PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Introducción a la lingüística computacional

César Antonio Aguilar

Facultad de Lenguas y Letras

03/08/2017

Cesar.Aguilar72@gmail.com

¿Qué es la lingüística computacional ? (1)

EL INGENIOSO
HIDALGO DON QVI-
XOTE DE LA MANCHA
Compuesto por Miguel de Ceruantes Saauedra.
DIRIGIDO AL DVQVE DE BEIAR, Marques de Gibraleon, Conde de Barcelona, y Banares, Vizconde de la Puebla de Alcozer, Señor de las villas de Capilla, Curiel, y Burgillos.

Año, 1605.

Con priuilegio de Castilla, Aragon, y Portugal.
EN MADRID, Por Iuan de la Cuesta.
Vendese en casa de Francisco de Robles, librero del Rey nro señor.

```latex
\documentclass[12pt]{article}
\title{\LaTeX}
\date{}
\begin{document}
  \maketitle \LaTeX{} is a document preparation system for the \TeX{}
  typesetting program. It offers programmable desktop publishing
  features and extensive facilities for automating most aspects of
  typesetting and desktop publishing, including numbering and
  cross-referencing, tables and figures, page layout, bibliographies,
  and much more. \LaTeX{} was originally written in 1984 by Leslie
  Lamport and has become the dominant method for using \TeX; few
  people write in plain \TeX{} anymore. The current version is
  \LaTeXe.
  \newline
  % This is a comment, it is not shown in the final output.
  % The following shows a little of the typesetting power of LaTeX
  \begin{eqnarray}
    E &=& mc^2                                  \\
    m &=& \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}
  \end{eqnarray}
\end{document}
```

LaTeX

LaTeX is a document preparation system for the TeX typesetting program. It offers programmable desktop publishing features and extensive facilities for automating most aspects of typesetting and desktop publishing, including numbering and cross-referencing, tables and figures, page layout, bibliographies, and much more. LaTeX was originally written in 1984 by Leslie Lamport and has become the dominant method for using TeX; few people write in plain TeX anymore. The current version is LaTeX 2_ε.

$$E = mc^2 \quad (1)$$

$$m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}} \quad (2)$$

¿Qué es la lingüística computacional ? (2)

<CD num= “1"> <PP tippp= "Aut"> M. Godron y G. Merriam </PP> entre otros, quienes <PVD><VD lema= "considerar" args= "3" mdo= "fin"> consideran </VD> a la <TERM term= "L" tipterm= "fn"> Ecología del Paisaje </TERM> <NX tipnx= "adv"> como </NX> </PVD> <MTD mdef= "dp" mt= ""> : </MTD> <DEF tipdef= "GD"> "la ecología de los sistemas movibles y heterogéneos, estudiando entonces la influencia de la estructura del paisaje sobre los procesos ecológicos, tanto a escala local como regional " </DEF> </CD>

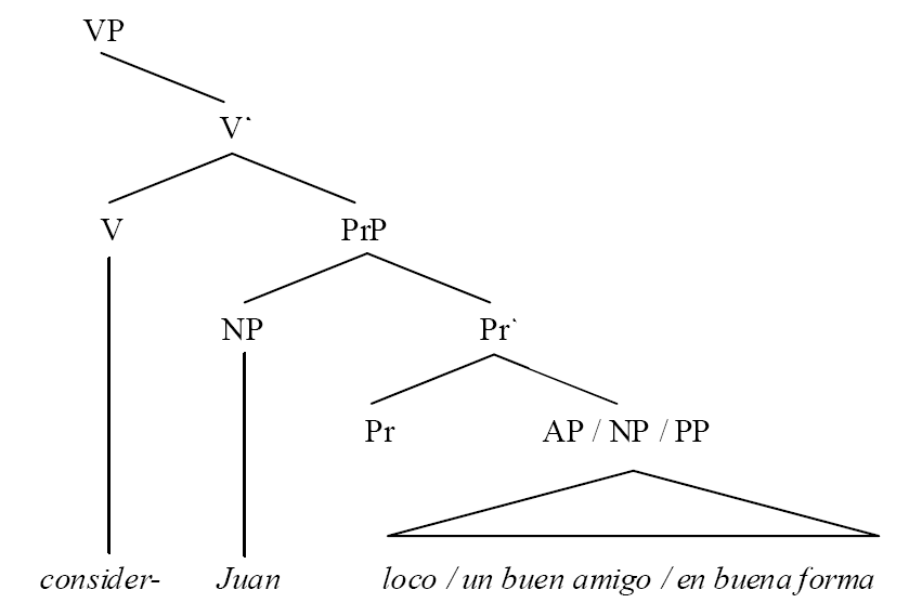
BRITISH NATIONAL CORPUS

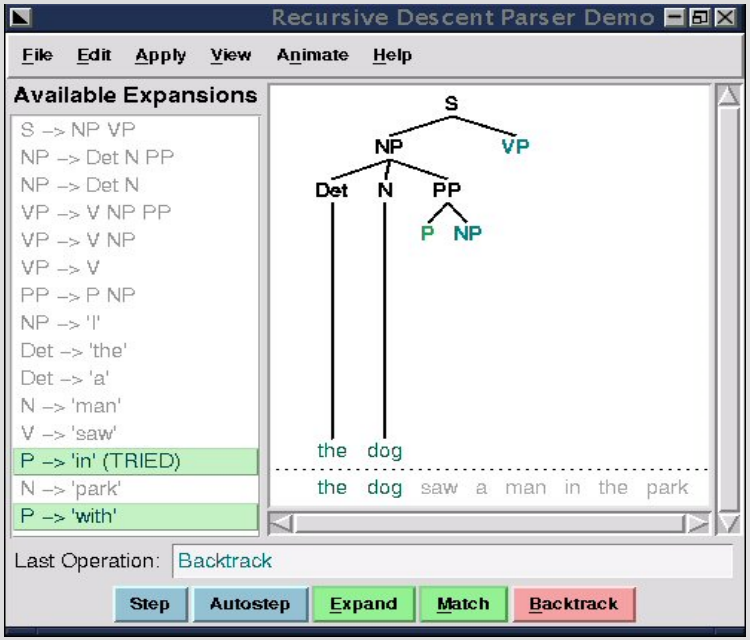
The Brown Corpus

REAL ACADEMIA ESPAÑOLA

CHEM Corpus Histórico del Español de México

¿Qué es la lingüística computacional ? (3)





Natural Language Toolkit

El test de Turing (1)

El que más ahondó en esta cuestión, e incluso logro crear la primera máquina inteligente de la historia moderna, fue un matemático británico llamado **Alan Turing**. Desarrolló su propuesta en un artículo titulado “Computer Machinery and Intelligence”.

En este artículo plantea que si una máquina es capaz de comprender y generar lenguaje humano, entonces se puede considerar que se trata de una máquina inteligente.

Ahora, el problema no es tanto diseñarla (según Turing), sino *probar que realmente este conocimiento es una prueba de cierta inteligencia*.

Alan Turing
(1912-1954)

El test de Turing (2)

Antes de meterse al problema de ingeniería que supone crear un computador avanzado, Turing diseñó primero un **test** o **prueba** que trata de determinar qué tanto una máquina puede hacernos pensar que tiene cierto grado de inteligencia. Este test se ha popularizado a través de muchas películas de ciencia ficción, como *Una Odisea espacial 2001*, *Blade Runner, Her* y más recientemente, *The Imitation Game*:

El test de Turing (3)

Mucha gente ha tratado de generar sistemas capaces de interactuar con humanos de una manera *natural*, es decir, situándose en un contexto de diálogo que sea familiar para las personas.

Uno de los mejores intentos ha sido el programa **Eliza**, desarrollado por **Joseph Weizenbaum** en el Laboratorio de Inteligencia Artificial del MIT (USA).

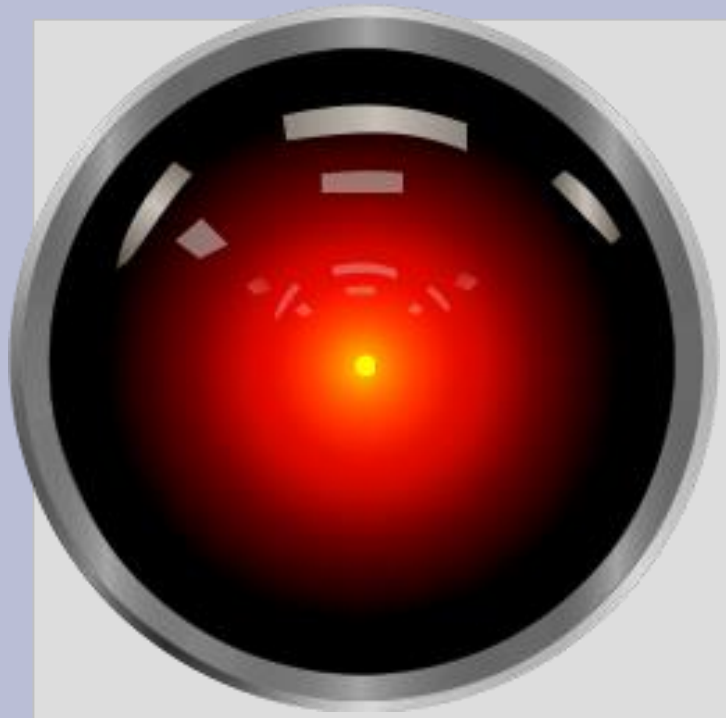
```
==============================================================================
          EEEEEEEE  L        IIIIIII  ZZZZZZZ      AAA
          E         L           I          Z      A   A
          E         L           I         Z      A     A
          EEEEE     L           I        Z       A     A
          E         L           I       Z        AAAAAAA
          E         L           I      Z         A     A
          EEEEEEEE  LLLLLLLL  IIIIIII  ZZZZZZZ    A     A

==============================================================================
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...

==============================================================================
```

http://elizagen.org/index.html

El Dr. Chandra platica con Sal, la hermana de Hal



Uno de los ejemplos más creíbles que podemos ver en el cine sobre lo que es la interacción de tipo humano-máquina lo podemos ver en una película hecha por Stanley Kubrick en 1968: *2001: A Space Odyssey*. En ella, vemos como el Dr. Sivasubramanian Chandrasegarampillai (o el Dr. Chandra) desarrolla en computador más avanzado de su tiempo, **Hall 9000**, capaz de realizar varias tareas complejas, entre ellas dialogar con astronautas.

Nueve años después, Hall tiene una hermana, **Sal 9000**, la cual en esencia tiene los mismos atributos que su hermano mayor. Un ejemplo de su funcionamiento lo pueden ver en la liga que anoté en la página del curso.

¿Se acuerdan de Threepio?

C-3PO Relaciones Interhumanas, también **C-3PO Relaciones Cibernético-humanas**, comúnmente **C-3PO**.

Se trata de un droide de protocolo, diseñado para interactuar con seres pensantes. Puede hablar y entender más de seis millones de formas de comunicación (¿idiomas?) .

Su principal función como droide de protocolo es ayudar a explicar las costumbres, de otras culturas, la traducción, etc. Es muy leal a sus amos. Desde su creación, ha servido a más de cuarenta.

Mayor información, ver:

http://starwars.wikia.com/wiki/C-3PO

La revolución de *Facebook* (1)

En una escena de la película *Ex Machina* (2015), los dos personajes principales (Caleb y Nathan) dialogan respecto a la tecnología necesaria para crear un autómata humanoide tan avanzado como EVE (el personaje principal).

Claro está, es necesario contar con un software y hardware que tengan una consistencia y nivel de operatividad cercano al cerebro humano:

La revolución de *Facebook* (2)

Empero, la creación de dichos software y hardware no son suficientes, pues hay que enseñarle a EVE a **actuar como si fuese humana**, más todavía si pretendemos que sea capaz de interactuar con nosotros en una conversación creíble. La pregunta es: ¿de dónde sacamos datos fiables que nos sirvan para ayudarle en su proceso de aprendizaje?

La respuesta es asombrosamente sencilla: estos datos ya los tenemos disponibles, y relativamente es muy fácil acceder a ellos usando… ¡**Facebook**!:

La revolución de *Facebook* (3)

Facebook, Twitter y otras redes sociales similares han impactado en la manera en como nos relacionamos con las personas de un modo **virtual**. Dichas redes son, en gran medida, una extensión de nuestras interacciones pragmáticas, y como tales, tenemos una extensa variedad de formatos que podemos identificar claramente como actos de habla y conversaciones, e igualmente aplicamos en ellos inferencias y suposiciones para poder comprenderlos y producirlos:

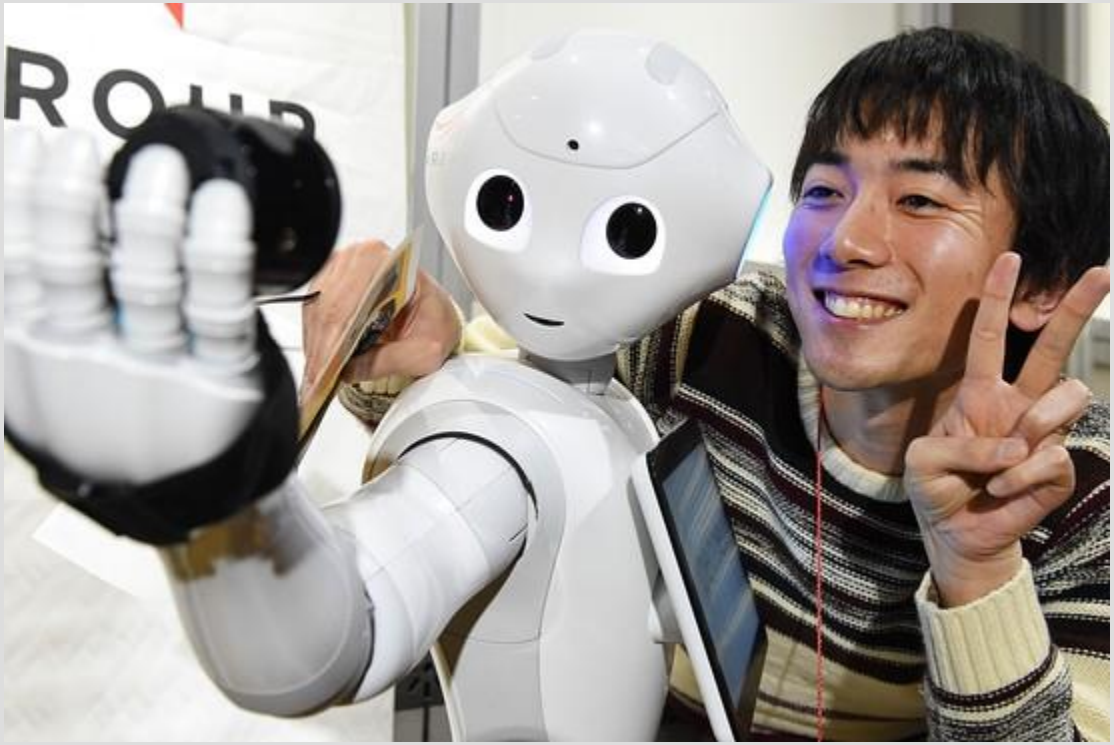
La revolución de *Facebook* (4)

Para la investigación pragmática, las redes sociales nos brindan dos posibles objetos de investigación:

I. Analizar el comportamiento real de las personas en situaciones de interacción, aprovechando al máximo la enorme cantidad de datos que se generan por medio de estas redes.

La revolución de *Facebook* (5)

II. Utilizar esos mismos datos para entrenar a computadores y robots respecto a los distintos modelos de interacción que tienen las personas entre sí. Si esto es viable, entonces debemos pensar en implementar **métodos de aprendizaje supervisado** y **no supervisado** que les permitan aprender a usar lenguaje natural:



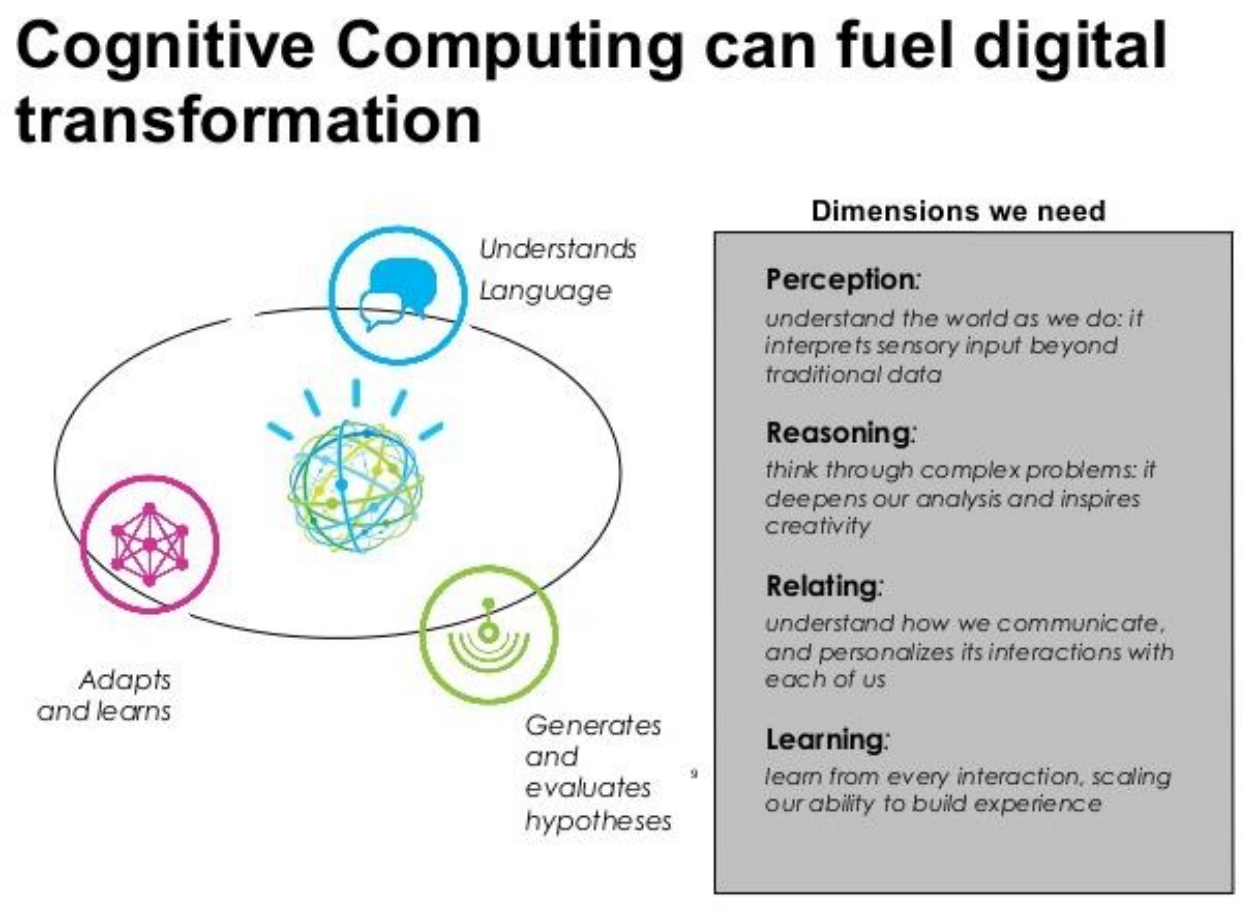
El giro paradigmático de Watson (1)

En el año 2011, IBM participó en el programa televisivo *Jeorpardy!*, en una serie de sesiones en donde sus competidores eran campeones de sorteos anteriores. La idea, a grandes rasgos, era demostrar que en ese momento existía la tecnología necesaria para desarrollar un cerebro electrónico capaz de responder preguntas y respuestas como lo hacen los humanos. Lo que ocurrió en ese concurso ya es parte de los anales de la historia de la inteligencia artificial:

www.youtube.com/watch?v=P18EdAKuC1U

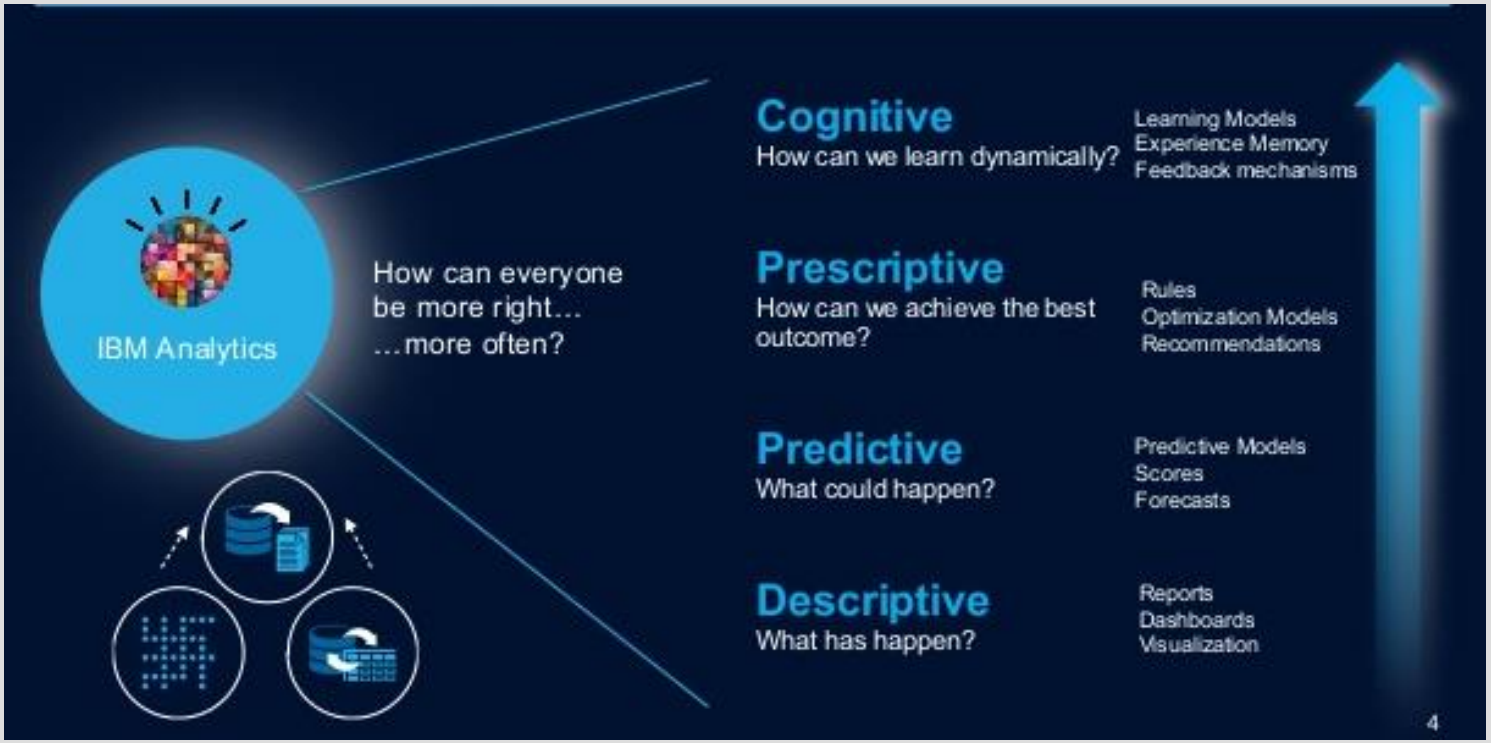
El giro paradigmático de Watson (2)

Muchos expertos en inteligencia artificial, tanto científicos académicos como industriales, consideran que Watson ha inaugurado un nuevo paradgima en las ciencias computacionales, denominado **computación cognitiva**, el cual se entiende del siguiente modo:



El giro paradigmático de Watson (3)

La computación cognitiva plantea una pregunta sumamente interesante e inquietante: ¿podemos crear máquinas capaces de aprender como lo hacen los humanos? Si esto es posible, el camino a seguir sería:



www.youtube.com/watch?v=_Xcmh1LQB9I

Estableciendo equivalencias (1)

Inteligencia artificial

Deducción y razonamiento

Representación del conocimiento

Aprendizaje automático

Percepción

Creatividad

Cibernética y robótica

Inteligencia social

Procesamiento del lenguaje natural

Procesamiento de voz

Procesamiento de texto

Estableciendo equivalencias (2)

Lingüística teórica

Fonética/ Fonología

Morfología

Lexicología

Sintaxis

Semántica

Análisis del discurso

Pragmática

¿Lingüística computacional?

Estableciendo equivalencias (3)

Lingüística aplicada

Psicolingüística

Neurolingüística

Terapia del lenguaje

Lexicografía

Adquisición y desarrollo del lenguaje

Traducción

Terminología

Análisis político del discurso

Sociolingüística

Enseñanza de lenguas

¿Ingeniería lingüística?

Estableciendo equivalencias (4)

Lingüística computacional

Procesamiento de voz

Lematización /tokenización

Lexicones

Chunking/ parsing

Relaciones semánticas

Representación del discurso

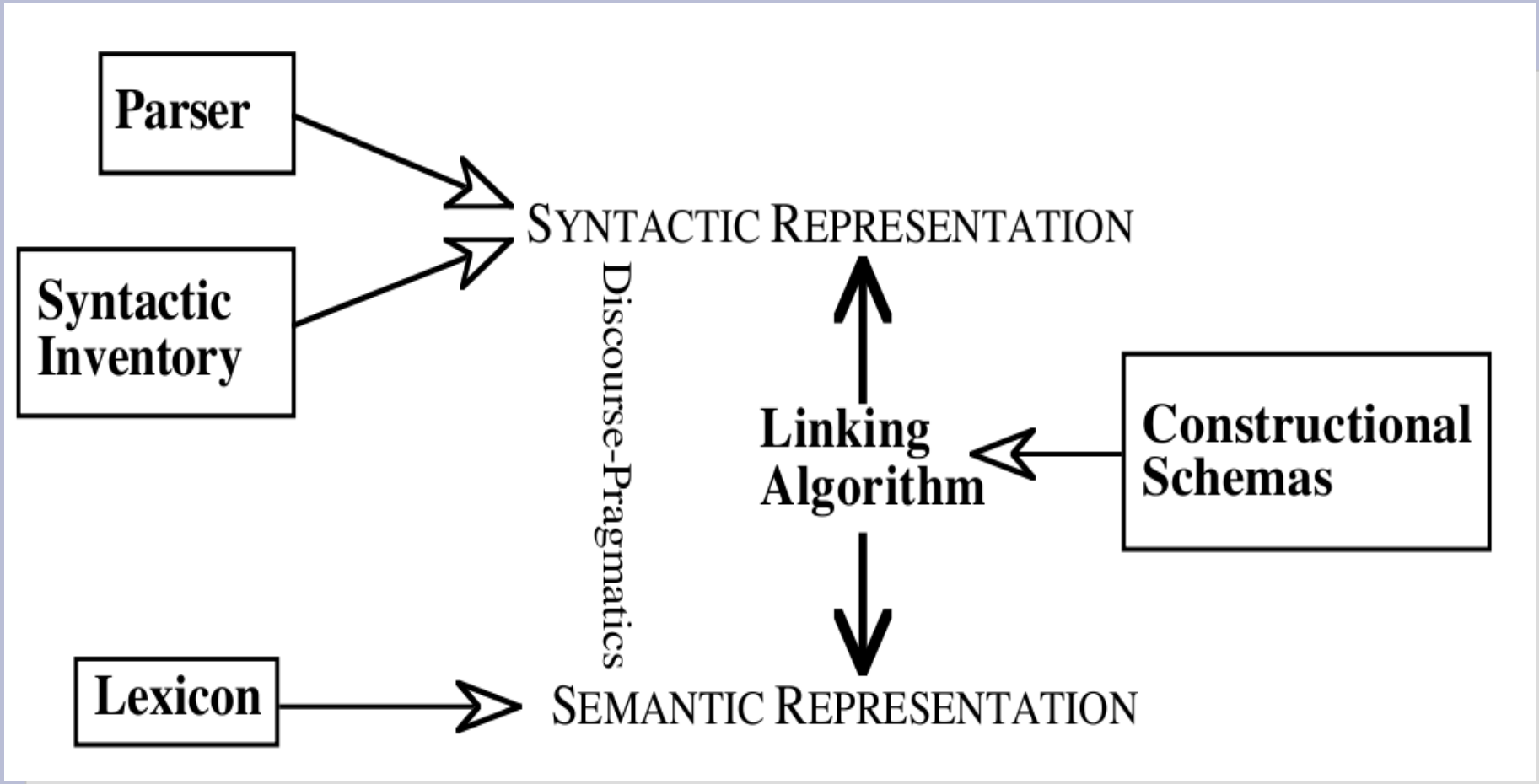
Inferencias pragmáticas

Estableciendo equivalencias (5)

Ingeniería lingüística

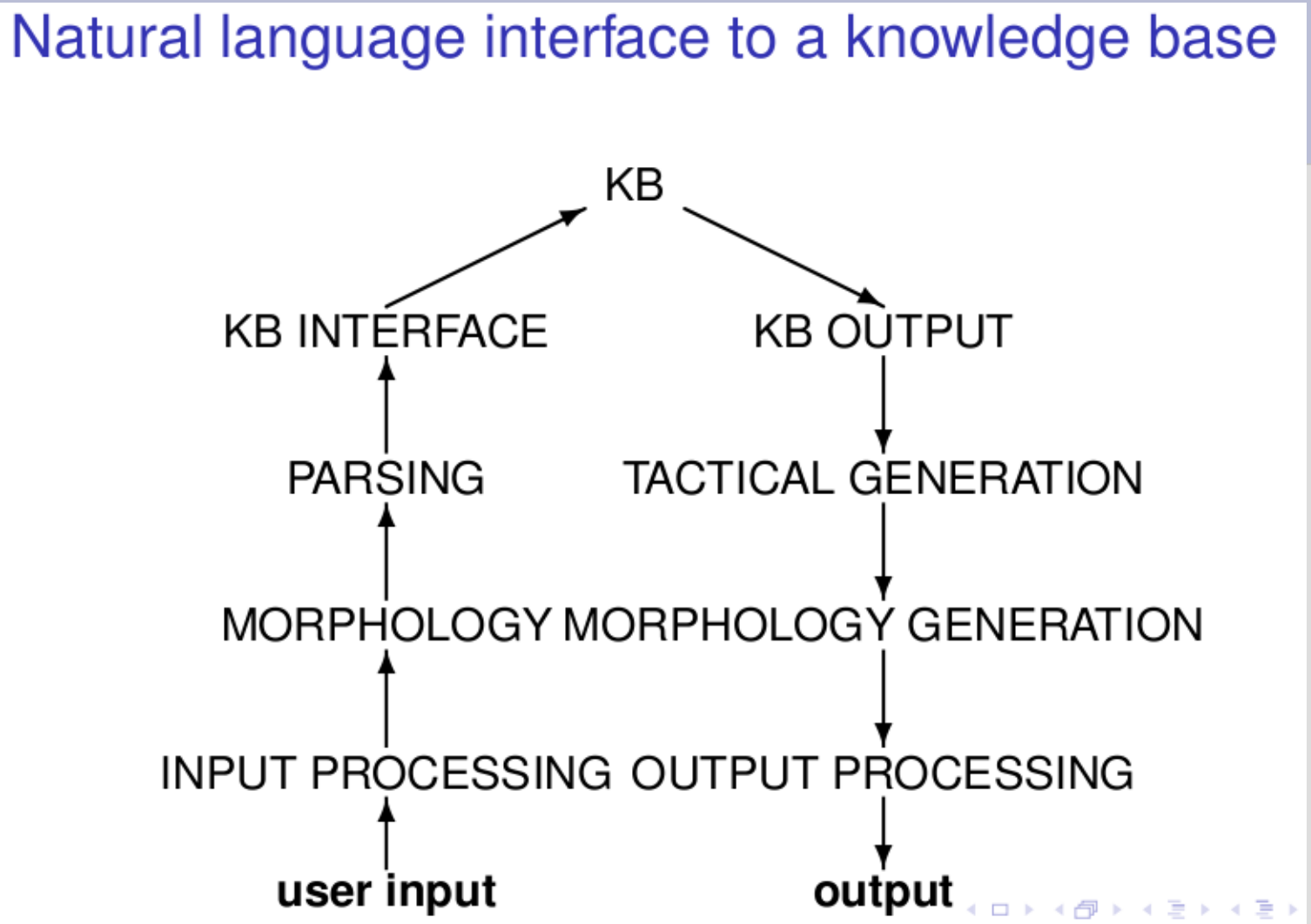
Reconocimiento/ generación de voz	Lexicografía computacional	Traducción automática	Corpus lingüísticos
Recuperación y extracción de información	Extracción terminológica	Sistemas pregunta-respuesta	Resumen automático
Clasificación de documentos	Minería de textos	Reconocimiento de entidades nombradas	Ontologías

¿La misma visión del mundo? (1)



Tomado de: Van Valin (2006): 264

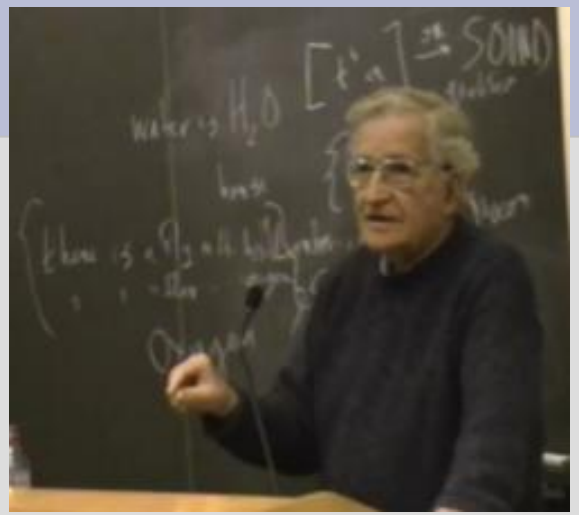
¿La misma visión del mundo? (2)



Tomado de: Copestake (2007)

¿Y qué hacemos? (1)

Como todo buen lingüista, nos interesa estudiar, describir y modelar las formas y las funciones que operan en toda lengua humana.



Como buenos computólogos, siguiendo la propuesta de Turing o Von Neumann, queremos ocupar ese conocimiento lingüístico para crear máquinas inteligentes.

¿Y qué hacemos? (2)

Un rasgo particular de los lingüistas computacionales respecto a los colegas lingüistas, es que ponemos énfasis en desarrollar modelos formales a partir de datos empíricos

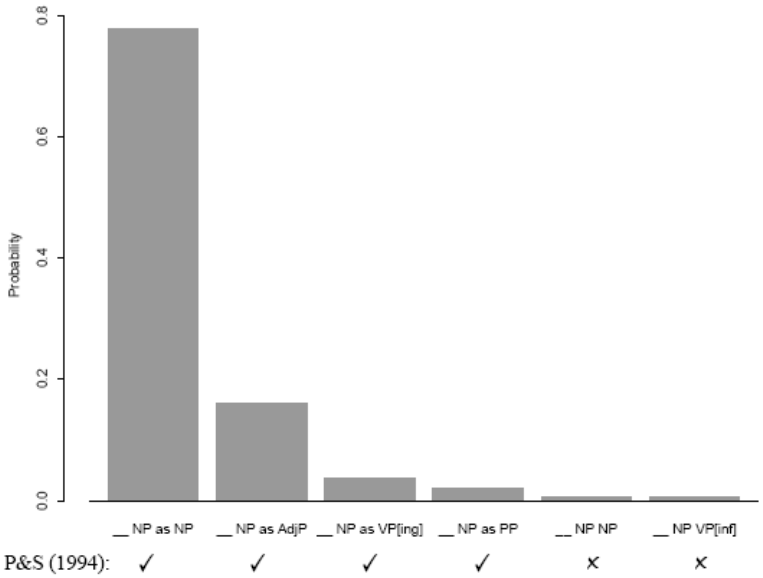


Figure 1: Estimated probability mass function (pmf) for subcategorizations of *regard* based on a 300 token *New York Times* corpus.

Un rasgo que distingue a los lingüistas computacionales de los colegas computólogos, es que trabajan con fenómenos dinámicos, que se ven afectados por un entorno social.

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/curso-de-procesamiento-del-lenguaje-natural.html