Tema 1: Procesamiento automático de lengua natural

Margarita Alonso Ramos

Master LUP

Octubre 2011

¿Qué se procesa?

Lengua

- lengua vs. lenguaje
- lenguaje:

la capacidad, específica de la especie humana, de comunicar por medio de un sistema de signos orales (o lengua) que pone en juego una técnica corporal compleja y que supone la existencia de una función simbólica y de centros corticales genéticamente especializados" (Dubois et al. 1994: 264).

- lenguaje e Inteligencia artificial
- lenguas: español, gallego, swahili, etc.

¿Qué se procesa?

Lengua natural

- lengua natural vs. lengua artificial
- aritmética, como ejemplo de lengua artificial

1) 23+45 = 68

2) *23 45 = 68 +

3) *23+45 = 69

(1) es una expresión sintácticamente correcta y semánticamente verdadera

(2) es sintácticamente incorrecta y no tiene valor de verdad

(3) es sintácticamente correcta pero semánticamente falsa.

- lenguajes de programación como Prolog

¿Qué se procesa?

Lengua restringida

- lengua restringida o controlada o “sublengua”
  - textos escritos, homogéneos y de dominos técnicos
  - aplicaciones funcionan sobre textos cortos, con sintaxis elemental, vocabulario poco polisémico
  - boletines meteorológicos, descripciones técnicas de aparatos, informes de bolsa, patentes, etc.
- Pero los avances que abarcan la **lengua general** están en aumento

¿Cómo se procesa?

Procesamiento

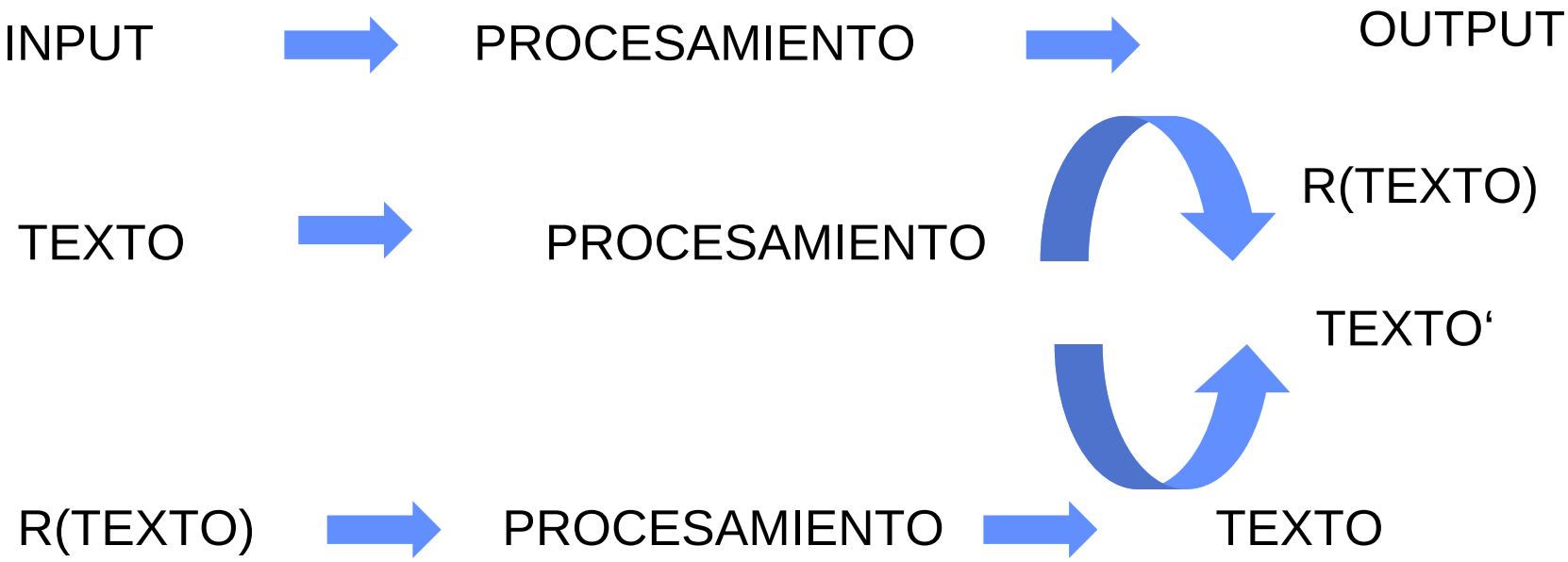
- *procesar*: actuar sobre un objeto, manipulándolo, transformándolo o incluso creándolo
- automático vs. manual

  máquina→ ordenador, máquina concebida para hacer cálculos

- procesamiento automático: secuencia de instrucciones (cálculos) que efectuará la máquina en un cierto orden cronológico, es decir, un programa.
- automatización total : procesamiento efectuado por el ordenador, sin intervención del humano en la ejecución de la tarea
- automatización parcial: efectuados en parte por el ordenador y en parte por el humano.

¿Cómo se procesa?

Procesamiento



¿Cómo se procesa?

Procesamiento

- Los textos son abordados como objetos que se pueden manipular
- para poder procesar automáticamente un objeto, hay que conocer los principios de constitución interna, es decir, debemos describirlo de manera operacional.
- Un texto debe poder ser descrito como un conjunto de correspondencias entre sentidos y formas, regido por reglas explicitables, las reglas de la lengua.

$$\{\text{SENTIDOi}\} \xleftarrow{\text{lengua}} \{\text{FORMAj}\} \mid 0<i, j< \infty$$

- EL PLN requiere
  - conocimientos lingüísticos formalizado en forma de reglas
  - programas informáticos que operan sobre las reglas

El instrumento

Ordenador

- una máquina programada para procesar automáticamente la información
- las informaciones contenidas en la memoria: caracteres codificados en código numérico binario, esto es, secuencias de 0 y 1 (bits)
- algoritmos: secuencias de instrucciones que permiten a la máquina acceder a los datos y manipularlos
- Lenguaje máquina ➔ Lenguaje de Programación ➔ Lengua Natural
- Informático analista
    - Informático programador
        - Ordenador traduce el LP a LM
        - por medio de un programa intérprete
        - por medio de un compilador

Disciplinas implicadas

Lingüística

Disciplina científica que se ocupa del desarrollo de modelos para la descripción de las lenguas naturales

Lingüística Computacional

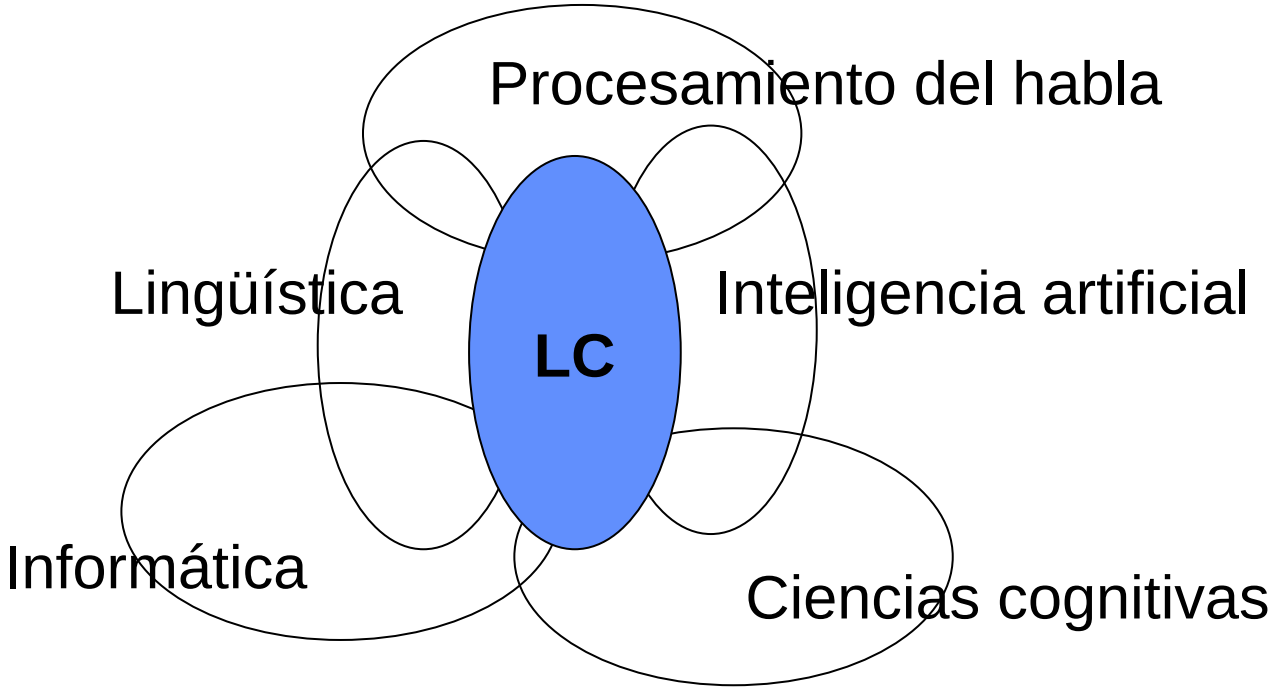
A. Lingüística Computacional Teórica

Disciplina científica que se ocupa del desarrollo de modelos para la descripción de las lenguas naturales y de su implementación

B. Lingüística Computacional Práctica (Aplicada)

Disciplina científica que se ocupa de la aplicación de modelos formales descriptivos de la lengua natural para resolver un amplío abanico de tareas de comunicación

Linguística computacional en su contexto



Breve Historia de la LC Aplicada, 1

Finales de los 50 e inicios de los 60:

- Traducción automática inglés-ruso usando técnicas de criptografía desarrolladas durante la segunda guerra mundial
- Traducción palabra a palabra

  Engl.The spirit is willing but the flesh is weak lit. ‘El espíritu está dispuesto pero la carne es débil’

  Rus. *Vodka xoroshaja, no mjaso protuxlo* lit. ‘El vodka es bueno pero la carne se pudre’

De principios a mediados de los 60:

- Con el nacimiento de la Inteligencia Artificial, aparece la división entre “Lingüística Computacional” (LC) y “Procesamiento del Lenguaje Natural” (PLN)
- LC usa modelos lingüísticos
- PLN usa modelos cognitivos y modelos de conocimiento *ad hoc*

Breve Historia de la LC Aplicada, 2

70s y principios de los 80:

- LC: se centra en el análisis de frases orientado a la superficie (sintaxis) y en la traducción automática (en Europa)
- PLN: se centra en la representación del conocimiento para el procesamiento del lenguaje, generación de texto a partir de representaciones de conocimiento y "question answering", …

A partir de mediados de los 80

- Paulatina conversión de la LC y el PLN en un solo campo.
- Se desarrollan nuevas teorías lingüísticas con especial atención a sus propiedades formales y a su implementación
- Uso cada vez mayor de métodos empíricos

A partir de finales de los 90:

- La distinción entre LC y PLN depende sólo de criterios ideológicos.

El porqué

¿Por qué es la LC / PLN un tema importante?

Vivimos en la **era de la información**. Cada día se espera que produzcamos información –hablada o escrita-, y también que consumamos información la cual puede llegarnos a través de distintos canales: radio, TV, email, internet, teléfono, medios impresos, etc. Es necesario desarrollar sistemas informáticos que sean capaces de ayudar a la gente a afrontar semejante reto.

Para desarrollar este tipo de sistemas hacen falta especialistas.

El porqué

En 1984, un responsable de la comunidad europea escribía:

"La lengua es un elemento primordial de la producción industrial: una central nuclear representa 100.000 páginas de documentación técnica. Un automóvil, una central telefónica, un ordenador, son productos cuya producción, reparación, exportación exigen un importante esfuerzo de redacción y de traducción técnica muy costoso".

Vertiente política:

En 1985, el secretario del Consejo de Europa declaraba:
"Las lenguas que no se industrialicen dejarán de ser vehiculares, de ser lenguas de civilización".

El porqué

Humanidades y Sociedad de la información

- las humanidades (lenguaje y cultura) son esenciales para acceder a la información en la sociedad globalizada y post-industrial
- aquellas lenguas que no estén digitalizadas y carezcan de presencia en la red encontrarán difícil sobrevivir en la comunicación socio-económica reinante
- Bill Gates: el lenguaje, en especial el lenguaje hablado, no es sólo el futuro de Windows, sino el futuro de la computación en general
- Nuevas tecnologías como catalizadoras de la innovación educativa
- Nuevos métodos científicos de acercamiento multidisciplinar

El porqué

Humanidades y Sociedad de la información

La revolución sigue siendo la misma: el contenido. En la pantalla que sea. La gente emplea múltiples pantallas: el televisor, la tableta, los móviles...Nuestra vida va de una a otra pantalla. Tenemos que asegurarnos de que Yahoo! transfiere en el momento adecuado la correcta información a cualquiera de estas pantallas. Mi trabajo es el contenido y hacerlo llegar a cualquier sitio. El negocio es el contenido.

Carol Bartz, consejera delegada de Yahoo!, 16 de febrero 2011

El buscador ideal sería aquel en el que podamos pedir en nuestra lengua lo siguiente:

Búscame documentos en cualquier idioma que hablen de tal asunto, selecciónname los más relevantes, clasifícalos de acuerdo con tal criterio y dame de cada uno un resumen en mi lengua.

El porqué

De la Información al Conocimiento

Este proceso depende de los sistemas lingüísticos en los que está codificada la información

Posibilidad de procesar la información en el sistema que ha sido formulada y devolver el resultado en ese mismo sistema o en uno semejante

Para llevar a cabo este proceso hacen falta

Tecnologías lingüísticas

Tecnologías lingüísticas

otro término: *ingeniería lingüística*

- TL : consisten en la aplicación del conocimiento sobre la lengua al desarrollo de sistemas informáticos capaces de reconocer, analizar, interpretar y generar textos
  - el resultado son máquinas que se comportan como si comprendieran el lenguaje humano

- contribuyen al mantenimiento y desarrollo de una sociedad plurilingüe en la que cada usuario pueda acceder a la información que necesita y recibirla en su propia lengua
  - sistemas que ayuden en el aprendizaje de idiomas asistido por ordenador,
  - programas que ayuden en las tareas de traducción, etc.

Tecnologías lingüísticas

- los avances en TL permiten
  - acceder a la información con eficacia
  - aprender otras lenguas y mejorar la propia
  - hacer negocios trabajando con sistemas informáticos de funcionamiento vocal

⇨ los dos principales objetivos

  - conseguir que nos relacionemos de forma natural con las máquinas
  - encontrar, seleccionar, resumir y presentar la información en la forma que desee el usuario

⇨ aplicaciones

  - sistemas que permiten al usuario comunicarse con el ordenador: consultas a bases de datos, recuperación y extracción de la info, interfaces hombre-máquina
  - sistemas que permiten a los humanos a comunicarse entre sí en diferentes lenguas: traducción automática, aprendizaje de lengua asistido por ordenador (ALAO)
  - sistemas de ayuda en tareas lingüísticas: herramientas de análisis textual y de corpus, de ayuda a la escritura y de creación de bases de datos lexicográficas

¿Por qué es tan difícil el PLN de alta calidad?

El PLN requiere:

⇨el procesamiento a varios niveles de un modelo de lenguaje

⇨ unidades y conjuntos de unidades que pueden contener ambigüedad (lo que dificulta la tarea de “comprensión”)

⇨ con características altamente idiosincrásicas (lo que dificulta la “producción”)

⇨ Herramientas:

- lematizadores, analizadores morfológicos y sintácticos, generadores

⇨ Recursos

- léxicos y corpus

Algunos tipos de ambigüedades

- **Pronunciación de palabras**:

  El padre quiere ca[*s*|*z*]*arla*

  (confusión "s" y "c": *cazar* vs. *casar*)

- **Separación de palabras:**

  [*type out*] vs. [*type*] [*out*]

- **Categoría gramatical de las palabras:**

  [*el*] $comunicado_{Noun}$ vs. $comunicado_{Part}$

- **Estructura de la oración:**

  Vi a un hombre con un telescopio

- **Significado de las palabras:**

  age = época | edad;

- **Semántica del discurso:**

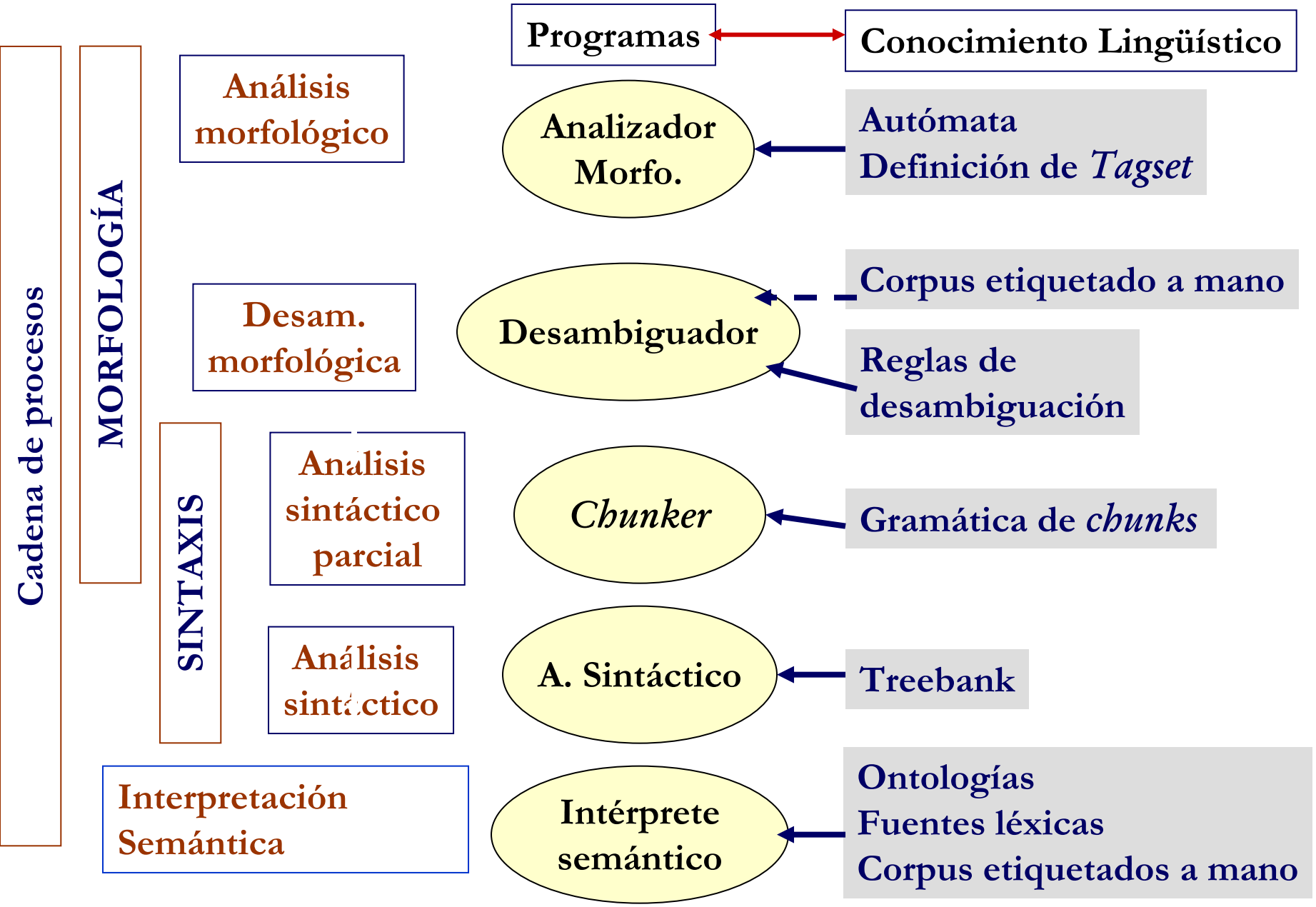
  Y, de hecho, son tontos

- **Pragmática:**

  Querría un billete para Barcelona

Bloques constituyentes y niveles de descripción linguística

Palabra	Estructura de las palabras (morfología)	Significado de las palabras (semántica léxica)
Oración	Estructura de las oraciones (sintaxis)	Significado de las oraciones (semántica sintáctica)
Texto	Estructura del texto (discurso)	Significado del texto (semántica del discurso)
Con-Texto	Estructura del con-texto	Significado del con-texto (pragmática)


Cadena de procesos
MORFOLOGÍA
SINTAXIS
Análisis morfológico
Desam. morfológica
Análisis sintáctico parcial
Análisis sintáctico
Interpretación Semántica
Programas
Conocimiento Lingüístico
Analizador Morfo.
Autómata
Definición de Tagset
Desambiguador
Corpus etiquetado a mano
Reglas de desambiguación
Chunker
Gramática de chunks
A. Sintáctico
Treebank
Intérprete semántico
Ontologías
Fuentes léxicas
Corpus etiquetados a mano

Procesos: Análisis morfológico

Qui qui pr0cn000 qui pt0cn000
va anar vmip3s0 anar vaip3s0
guanyar guanyar vmn0000
la el da0fs0 ell pp3fs000 la ncms000
Copa_Davis Copa_Davis NP00000
l' el da0cs0 ell pp3cs000
any_1968 1968 W
? ? Fit

Procesos: desambiguación morfológica

Qui qui pt0cn000
va anar vaip3s0
guanyar guanyar vmn0000
la el da0fs0
Copa_Davis Copa_Davis NP00000
l’ el da0cs0
any_1968 1968 W
? ? Fit

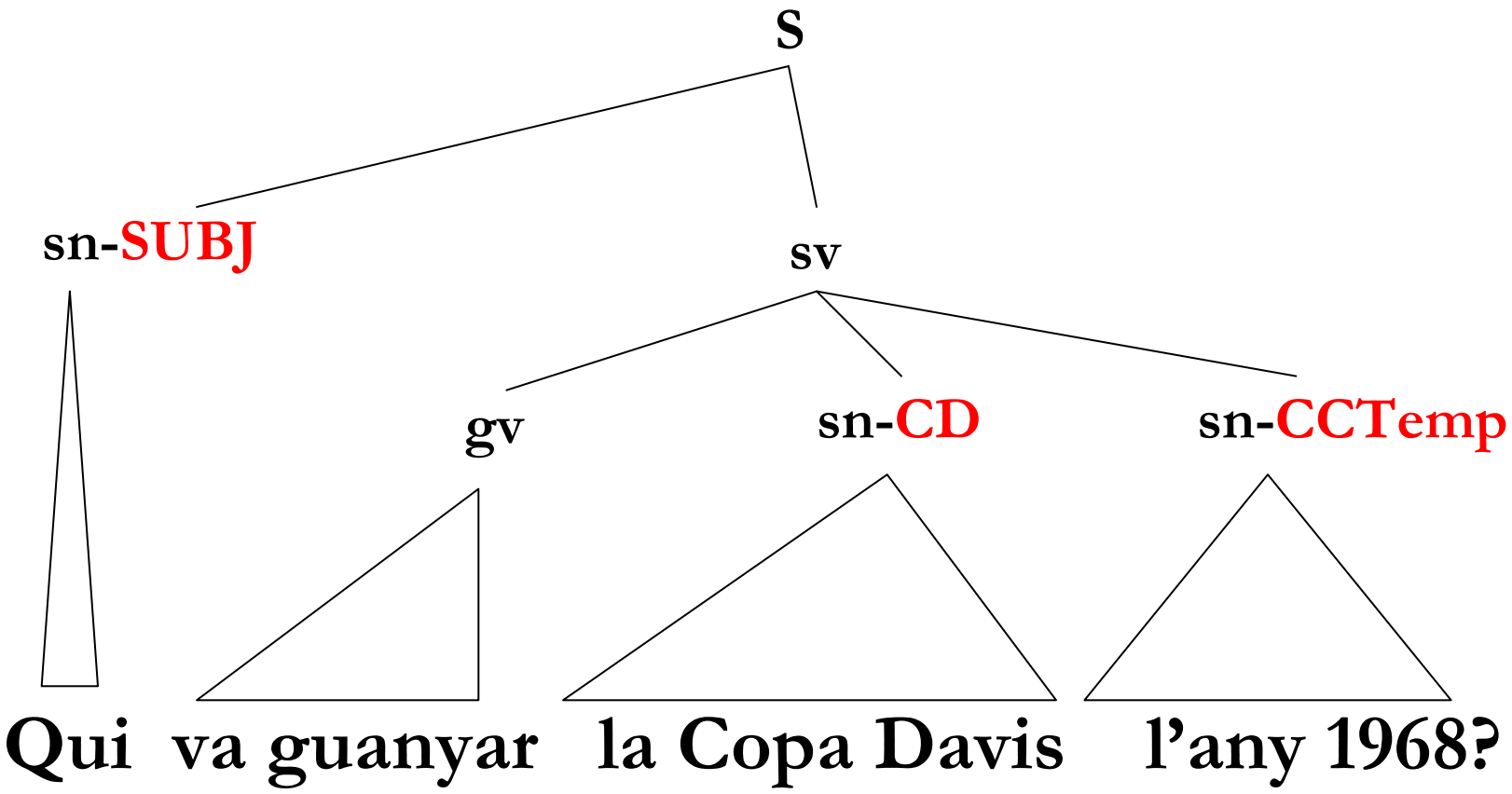
Procesos: Chunking

Qui va guanyar la Copa Davis l'any 1968?

```
S_{
    sn_{ Qui }
    sv_{ va guanyar }
    sn_{ la Copa_Davis }
    sn_{ l'any_1968 }
    ?
    }
```

Procesos: análisis sintáctico

Qui va guanyar la Copa Davis l'any 1968?



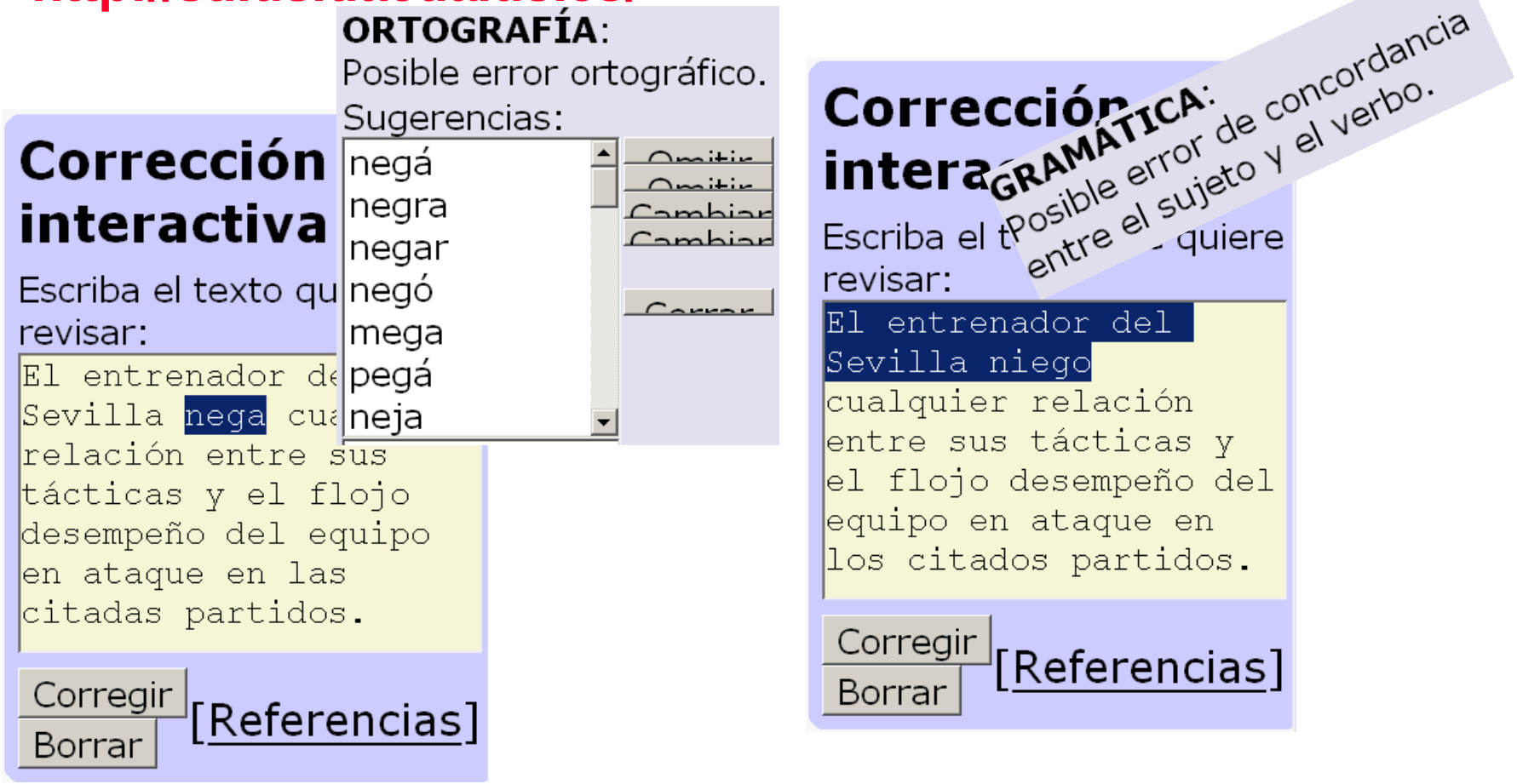
Principales aplicaciones del PLN

i. Programas de ayuda a la escritura (corrector ortográfico y corrector de estilo)
ii. Reconocimiento del lengua hablada y síntesis de la voz.
iii. Traducción automática
iv. Generación de texto (o de documentos multimedia)
v. Resumen automático
vi. Sistemas de respuestas a preguntas (“Question answering”)
vii. Recuperación y Extracción de la información
viii. Aprendizaje de lenguas asistido por ordenador (ALAO)
ix. Adquisición de recursos lingüísticos (diccionarios, gramáticas) (p. ej. a partir de corpus)

Un vistazo a las aplicaciones actuales

Correctores ortográficos y correctores de estilo

http://stilus.daedalus.es/

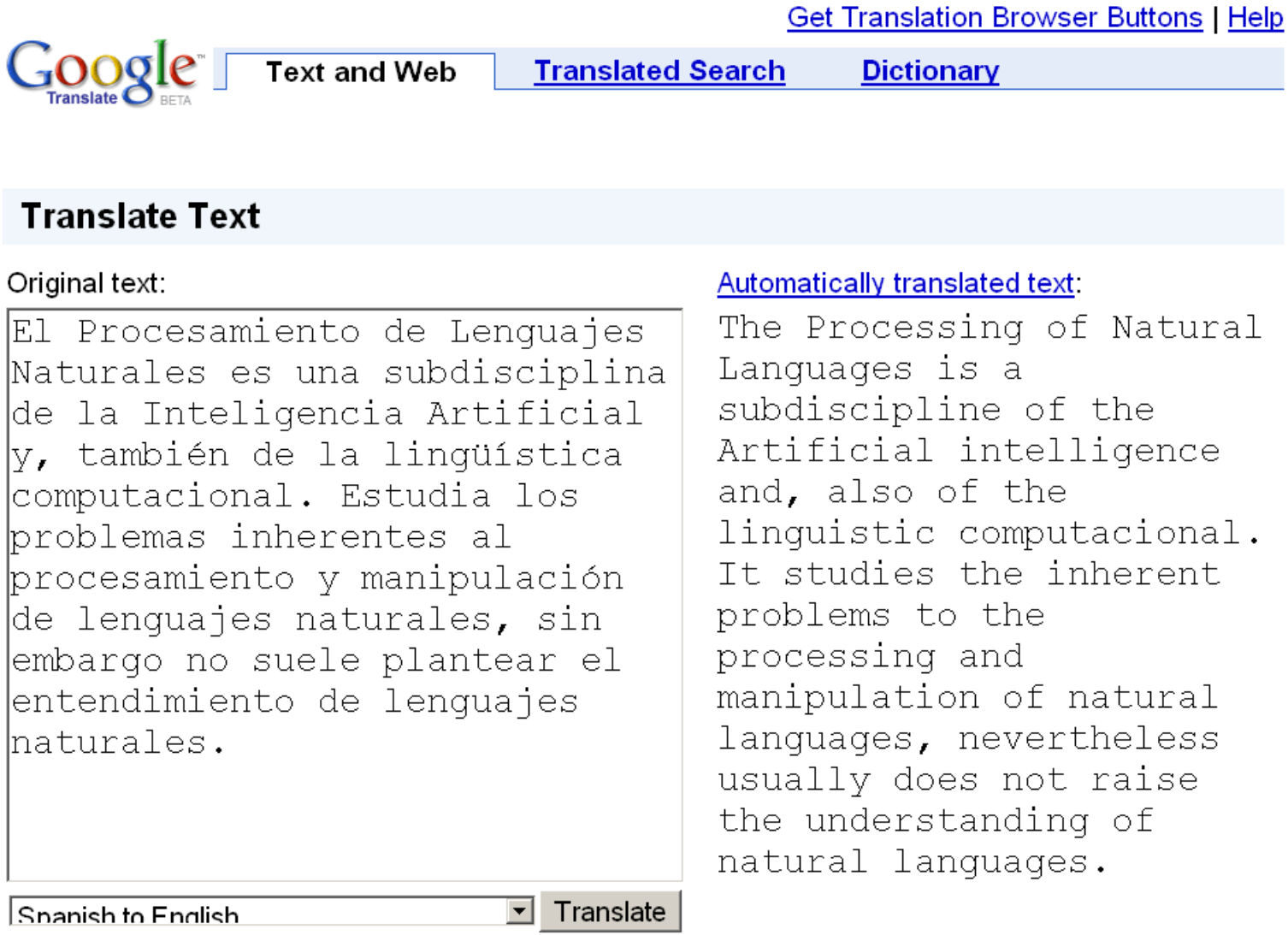


Un vistazo a las aplicaciones actuales

Correctores ortográficos

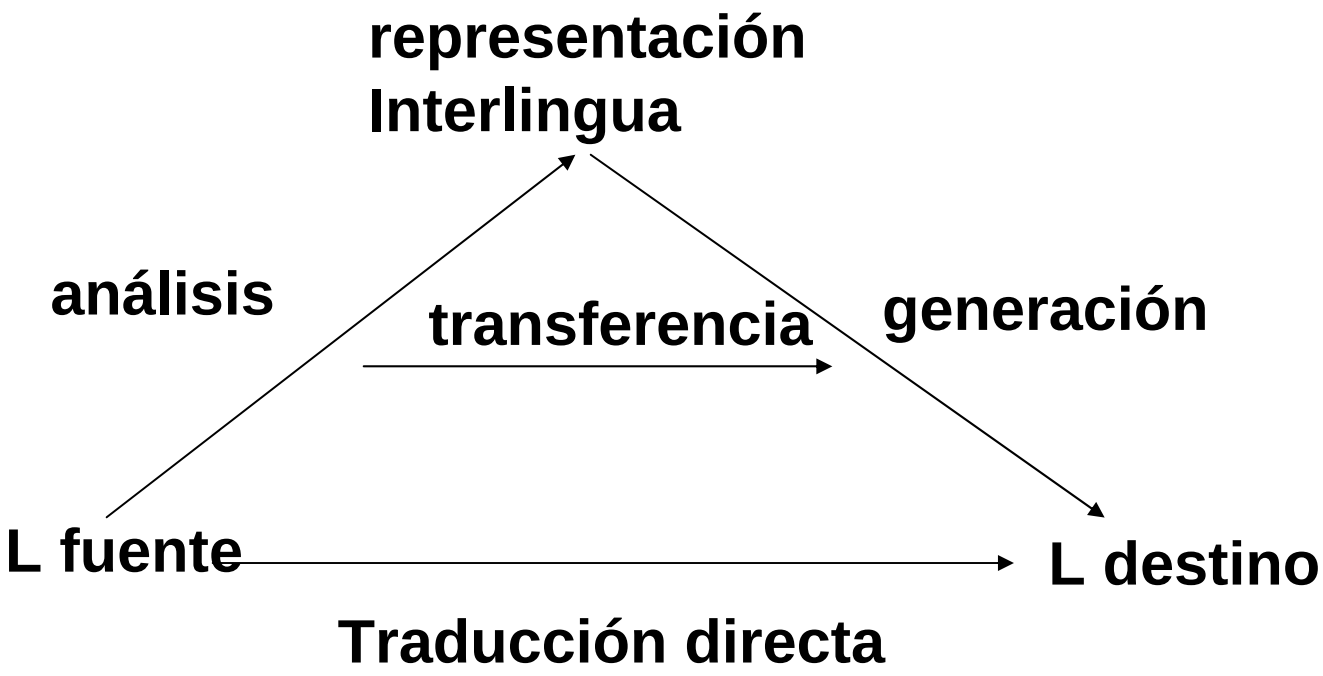
Un vistazo a las aplicaciones actuales

Programas de traducción automática



Nivel de procesamiento en relación a una aplicación

Traducción automática

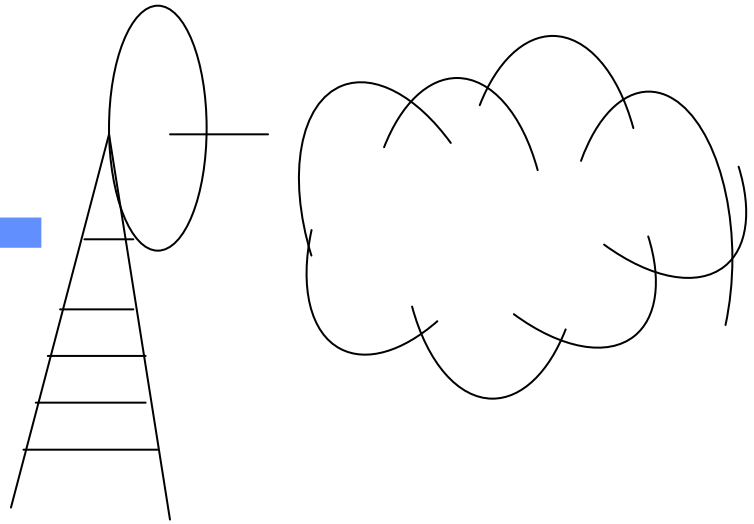


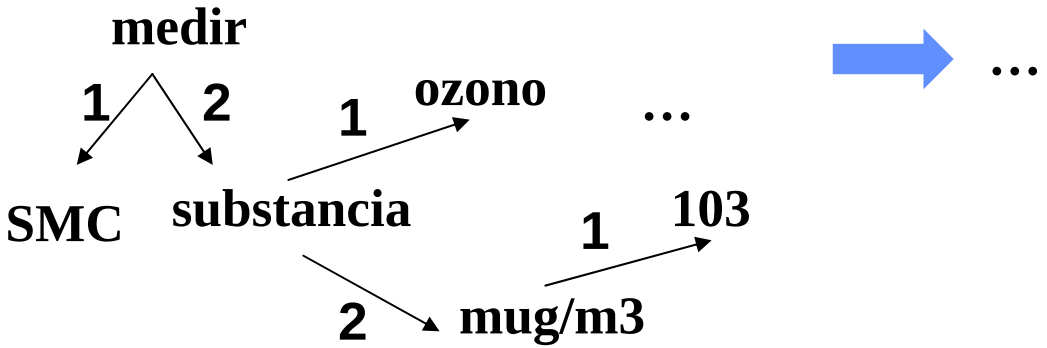
Nivel de procesamiento en relación a una aplicación

Generación de textos:

action: monitor
(unit: hour)
(time: 12)
(value: 103
(unit: mug/m3))
(substance: ozone)
(agent: SMC)
(location: V.Laietana)

(t = 07; v = 0)
(t = 08; v = 2)
(t = 10; v = 54)
(t = 11; v = 94)
(t = 09; v = 11)
(t = 12; v = 103)





Tratamiento parcial en relación a una aplicación

Extracción de información

Bagdad. (EP/AP).- Al menos seis personas murieron y otras 20 resultaron heridas en un atentado suicida contra un centro de reclutamiento del Ejército iraquí, según informó el brigadier de policía, Said Ahmed al Jibori, precisando que el agresor, que llevaba los explosivos bajo su ropa, los detonó mientras se encontraba en medio de demandantes de empleo en Tal Afar, a 150 kilómetros al este de la frontera siria y 420 kilómetros al noroeste de Bagdad.

- Búsqueda de las palabras clave y análisis parcial de las frases (del régimen de los predicados relevantes)

(evento: atentado suicida

número de víctimas: 6 muertos y 20 heridos

lugar: centro de reclutamiento del Ejército iraquí en Tal Afar)

Aplicaciones
Entornos de aprendizaje

1 *tr.-prnl.* Poner en acto o acción:

2 *intr.* Ejercer una persona o cosa actos propios de su

3 Ejercer las funcion pias de un oficio: ~ de secr

4 Representar en el teatro o en cine.

5 Trabajar en un espectáculo pú o.

6 Defender, oposición.

7 DER. Rea

En la misma situación que el goleador azu n la mayoría de jugadores suramericanos que *actúan* er **n 3**

Cardenal *actuará* contra quienes dijeror nó a Geresta.

Acepción 3

Clinton añadió: "*Actuando* ahora, defen ores , protegemos nuestros intereses e impulsamos la causa **n 3**

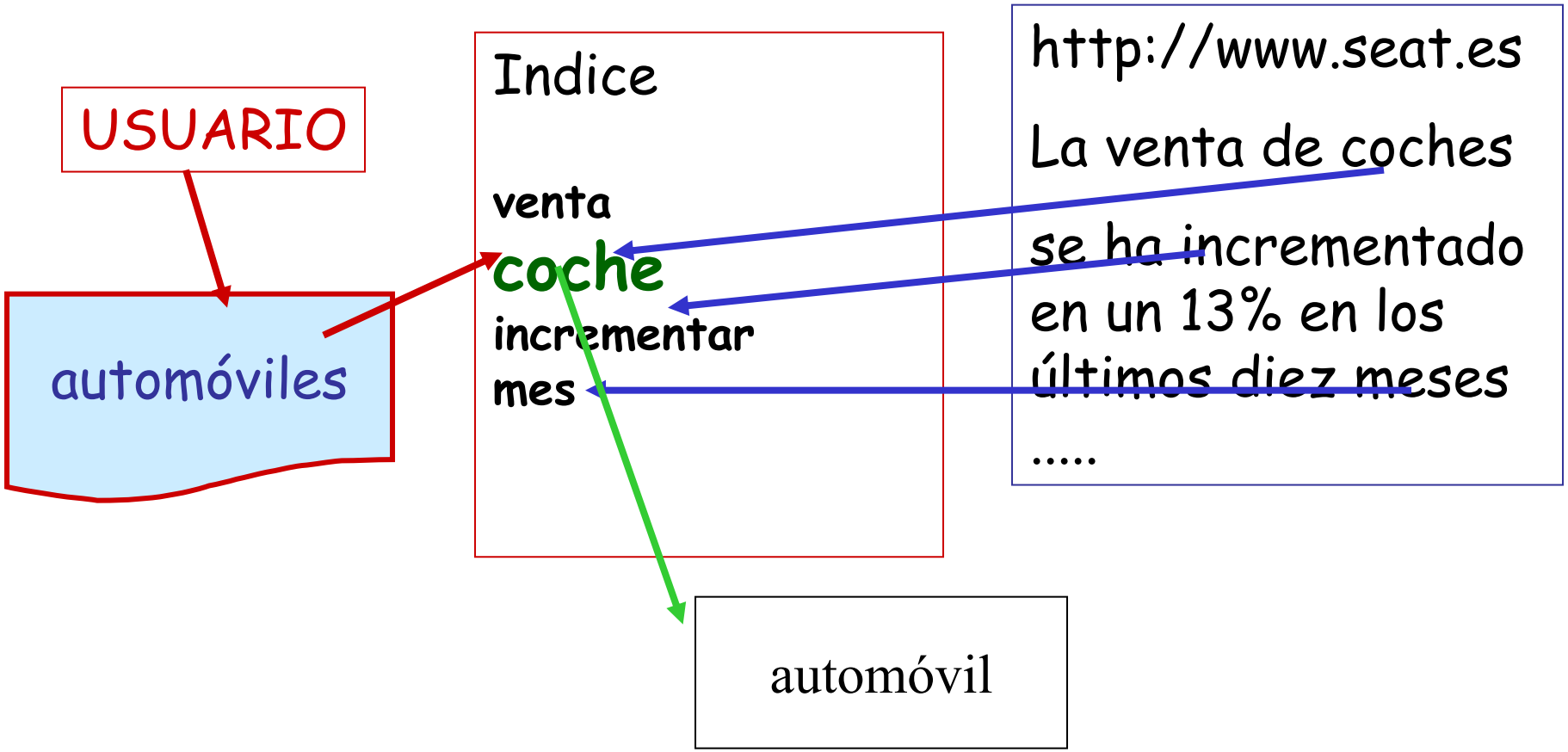
actuar

Presente
actúo
actúas
...
Imperfecto
actuaba
actuabas
...
Futuro
actuaré
actuarás

hacer

actuar → representar

Aplicaciones
Recuperación de Información



Aplicaciones
Recuperación de Información