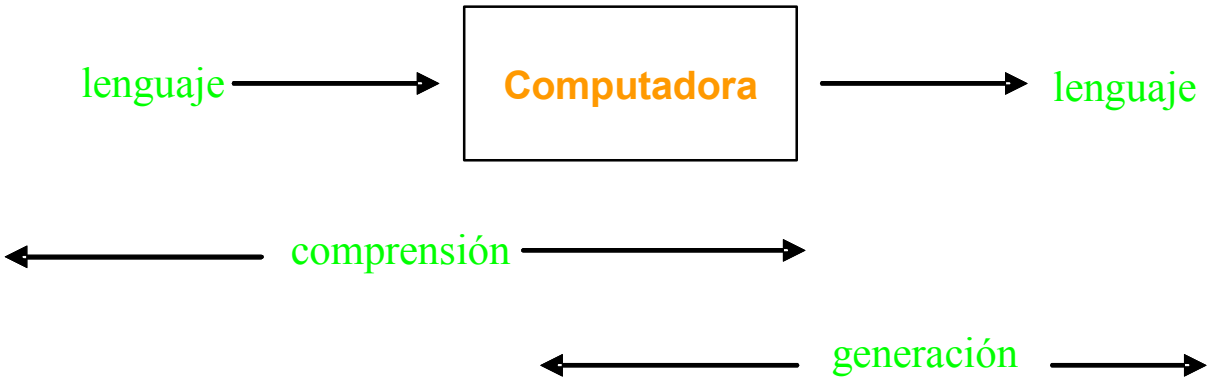
- Generalidades del curso
- Introducción al Procesamiento de Lenguaje Natural

Temario

- ¿Qué es el PLN?
- 6 niveles de procesamiento.
- Un poco de historia, éxitos y desafíos.
- Proyectos del grupo PLN del InCo.
- El curso GFLN.

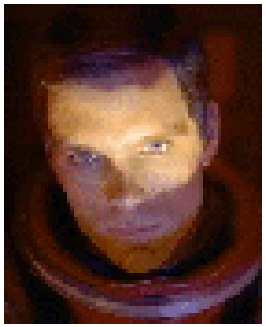
Temario

- ¿Qué es el PLN?
  - Conjunto de métodos y técnicas eficientes desde un punto de vista computacional para la **comprensión** y **generación** de lenguaje natural.
  - Subdisciplina de la IA.



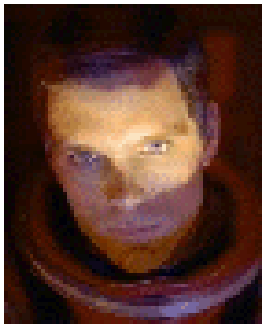
2001: Odisea del Espacio

Stanley Kubrick 1968



2001: Odisea del Espacio

Stanley Kubrick 1968



- *Dave: Open the pod bay doors, HAL.*
- *HAL: I'm sorry Dave. I'm affraid I can't do that.*

- Dave: Abre las compuertas, HAL.
- HAL: Lo siento, Dave. Me temo que no puedo hacerlo.

HAL - 2001, Odisea del Espacio

Habilidades de HAL (1967)

- comprensión de humanos vía:
  - reconocimiento del habla
  - comprensión de lenguaje natural
- comunicación con humanos vía:
  - generación de lenguaje natural
  - síntesis del habla
- pero también:
  - capacidades gráficas
  - juega al ajedrez
  - percepción visual

Habilidades de HAL

Señal sonora  Secuencia de palabras

Reconocer/Generar

Conocimientos de:

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

Habilidades de HAL

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.

Habilidades de HAL

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

- Por ejemplo: **(*)** *Lo puedo Dave siento que no temo me hacerlo.*
- Sin embargo: *Dave, lo siento. Que no puedo hacerlo, me temo.*

Conocimientos de **Sintaxis**:

estudio de la estructuración (orden y agrupamiento) de las palabras en unidades mayores.

Habilidades de HAL

- La sintaxis no es suficiente:
  - Abre las compuertas, HAL. (*Estructura: VC + ART + SUST + SP + SUST*)
  - Baja las persianas, HAL.
  - Saca los dados, HAL.
  - Suelta los perros, 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 significados mayores: **Semántica Composicional**

Habilidades de HAL

Adicionalmente, HAL presenta una utilización educada del lenguaje:

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

Significa, en realidad: (1) no lo siente y (2) puede abrir las compuertas

HAL podría haber respondido:

- *No.*
- *De ninguna manera.*

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.

6 niveles de procesamiento

- ***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.
- ***Discurso:*** *estudio de las unidades mayores a la oración.*
- ***Pragmática***: estudio de cómo el lenguaje se utiliza para cumplir objetivos.

Ambigüedad: el mayor problema en PLN

Ambiguo: que admite distintas interpretaciones.

Fuentes de ambigüedad

-
-

Ambigüedad en los niveles de análisis

Ambigüedad a nivel fonético

Homofonía

- ola / hola
- as / has / haz

Segmentación

- Ató dos palos. / A todos, palos.
- Entre el clavel y la rosa, su majestad escoja. (Quevedo)

Ambigüedad en los niveles de análisis

Ambigüedad a nivel morfológico

Nosotros plantamos papas.

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

Ambigüedad en los niveles de análisis

Ambigüedad a nivel sintáctico

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 en los niveles de análisis

Ambigüedad a nivel semántico

- *Homonimia: dos palabras con misma forma que tienen distintos significados*

  (distinta etimología, distintas entradas en el diccionario).
  - Homografía: *vino (bebida) / vino (llegó*)
  - Homofonía: o*la / hola, as / has / haz, cocer / coser*.

- *Polisemia: una palabra con múltiples significados (relacionados)*

  (una entrada en el diccionario con distintos significados).
  - *El hombre **desciende** del mono y el mono **desciende** del árbol.*
  - *Plantó un **árbol*** vs. *Recorrida DFS de un **árbol** binario.*

Ambigüedad en los niveles de análisis

Ambigüedad a nivel semántico

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 en los niveles de análisis

Ambigüedad a nivel de discurso

...

El hijo se va a jugar al billar, y en el momento en que va a tirar una carambola sencillísima, el otro jugador le dice:

-Te apuesto un peso a que no la haces.

Todos se ríen. Él se ríe. Tira la carambola y no la hace. Paga su peso y todos le preguntan qué pasó, si era una carambola sencilla.

...

Algo muy grave va a suceder en este pueblo.

Gabriel García Márquez

Ambigüedad en los niveles de análisis

Ambigüedad a nivel de discurso

...

El hijo se va a jugar al billar, y en el momento en que va a tirar una carambola sencillísima, el otro jugador le dice:

-Te apuesto un peso a que no la haces.

Todos se ríen. Él se ríe. Tira la carambola y no la hace. Paga su peso y todos le preguntan qué pasó, si era una carambola sencilla.

...

Algo muy grave va a suceder en este pueblo.

Gabriel García Márquez

Ambigüedad en los niveles de análisis

Ambigüedad a nivel de discurso

La Celeste se reencuentra con su gente un mes después de la goleada 4-0 sobre Paraguay en el Centenario, un reducto que se ha vuelto inexpugnable en estas Eliminatorias y que viene siendo el principal sostén de una notable campaña.

Uruguay tiene el ataque más efectivo (16 al igual que Brasil), la valla menos goleada (cinco tantos recibidos y ninguno en casa), el máximo anotador del torneo (Edinson Cavani con cinco) y el líder en asistencias (Carlos Sánchez con cuatro).

Esos números habrá que ratificarlos esta noche para romper otros. Y es que Venezuela se ha vuelto una piedra en el zapato de la Celeste, que no la derrota en Montevideo desde 2000. Además, el último duelo, en la Copa América Centenario, terminó con triunfo vinotinto 1-0.

El equipo de Tabárez, el único hasta el momento que ganó cinco partidos en el camino a Rusia 2018, saldrá a la cancha con Fernando Muslera; Mathias Corujo, Diego Godín, Sebastián Coates y Gastón Silva; Carlos Sánchez, Egidio Arévalo Ríos, Nicolás Lodeiro y Cristian Rodríguez; Luis Suárez y Edinson Cavani.

Ambigüedad en los niveles de análisis

Ambigüedad a nivel de discurso

La Celeste se reencuentra con su gente un mes después de la goleada 4-0 sobre Paraguay en el Centenario, un reducto que se ha vuelto inexpugnable en estas Eliminatorias y que viene siendo el principal sostén de una notable campaña.

Uruguay tiene el ataque más efectivo (16 al igual que Brasil), la valla menos goleada (cinco tantos recibidos y ninguno en casa), el máximo anotador del torneo (Edinson Cavani con cinco) y el líder en asistencias (Carlos Sánchez con cuatro).

Esos números habrá que ratificarlos esta noche para romper otros. Y es que Venezuela se ha vuelto una piedra en el zapato de la Celeste, que no la derrota en Montevideo desde 2000. Además, el último duelo, en la Copa América Centenario, terminó con triunfo vinotinto 1-0.

El equipo de Tabárez, el único hasta el momento que ganó cinco partidos en el camino a Rusia 2018, saldrá a la cancha con Fernando Muslera; Mathias Corujo, Diego Godín, Sebastián Coates y Gastón Silva; Carlos Sánchez, Egidio Arévalo Ríos, Nicolás Lodeiro y Cristian Rodríguez; Luis Suárez y Edinson Cavani.

Ambigüedad en los niveles de análisis

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**

¿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 procesamiento de lenguaje 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

Un poco de historia…

Breve historia : 50s, 60s

Primeras aplicaciones en computadoras menos poderosas que una calculadora

- Traducción Automática del Ruso al Inglés (Guerra Fría).
  - Famosa leyenda urbana:
    - (Original) "The spirit is willing, but the flesh is weak." (El espíritu es fuerte pero la carne es débil)
    - (Doble traducción) "The vodka is strong, but the meat is rotten." (El vodka está bueno pero la carne es muy mala)

- Trabajo fundacional en Autómatas, Lenguajes Formales, Probabilidades y Teoría de la Información

Breve historia : 70s, 80s

- Primer sistema de comprensión completa en un dominio limitado (Winograd, SHRDLU, 1971).
    - ¿La pirámide verde está sobre el cubo rojo?
- Separación de procesamiento (parsers) y descripción del conocimiento lingüístico.
- Explicitación de nivel de representación semántica.
- Se percibe necesidad de utilizar conocimiento sobre el mundo (proyecto CYC, Lenat).
- Traducción automática en dominios limitados (meteorología).

Breve historia : 90s

- Métodos de estado finito: gran eficiencia
  - Karttunen, Kaplan & Kay, FST
- La disponibilidad de grandes cantidades de texto (Web) reorienta el área.
- Primeros resultados robustos con métodos probabilísticos.
- Utilización de aprendizaje automático.

Breve historia : 2000s

- Enfásis en semántica y representación del conocimiento.
- Enfásis en discurso y diálogo.
- Integración de técnicas simbólicas y probabilísticas.
- Mayor integración de componentes LN en otros sistemas.
- Pero también : proliferación de aplicaciones “guiadas por patrones”, sin análisis profundo.

Algunas aplicaciones

Traducción Automática

Actualmente

- Original: *el día que las vacas vuelen*
- Doble Traducción (español-> inglés -> español) con Google
  - ***el día que las vacas lo vuelan*** (2008?)
  - ***las vacas día volar*** (2009)
  - ***el día que las vacas vuelen*** (2012)

  (traducción intermedia: *the day the cows come home* -> frase hecha)
  - ***el día que las vacas vuelvan a casa*** (2014) !!
  - ***el día que las vacas vuelan*** (2017)

Traducción Automática

- Cuestionamiento: con tasas de error elevadas, ¿es realmente útil la traducción automática?

- Ejercicio: interprete el siguiente texto en chino mandarín simplificado:

在加纳村惨剧后，暂停对黎南空袭48小时的以色列军队在8月1日恢复空袭，以色列内阁也通过决议扩大以军在黎巴嫩南部的地面攻势。同时，以色列开始大规模征召预备役人员。这一切表明，黎巴嫩南部的战火和硝烟在短期内难以平息。

Traducción Automática

- Cuestionamiento: con tasas de error tan elevadas, ¿es realmente útil la traducción automática?

- Ejercicio: interprete el siguiente texto en chino mandarín simplificado:

在加纳村惨剧后，暂停对黎南空袭48小时的以色列军队在8月1日恢复空袭，以色列内阁也通过决议扩大以军在黎巴嫩南部的地面攻势。同时，以色列开始大规模征召预备役人员。这一切表明，黎巴嫩南部的战火和硝烟在短期内难以平息。

(Traducción de Google) Ghana tragedy in the village, 48-hour suspension of air strikes against Lina in the **Israeli army** resumed air strikes on August 1, the Israeli cabinet passed a resolution to expand Israeli ground offensive in **southern Lebanon**. At the same time, Israel began a large-scale recruitment of reservists. All this shows that the fighting in southern Lebanon and smoke in the short term it is difficult to quell.

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’.

Extracción de Información

Texto Original

Restaurante Español cerca de Manchester en Inglaterra, busca camareros o camareras de salad con conocimiento de cocktelería y barra, deben saber flambear y tener un mínimo de tres años de experiencia con un manejo de Inglés a nivel medio, conocimientos de vinos Españoles y resto del mundo una ventaja. Salario mínimo 1500 euros mes con propinas. Cinco días por semanas de unas 50/55 horas.

→

Ficha

Industria: Restauración.

Puesto: Camarero/a.

Lugar: Manchester, Inglaterra.

Compañía: Restaurante Español

Salario: 1500 euros/mes.

Dedicación: 50/55 hs. Semanales.

Extracción de Información

- Objetivo: mapear una colección de documentos a una base de datos estructurados.

- Motivaciones:

  – Permitir búsquedas complejas: quiero trabajos en restauración en Manchester que paguen por lo menos 1200 euros al mes.
  – Permitir consultas estadísticas: ¿el número de trabajos en restauración creció en los últimos cinco años?

Interfaces a BD

- **Usuario**: Necesito un tren nocturno de París a Viena que llegue alrededor de las 10 de la mañana.
- **Sistema**: ¿Qué día desea viajar?
- **Usuario**: Mañana.
- **Sistema**: Los trenes disponibles son…

- Análisis de la entrada y "traducción" a una consulta.
  - P.ej: $\exists x(\text{tren}(x) \wedge \text{nocturno}(x) \wedge \text{recorrido}(x, \text{París}, \text{Viena}) \wedge \exists y \exists z(\text{horario}(x, y, z) \wedge \text{alrededor}(z, 10)))$
- El enfoque funciona bien con léxico y sintaxis restringidos.

Más aplicaciones

- Recuperación de información
- Verificadores de gramática y estilo
- Categorización de documentos
- Respuesta a preguntas
- Implicación / paráfrasis textual
- Determinación de plagio
- Determinación de autoría
- Análisis de redes sociales (Facebook, Tweeter, foros en general): opiniones y sentimientos
- ...

Grupo PLN – InCo - UDELAR

- Análisis sintáctico
    - Segmentación de oraciones en proposiciones
    - Desambiguación de comas
- Reconocimiento de eventos
    - ¿Cuáles son los eventos a los que se hace referencia en un texto?
    - ¿Ocurrieron efectivamente?
- Análisis temporal de textos

  Ubicación temporal y ordenamiento de los eventos mencionados.
- Opiniones

  ¿Quién opinó sobre el tema X? ¿Qué dijo? ¿Opinó a favor o en contra?
- BIO-NLP
- Recursos y herramientas (Proyecto RITA)
- Proyectos de grado: generación de crucigramas, clasificación de humor, análisis de tweets, …
- Actualmente: representaciones distribuidas y redes neuronales

Algunas herramientas y recursos

- FreeLing: etiquetador morfo-sintáctico, analizador sintáctico, distribución libre (Universitat Politècnica de Catalunya)
- Clatex: segmentador en proposiciones (PLN-InCo)
- Editor de reglas contextuales: (PLN-InCo)
- Anotación de textos: Clark, Knowtator-Protégé, MMAX2
- NLTK: kit de herramientas de PLN para Python
- Spanish WordNet: (Universitat Politècnica de Catalunya, licencia gratuita para usos académicos)
- Corpus: Corin (Lingüística), CREA (RAE), “Corpus del Español”, Temantex (PLN-InCo), Opiniones (PLN-InCo), Ancora (Grial), Corpus de prensa uruguaya (PLN-InCo), Wikipedia en español
- Maltparser: Analizador de dependencias español (IULA-Pompeu Fabra)
- Representaciones vectoriales: Diferentes conjuntos de vectores de palabras para el español generados en el marco del grupo PLN-InCo

Gramáticas Formales para el Lenguaje Natural

1. Introducción al Procesamiento de Lenguaje Natural
2. Generalidades del curso

Objetivos del curso

- Conocer formalismos gramaticales de uso actual en PLN, sus fundamentos teóricos, su poder expresivo, los métodos de análisis sintáctico (*parsing*).
- Experimentar con diversas herramientas y datos (*corpus*) disponibles.
- Profundizar en algún aspecto de gramáticas formales y tecnologías de *parsing.*

Cronograma tentativo

Semanas	Temas
1	Introducción
1, 2	Elementos de la gramática del español
3, 4	**GLC** (Gramáticas Libres de contexto)
5, 7	GLC a partir de corpus, GLCP, *parsing* probabilístico
7, 8, 10	X barra, TFS, **HPSG** (Gramáticas Sintagmáticas Guiadas por el Núcleo)
10, 11, 12	**Gramáticas de Dependencias**, *parsing* basado en dependencias.
13, 14	**Gramáticas Combinatorias Categoriales**, *parsing* basado en gramáticas categoriales.
17	¿Prueba escrita?

Nota: La semana 6 es Turismo y la semana 9 no hay clases por parciales.

Modalidad y evaluación

- **Materiales en EVA**: presentaciones utilizadas en clase, clases filmadas (curso 2016), bibliogafía complementaria, prácticos.

- **Foro de discusión en EVA**

- **Evaluación:**
  - 4 entregas obligatorias y en grupo, relativas a los conceptos básicos del curso y de familiarización con recursos y herramientas.
  - **Primera entrega sobre gramática del español tercera semana.**
  - Mini-proyecto (solo en modalidad posgrado).
  - Prueba escrita individual.

Docentes

- Aiala Rosá (responsable del curso)
- Mathias Etcheverry (responsable del laboratorio)

Retomamos: 6 niveles de análisis

- ***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.

6 niveles

- ***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.

Sintaxis

Pedró comió arroz con tenedor.

Pedro comió arroz con queso.

Diferentes estructuras, la sintaxis (computacional) debe:

Sintaxis

Pedró comió [arroz] [con tenedor].

Pedro comió [arroz [con queso]].

Diferentes estructuras, la sintaxis (computacional) debe:

1- especificar que ambas opciones son posibles,

2- ir acompañada por algoritmos que permitan obtener cualquiera de los 2 análisis.

Análisis sintáctico (*parsing*)

- Debe dar la estructura adecuada en cada caso.
- Actualmente se aplican métodos probabilísticos para desambiguar.
- La sintaxis es un paso intermedio para llegar a la semántica.
- Es imprescindible en muchas aplicaciones:
  - Preguntas y respuestas, inferencia

    Juan prometió ir al acto / convenció a María de ir al acto.

    ¿Quién fue al acto?
  - Traducción, ya que muchas veces el resultado es agramatical.

Gramáticas formales

- Formalismos computacionales.
- Describen de algún modo la estructura de las oraciones.
- Tienen asociado mecanismos de cálculo de la estructura a partir de una entrada (en general, una oración). Estos mecanismos son llamados *parsers*, realizan *parsing.*
- El estado del arte actual se basa en el *parsing* probabilístico.