

ELiC11: Transductores

Clase 1: Introducción

Carlos Areces

carlos.areces@gmail.com

La clase de hoy

- Qué es LC (Linguística Computacional)?
- Qué veremos en este curso?
- Qué son los transductores?
- Lenguajes Regulares
 - Expresiones Regulares
 - Autómatas de Estados Finitos



Qué es LC?



Deckard: Enhance 224 to 176. Enhance, stop. Move in, stop. Pull out, track right, stop. Center in, pull back. Stop. Track 45 right. Stop. Center and stop. Enhance 34 to 36. Pan right and pull back. Stop. Enhance 34 to 46. Pull back. Wait a minute, go right, stop. Enhance 57 to 19. Track 45 left. Stop. Enhance 15 to 23. Give me a hard copy right there.

—*Blade Runner*, 1982

***Por qué la LC
es difícil?***

“I made her duck”

Qué significa esta oración?

- “duck”: sustantivo o verbo?
- “make”: “cocinar X”?, o “hacer que X haga Y”?
- “her”: “para ella” o “a ella”?

Otros ejemplos de ambigüedad:

- “I eat sushi with tuna” o “I eat sushi with chopsticks”
- “I saw the man with the telescope on the hill”
- “John saw Jim. He was drinking coffee.”

Y “I made her duck cassoulet” ?

- Cuántas palabras en inglés tiene la oración?

***Pero qué es LC
realmente?***

Un psicoanalista?

Computer: TELL ME MORE ABOUT YOUR FAMILY

Human: My mother takes care of me.

Computer: WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

Human: My father.

Computer: YOUR FATHER

Human: You are like my father in some ways.

Computer: WHAT RESEMBLANCE DO YOU SEE

Human: You are not very aggressive but I think you don't want me to notice that.

Computer: WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE


Human: You don't argue with me.

Computer: WHY DO YOU THINK I DON'T ARGUE WITH YOU

Human: You are afraid of me.

Weizenbaum (1966), ELIZA

LC en el mercado

 **CoGenTex, Inc.**

Active Sales Assistant™ personalized product recommendations from smart virtual sales assistants.

SHOPPERS

These virtual sales assistants give you the best product recommendations based on your preferences, for free.

You get: Recommendations ranked from best fit to worst, plus prices from leading retailers.

1

Toshiba SD-275

2

Onkyo DV-S555


3

Sony DVP-F21

BUSINESSES

Increase sales on **your site** with Active Sales Assistant! Our clients typically **double their sales conversion rates**.



Free report
the **top 5 secrets** to great online selling

**These Digital Cameras best suit your needs...**

<<refine your search if you wish, or start a new search

compare

* red means you didn't want that feature but the product may still be a very good fit otherwise

	Rank	Brand & Model	Avg. Street Price	Optical Zoom	Resolution
<input type="checkbox"/>	1 why?	 Toshiba PDR-M25 more info	\$240.00 WHERE TO BUY	3X	1792 x 1200 pixels
<input type="checkbox"/>	2 why?			3X	1280 x 960 pixels
<input type="checkbox"/>	3 why?	 1400 more info	\$249.00 WHERE TO BUY	3X	1280 x 960 pixels

The PDR-M25 gets our #1 ranking, based on your needs. Although it does not have an MP3 player, it has some of the best scores for price and resolution of the five top-ranked models, and has an acceptable rating for optical zoom.

LC en investigación

Text: *More than a decade ago, Carl Lewis stood on the threshold of what was to become the greatest athletics career in history. He had just broken two of the legendary Jesse Owens' college records, but never believed he would become a corporate icon, the focus of hundreds of millions of dollars in advertising. His sport was still nominally amateur. Eighteen Olympic and World Championship gold medals and 21 world records later, Lewis has become the richest man in the history of track and field -- a multi-millionaire.*

Hypothesis: *Carl Lewis won eighteen gold medals and set 21 world records.*
(Source: The 3rd Recognizing Textual Entailment Challenge (2006/7))

El texto implica la hipótesis? Si/No

Esta es un área de investigación muy activa

Traducción automática

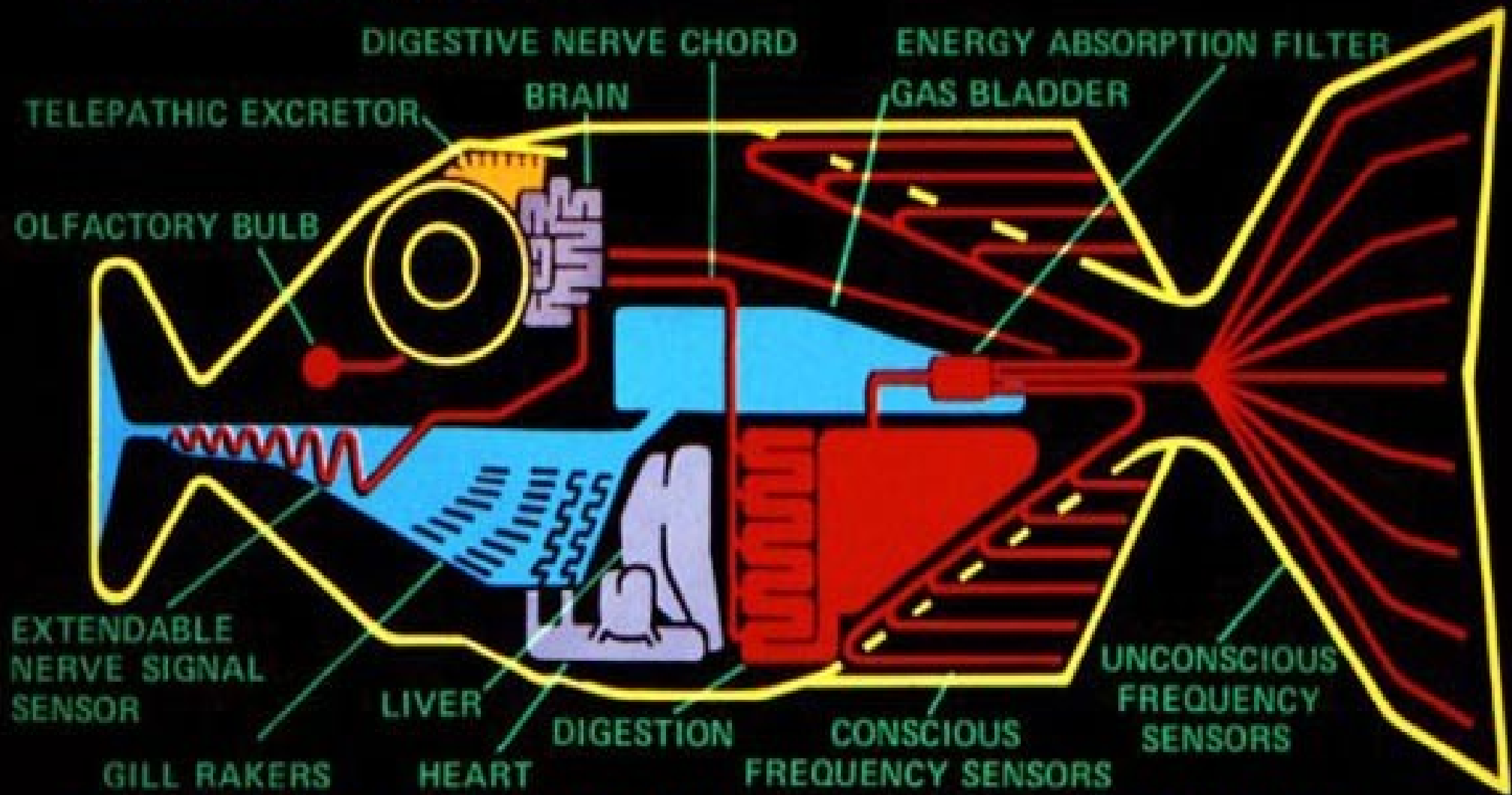


Quién no ha usado google translate?



***Qué hay detrás
de LC?***

BABEL FISH



THE BABEL FISH IS SMALL, YELLOW, LEECHLIKE, AND PROBABLY THE ODDEST THING IN THE UNIVERSE. IT FEEDS ON BRAIN WAVE ENERGY, ABSORBING ALL

***Qué hay detrás
de LC
realmente?***

Areas de lingüística

- ***Cómo funciona el lenguaje?***
(lingüística fundamental)
- ***Cómo enseñar/traducir/etc lenguajes?***
(lingüística aplicada)
- ***Cómo aprendemos y procesamos lenguajes?***
(psico-lingüística)
- ***En qué lugar del cerebro procesamos lenguajes?***
(neuro-lingüística)
- ***Cómo cambian los lenguajes en el tiempo?***
(lingüística histórica)
- ***Cómo expresa el lenguaje el status social?***
(socio-lingüística)

Niveles de lingüística

- **Fonética:** *Qué sonidos usamos?*
- **Fonología:** *Como combinamos sonidos?*
- **Morfología:** *Cómo formamos palabras?*
- **Sintaxis:** *Cómo formamos oraciones?*
- **Semántica:** *Cómo transmitimos significados?*
- **Pragmática:** *Para qué usamos el lenguaje?*

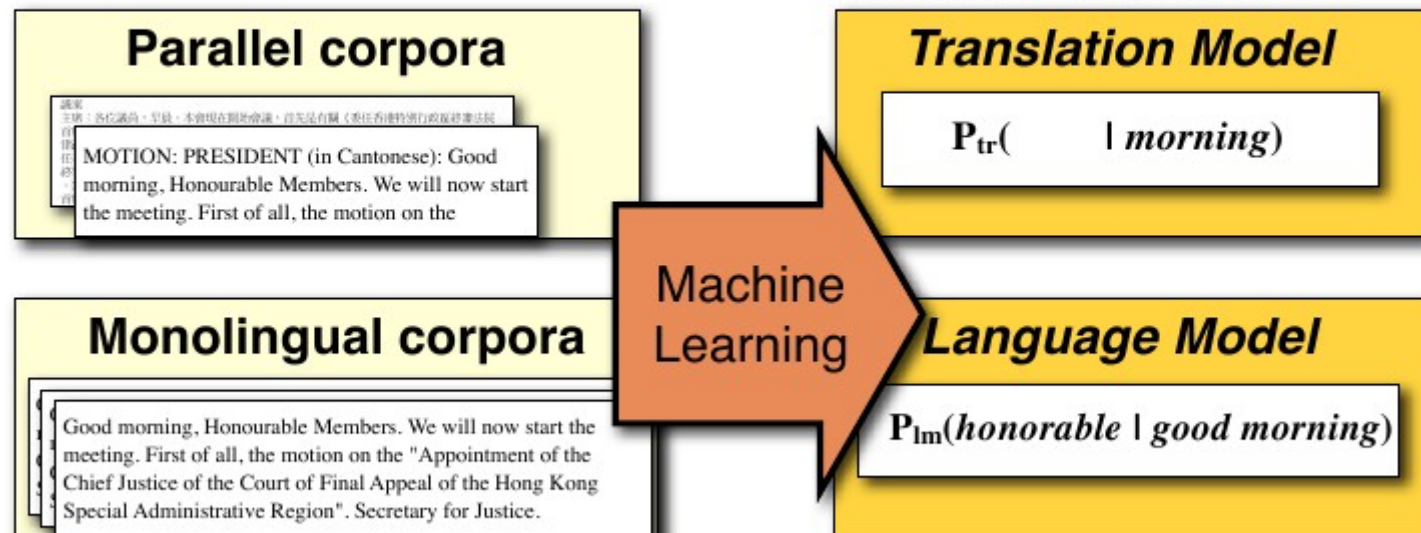
Tareas de lingüística

- **Reconocedor y Sintetizador de Speech**
para procesar señales de audio y generar speech
- **Tokenizer/Segmentador**
para dividir las señales de audio en palabras
- **Parser Sintáctico/Semántico + Gramática**
para inferir la estructura/significado de la oración
- **Generación de Lenguaje Natural**
para formular respuestas
- **Modelo del Discurso**
para hacer contribuciones coherentes

Mucho aprendizaje automático

The image shows a screenshot of the Google Translate interface. On the left, there is a Chinese article titled "穿越灾难 迎接光荣——写在四川汶川大地震抗震救灾百日之际" (Through disaster to meet glory - written on the 100th day of the earthquake relief efforts in Sichuan Wenchuan). The article discusses the relief efforts and the resilience of the Chinese people. On the right, the English translation is shown, titled "Through the disaster to meet the glorious - The earthquake in four Chuan Wenchuan 100 days of the earthquake relief". A large red arrow with the text "Google Translate" points from the Chinese text to the English text.

$$\operatorname{argmax}_{Eng} P(Eng|Chin) = \operatorname{argmax}_{Eng} \underbrace{P(Chin|Eng)}_{\text{Translation Model}} \times \underbrace{P(Eng)}_{\text{LanguageModel}}$$



Muchas sustituciones

***Human:** You don't argue with me.*

***Computer:** WHY DO YOU THINK I DON'T ARGUE WITH YOU*

- 1.** Replace **you** with **I** and **me** with **you**:
I don't argue with you.
 - 2.** Replace **X** with Why do you think **X**:
Why do you think I don't argue with you.
- Los patrones que se usan para sustituir en ELIZA, pueden ser expresados con expresiones regulares
 - Las expresiones regulares son equivalentes a los autómatas finitos

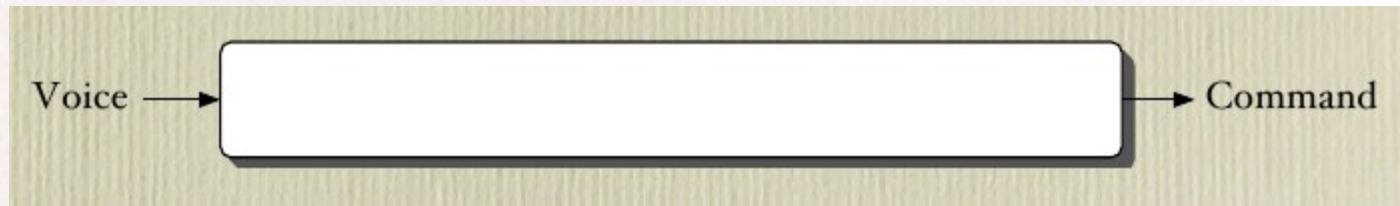
CL y teoría de autómatas

Qué clase de autómata/gramática hace falta para analizar lenguaje natural?

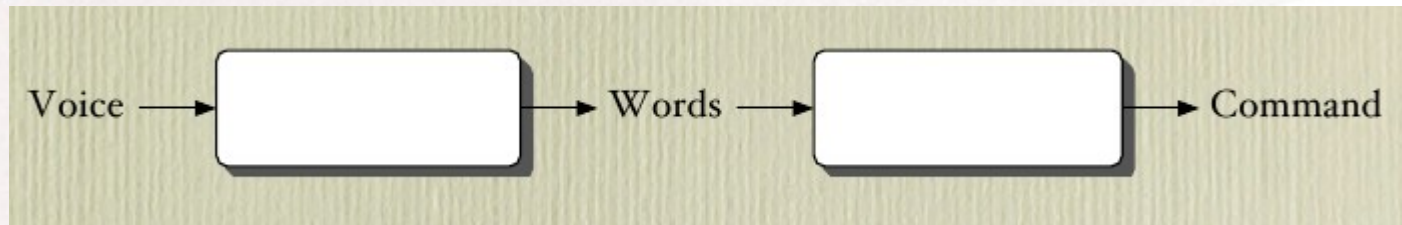
En qué clase de lenguaje clasificamos a los lenguajes naturales?

La jerarquía de lenguajes formales de Chomsky (1956) fue desarrollada originalmente para contestar estas preguntas.

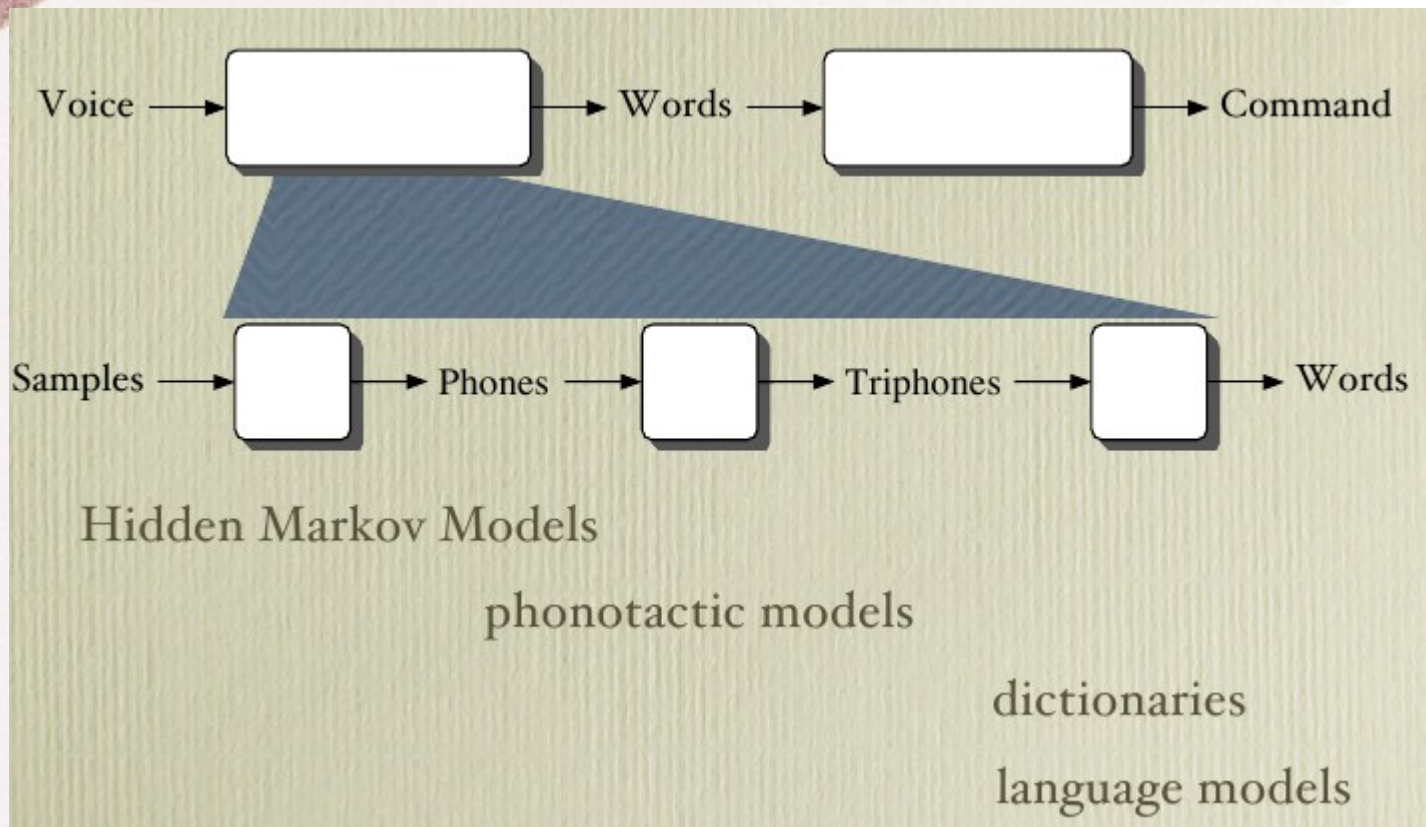
Blade Runner: De voz a comandos



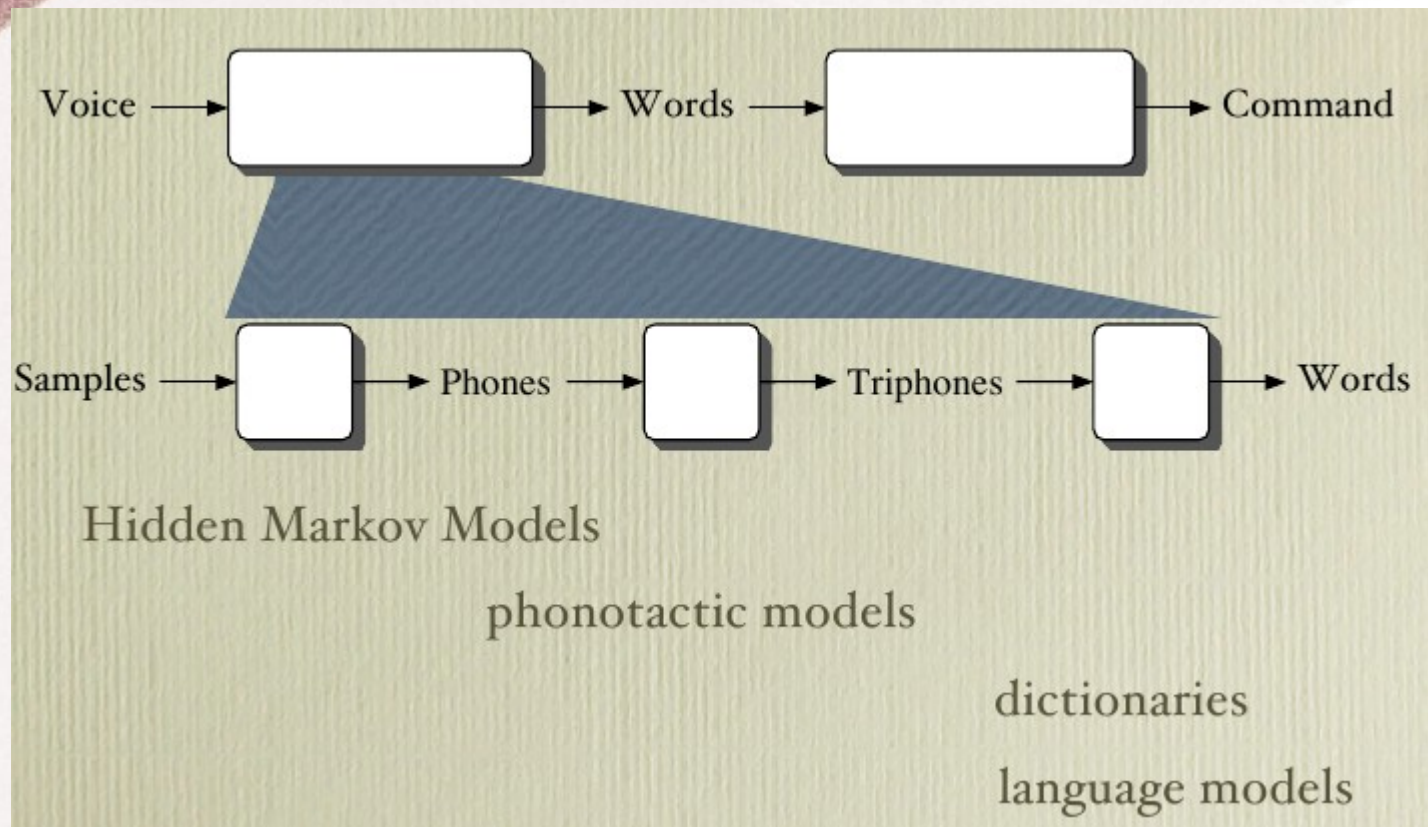
Blade Runner: De voz a comandos



Blade Runner: De voz a comandos

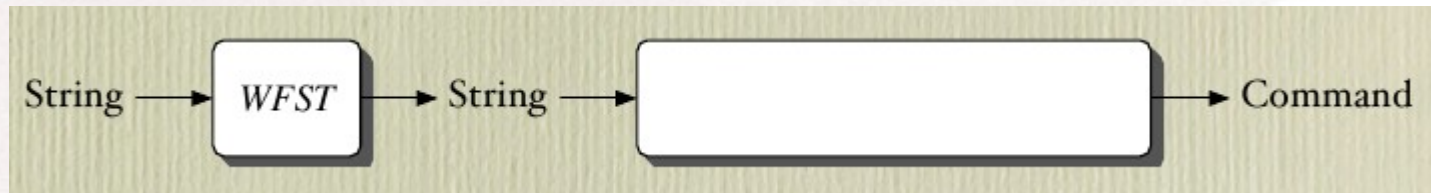


Blade Runner: De voz a comandos

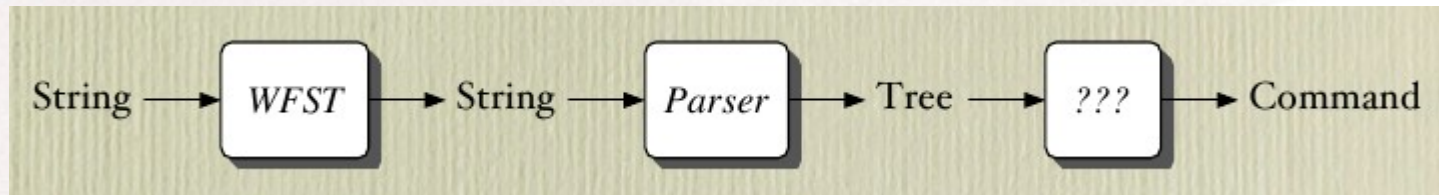


Weighted Finite State Transducers (WFST)

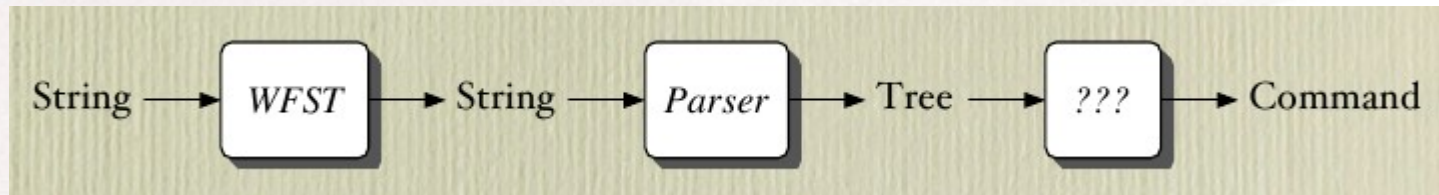
Blade Runner: De voz a comandos



Blade Runner: De voz a comandos



Blade Runner: De voz a comandos



Cadenas son a **WFST**
lo que
Árboles son a **???**

Plan del curso

- Autómatas de estados finitos
- Transductores de estados finitos

Plan del curso

- Autómatas de estados finitos
 - Lenguajes regulares
 - Autómatas
 - Construcción de Thompson
 - Operaciones
 - Left/right reversal
 - Granularidad
 - Eliminación de ϵ
 - Determinización
 - Minimización
- Transductores de estados finitos

Plan del curso

- Autómatas de estados finitos
- Transductores de estados finitos
 - Relaciones regulares
 - Operaciones
 - Composición
 - Inversión
 - Aplicaciones
 - Pesos
 - Transductores de árboles



Lenguajes regulares y autómatas de estados finitos

Lenguajes regulares

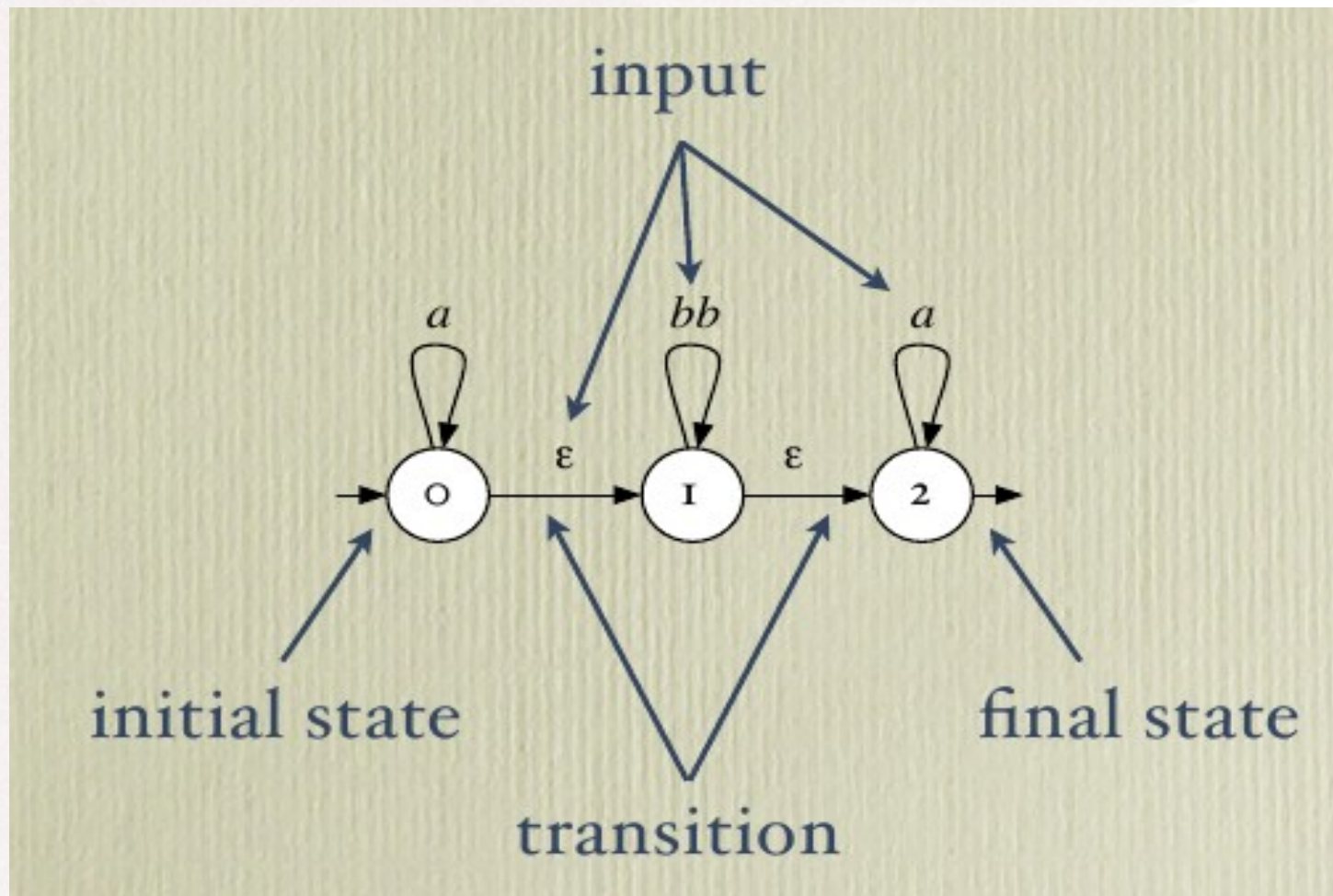
- Un lenguaje es un conjunto de cadenas
- Los **lenguajes regulares** son el menor conjunto de lenguajes que incluye
 - El lenguaje vacío
 - Los lenguajes unitarios
 - Y están cerrados bajo
 - Unión ($L_1 \cup L_2$)
 - Concatenación ($L_1 \cdot L_2$)
 - Iteración (L^*)

Expresiones regulares

- Notación para expresiones regulares
 - Cadena vacía: ϵ
 - Lenguajes unitarios: a
 - Unión: $x|y$
 - Concatenación: xy
 - Iteración: x^*
- Ejemplo (alfabeto = $\{a, \dots, z, \square\}$)

$$((be \mid it \mid let)^\square)^* \ni let^\square it^\square be$$

Automatas finitos



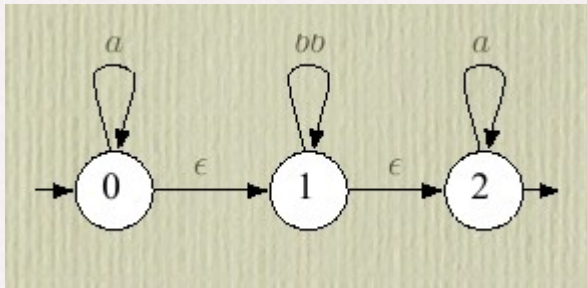
Autómatas de estados finitos

- Un autómata de estados finitos (FSA) es una tupla $\langle Q, \Sigma, \Delta, q_0 \rangle$ donde
 - Q es un conjunto finito de estados
 - Σ es un conjunto finito de **símbolos terminales**, (que no incluye el símbolo especial #).
 - q_0 es el **estado inicial**.
 - Δ es la **función de transición**.

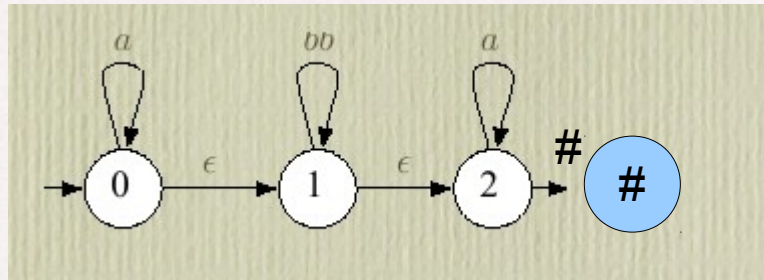
La función de transición

- Sea Σ^* el conjunto de todas las cadenas sobre Σ ; la notación $\Sigma^\#$ representa el conjunto de cadenas de Σ^* con un símbolo $\#$ opcional al final.
- $\Delta \in Q \times \Sigma^\# \times (Q \cup \{\#\})$ es un conjunto de transiciones, cada una linkeando un estado **fuente** (de Q) y un estado **destino** (de $Q \cup \{\#\}$) mediante una cadena (quizás vacía) de $\Sigma^\#$, y tal que el estado destino es $\#$ si la cadena termina con $\#$
- Estados con transiciones a $\#$ se llaman **estados finales**.

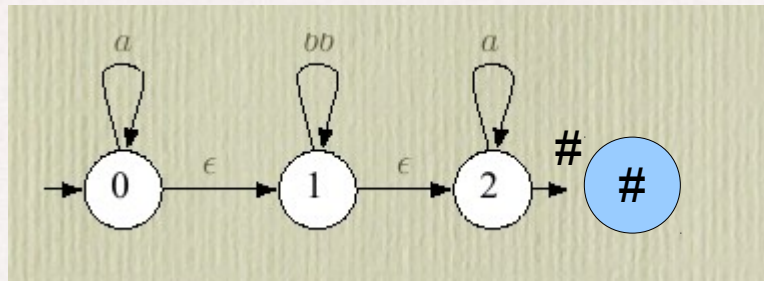
Derivaciones y cadenas aceptadas



Derivaciones y cadenas aceptadas



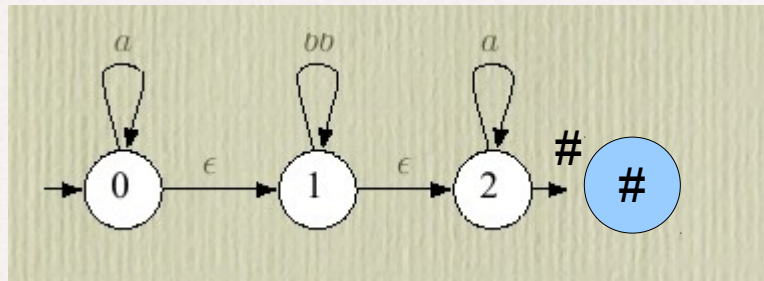
Derivaciones y cadenas aceptadas



Función de transición

$q_0 a \vdash q_0$	$q_0 \vdash q_1$
$q_1 bb \vdash q_1$	$q_1 \vdash q_2$
$q_2 a \vdash q_2$	$q_2 \# \vdash \#$

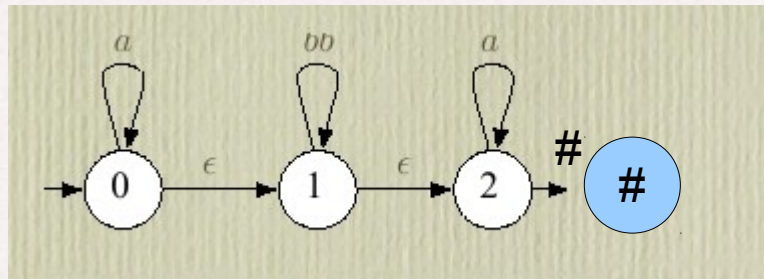
Derivaciones y cadenas aceptadas



Reglas de reescritura

$q_0 a \vdash q_0$	$q_0 \vdash q_1$
$q_1 bb \vdash q_1$	$q_1 \vdash q_2$
$q_2 a \vdash q_2$	$q_2 \# \vdash \#$

Derivaciones y cadenas aceptadas



Reglas de reescritura

$q_0 a \vdash q_0$	$q_0 \vdash q_1$
$q_1 bb \vdash q_1$	$q_1 \vdash q_2$
$q_2 a \vdash q_2$	$q_2 \# \vdash \#$

Una derivación:

$$\begin{aligned}
 & q_0 aaabba\# \vdash q_0 aabba\# \vdash q_0 abba\# \vdash \\
 & \vdash q_0 abba\# \vdash q_1 bba\# \vdash q_1 a\# \\
 & \vdash q_2 a\# \vdash q_2 \# \vdash \#
 \end{aligned}$$

Def. de aceptación: aceptar w sii $q_0 w\# \vdash^* \#$

De expresiones regulares a automatas

