

Georges Antoniadis
Université Stendhal – Grenoble 3
Laboratoire LIDILEM
BP 25
38040 Grenoble Cedex 9, France
Georges.Antoniadis@u-grenoble3.fr

Enseñanza :

Informática, Tratamiento Automático de la Lengua (TAL),
Generación Automática de Textos (GAT), Aplicaciones del
TAL para el Aprendizaje de Lenguas Asistido por
Ordenador (ALAO)

Investigaciones :

TAL, GAT, Sistemas de ALAO y TAL

Plan

- ▶ Industrias de la lengua y tratamiento automático de las lenguas (TAL)
- ▶ Informática y TAL, “calculabilidad” de la lengua
- ▶ Histórico del TAL
- ▶ Definiciones y problemática del TAL
- ▶ Reconocimiento automático de textos
 - Recorte en formes, aplicaciones
 - Análisis et analizadores morfológicos, aplicaciones
 - Análisis et analizadores sintácticos, aplicaciones
 - Análisis semántico et pragmático automáticas
- ▶ Generación automática de textos
 - Problemática, particularidades de la generación
 - Aplicaciones de la generación
 - El « Que decir », y el « Cómo decirlo »
 - generación profunde
 - generación de superficie
- ▶ Arquitectura informática des sistemas
- ▶ TAL et aprendizaje de idiomas

Industrias de la Lengua

Ejemplos de productos : traductores automáticos, diccionarios informatizados de toda forma, corrector de errores, clasificación automática de informaciones, búsqueda automática de informaciones, contestadores vocales, productos para minusválidos, comunicación con máquinas, etc.

A la base de todo producto de las industrias de la lengua se encuentra el Tratamiento Automático de la Lengua (TAL)

Informática :

Tratamiento automático de la información

Tratamiento automático de la lengua (TAL)

Automático : sin intervención humana, hecho por máquinas

Tratamiento : recolectar, almacenar, clasificar, usar,
comunicar

Información : elemento del conocimiento; aspectos
inmateriales et materiales (forma)

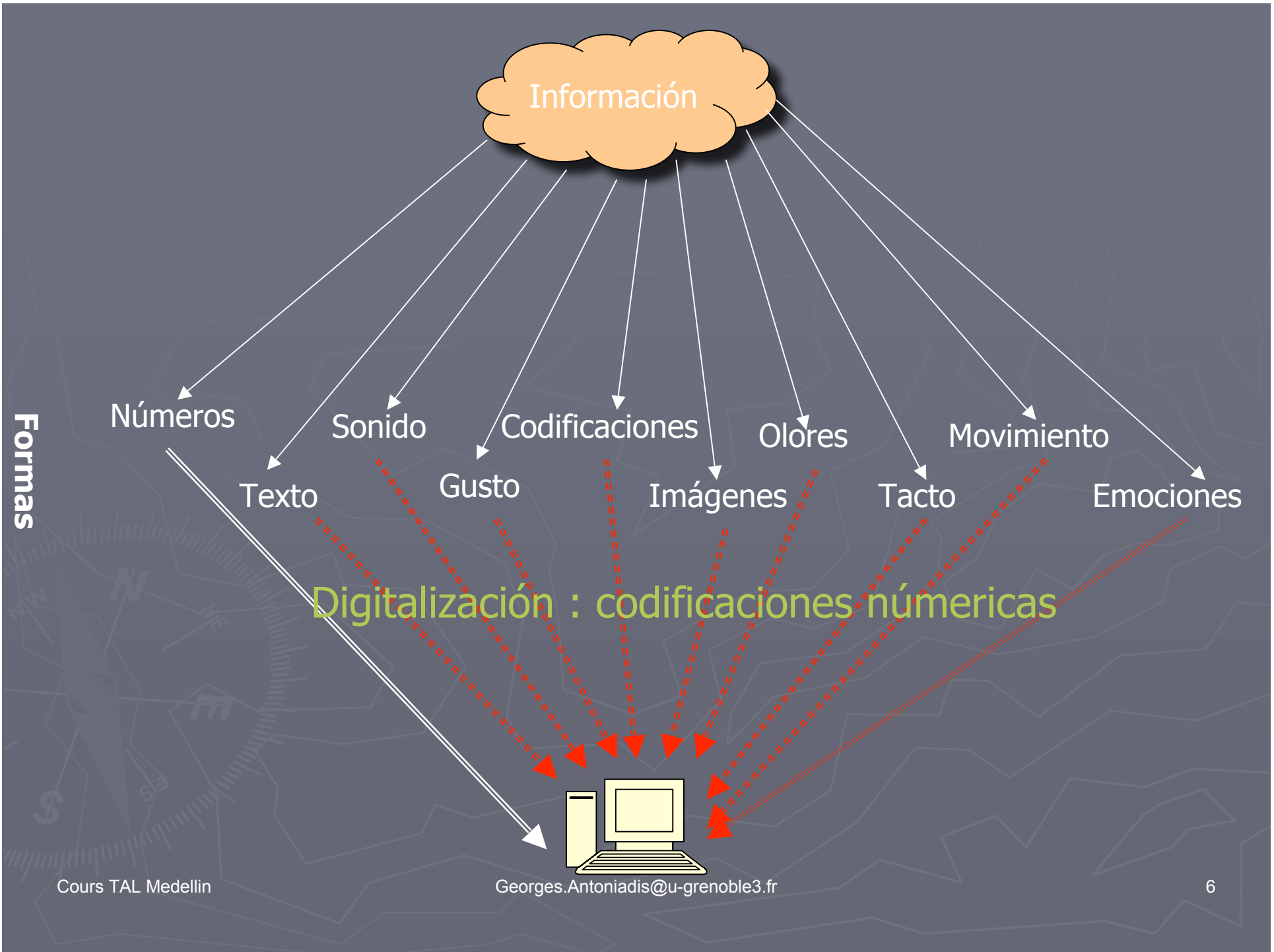
La informática solo trata de forma

Varios tipos de forma :

números, caracteres (texto), toda forma de codificación, imágenes, sonidos (incluido habla), tacto, movimientos, gusto, olores, emociones

Las maquinas informáticas solo tratan de la forma

Para tratar otras formas, hay que transformar les en formas numéricas



Uso de formas de información digitalizadas por la informática

- ▶ Para resolver problemas con ayuda de programas
- ▶ Informática no puede resolver cualquier problema tratando formas digitalizadas de información ; el problema tiene que ser calculable

Problema calculable

Un problema es calculable si dos condiciones están satisfechas :

- 1) Se puede descomponerlo en subproblemas elementaros de modo que la maquina pueda resolver cada subproblema.
- 2) Se obtiene la solución del problema en un tiempo finito.

Ejemplos : Usando las letras del alfabeto formar : todas las secuencias posibles ; todas las palabras del castellano ; todas las secuencias posibles formadas de 3 caracteres máximo.

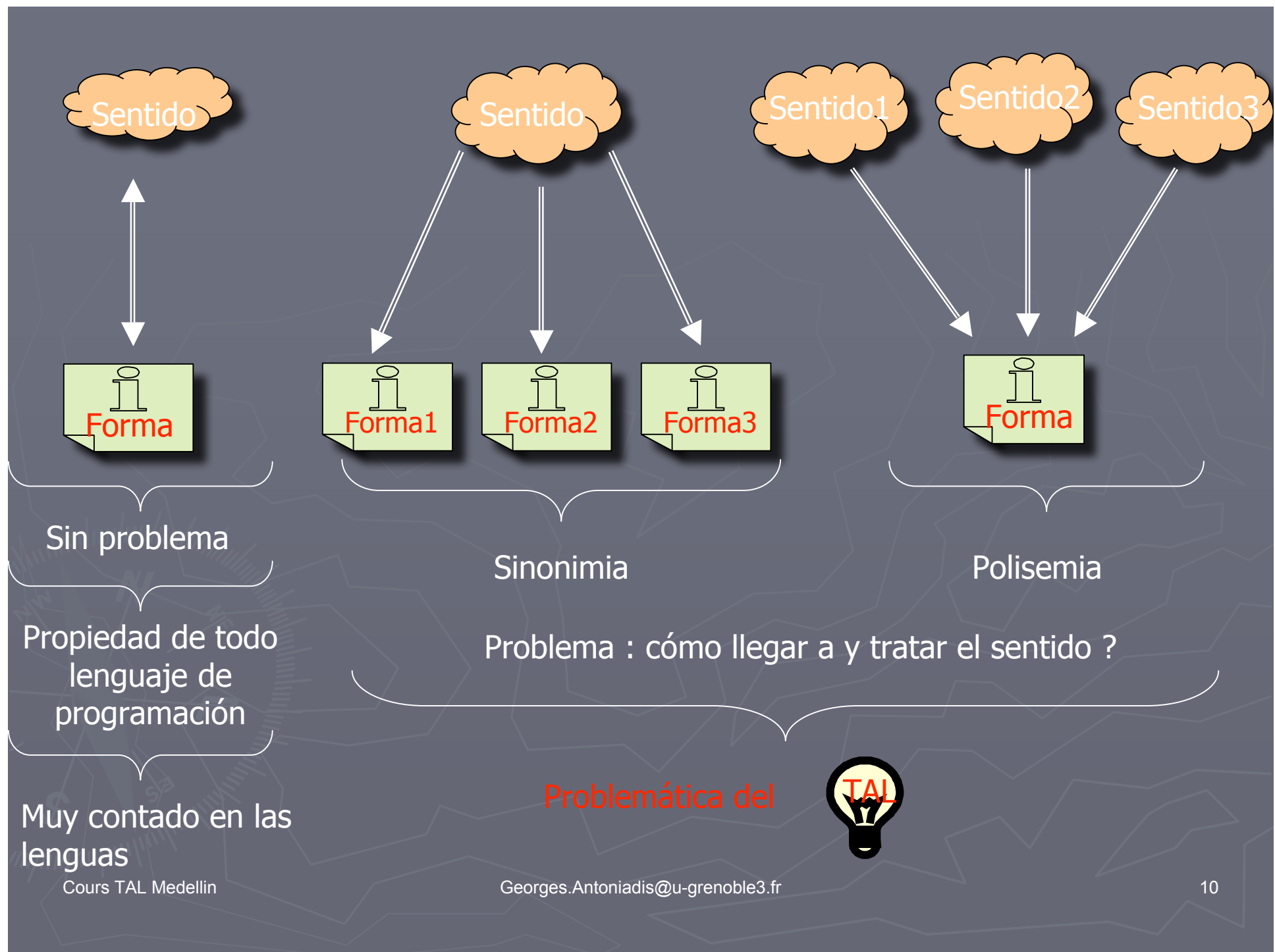
Lengua :

- ▶ sistema de conceptos doblado de un sistema de formas
- ▶ subconjunto de la información



Informática solo puede tratar la forma de la lengua

Problema : cómo tratar el sistema de conceptos (sentido) también ?



Tratamiento Automático de la Lengua

Elaboración de programas informáticos que manipulan la forma lingüísticas y que son capaz, a través la forma, de tratar el sentido asociado, para resolver problemas específicos

El objetivo del TAL es hacer que los problemas implicando informaciones lingüísticas se vuelvan calculables

Problema central del TAL : Tratar la ambigüedad.
Todas las lenguas son ambiguas, por suerte.

Las dos formas del idioma

► Escrita : No hay problema de codificación, tampoco de identificación ; una palabra o una frase se escribe igual, no importa el scriptor.

► Habla (sonido) : Problemas específicos suplementarios para identificar y aislar la parte lingüística dentro de la señal sonora.

Cada forma tiene sus propias dificultades de tratamiento.

Ejemplos : couvent, portions, des sous, etc.

Una gran parte de los problemas y de los tratamientos son comunes

Histórico

- ▶ El TAL empezó en los años 50 en los EE. UU. con el objetivo la traducción automática ruso / inglés (guerra fría)
- ▶ Se pensaba que la traducción automática funcionaria dentro de 15 años
- ▶ Traducción con diccionarios con algunos arreglos de superficie
- ▶ Los resultados no llegaban

Histórico

Comisión, principio de los años 60, para evaluar el estado de los trabajos

Evaluación de los sistemas con doble traducción automática : inglés/ruso/inglés y comparación de las frases en entrada y salida

Ejemplo famoso :

« The spirit is willing but the flesh is weak » →

« The vodka is strong but the meat is rotten »

Reporte ALPAC con la conclusión : *traducción automática es una utopía*

Consecuencia : no más fondos, no más trabajos de TA, transición a la TAC (Traducción Asistida por Computadora)

Histórico

Ya al fin de los años 50 conclusión de Bar-Hillel :

« Hay que tomar en cuenta conocimientos contextuales, contextuales y enciclopédicas »

Otra constatación :

El problema del TAL no es informático pero pluridisciplinario : informática, lingüística formal, lógica, psicología, incluso matemáticas, estadísticas, sociología, antropología, etc.

El “camino recorrido”

- ▶ Al principio sola se trabajaba con un lexico
- ▶ Al fin de los años 50 se tiene en cuenta la sintaxis
 - Trabajos de Chomsky (Teorías de las gramáticas formales)
 - Muy usados para los lenguajes informáticos
 - Problemas con la ambigüedad: frase de 17 palabras → 572 soluciones diferentes

El “camino recorrido”

- ▶ Fin de los años 60 et principio de los años 70, tomada en cuenta de la semántica (primeros trabajos)
 - Fillmore (68): gramáticas de caso (vínculo verbo / objeto)
 - Quillian (68): redes semánticas de palabras
- ▶ Modelos semánticos
 - Schank (72 y después), Wilks (73 y después), Minsky (74) ...
 - Se empieza a querer interesarse al nivel más allá de la frase.
 - Uso y modelización de micro-mundos
- ▶ Programmas famosos
 - Weizenbaum (72): ELIZA, primera herramienta que dialoga libremente en lengua natural
 - Versión Javascript de Eliza : <http://www.manifestation.com/neurotoys/eliza.php3>
 - Winograd (72): SHRDLU, primera herramienta que dialoga en un micro-mundo compuesto de cubos, pirámides, cilindros, que se puede designar, situar en respecto a los otros, mover, etc.
 - <http://hci.stanford.edu/~winograd/shrdlu/>
- ▶ Se investiga desde entonces cómo calcular y implementar informaciones semánticas

Tendencia actual

Transición del tratamiento automático de toda la lengua, al tratamiento específico para problemas particulares

De « querer comprender el sentido de un texto » a « obtener informaciones suficientes para el problema tratado »

Ejemplo : No se necesita entender el sentido de un texto para detectar los errores de ortografía que contiene

Problema muy difícil :

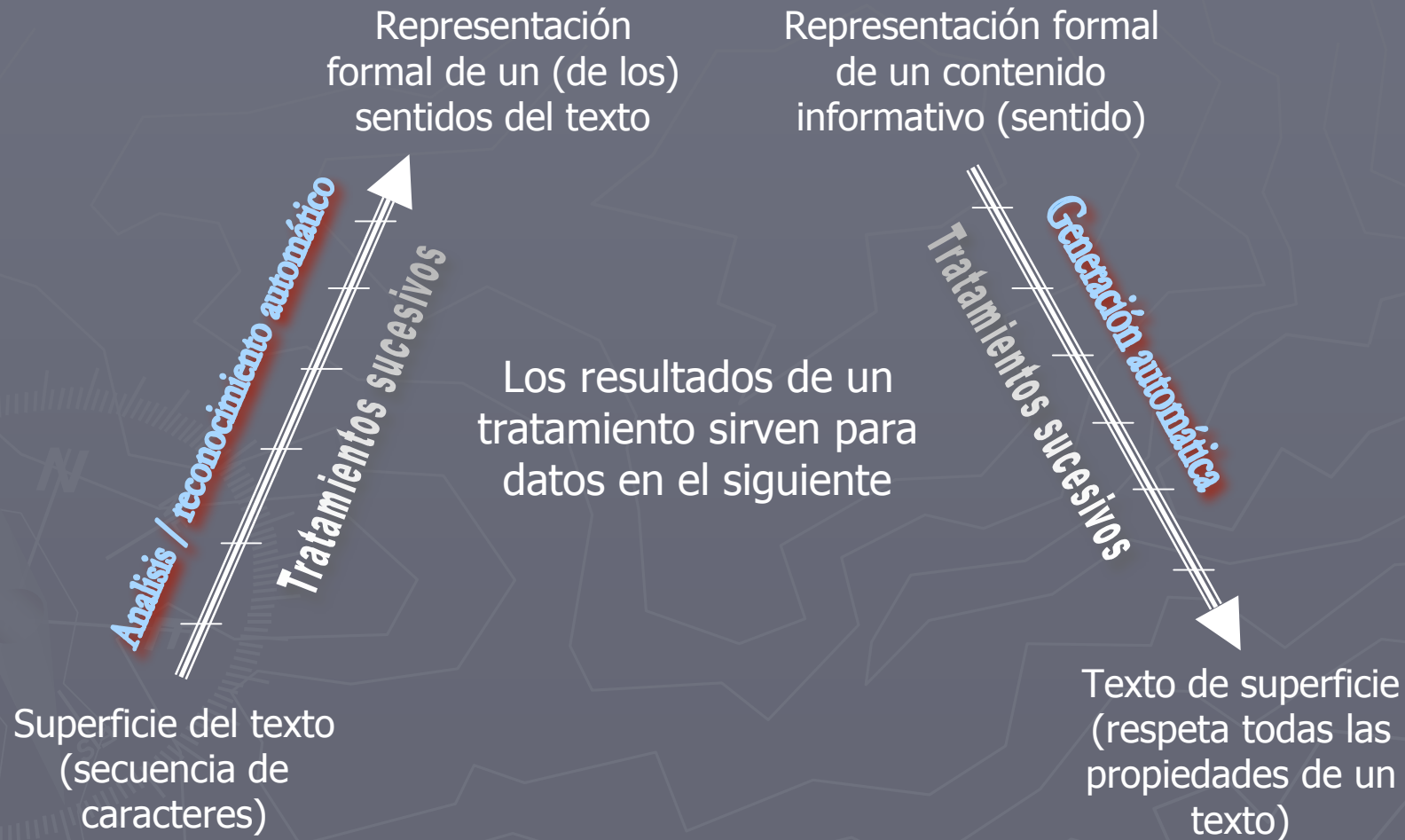
Definición et representación del sentido

más fácil : definir y representar aspectos del sentido

Ejemplo : (Hablaemos: verbo hablar, futuro, 1ra persona del plural).

Estos aspectos del sentido bastan para : frecuencia de las palabras de un texto, detección de errores, búsqueda de ciertas informaciones...

Tratamiento Automático de la Lengua



Reconocimiento automático de textos



Uno o varios sentidos?

Cual representación formal ?

Aplicaciones posibles despues de tratamiento

Generación automática de textos

Definición imperativa del punto de inicio

Las características del resultado buscado son predeterminadas

Casi ninguna aplicación para los resultados intermedios

Hay una manera de expresar una idea en la lengua ?

Sin lector, no existe el texto

Necesidad de anticipar la comprensión del texto por el (los) lector(es)

Representación formal
de un contenido
informativo (sentido)

Tratamientos sucesivos
Generación automática

Texto de superficie
(respeto todas las
propiedades de un
texto)

Lo entendí
todo



Tratamientos

Para cada tratamiento solo hay dos posibilidades :

► Utilización de diccionarios

- Un diccionario enumera todas las posibilidades para un nivel de tratamiento
- Ejemplo : lista de todas las formas del verbo "hablar"
- Definición de un lenguaje por extensión (enumeración de todos los términos del lenguaje)

► Utilización de gramáticas

- Una gramática es un conjunto finito de reglas
- Una gramática permite de construir todas las posibilidades para un nivel de tratamiento (a menudo un número casi infinito) y de averiguar si un término dado puede ser generado
- Ejemplo : gramática (reglas) de construcción de todas las formas del verbo "hablar"
- Definición de un lenguaje por intención (Definición de los procesos permitiendo el reconocimiento y la generación de cada termino del lenguaje)

► En general el tratamiento es mixto

Reconocimiento automático de textos

- Si fuéramos la computadora, un texto podría ser :

9Δ.∇İ.Δωρ▷▷▷Δ▷Δ▷Δ.Λ.Ç.▷Δρ

Posibilidades de la informática : hacer la lista de los caracteres utilizados y marcar las repeticiones

- Necesidad de otros conocimientos para “entender” más
- Utilización de conocimientos de la estructura y del funcionamiento de la lengua
 - Estos conocimientos tienen que estar formalizados y expresados con objetos lingüísticos que somos capaz de definir a este nivel (en este caso caracteres)

Recorte en formas

- ▶ Secuencia de caracteres → secuencia de formas
- ▶ Formas ligüísticas y non ligüísticas
 - números, símbolos matemáticos, químicos..., palabras de otro idioma, tablas, imágenes, formulas, etc.
- ▶ Definición de las formas por gramática ?
- ▶ Definición de las formas por diccionario ?
- ▶ Definición, por gramática, de lo que no es forma
 - Reglas para : , - ' . etc.
 - Problema con las siglas, abrevaciones, partes non lingüísticas
- ▶ Resultado : secuencia de formas
- ▶ Aplicaciones : estudio de textos, detección de errores ortográficas, aprendizaje de un idioma
 - Problemas debidos al no-conocimiento de la naturaleza gramatical de las forma

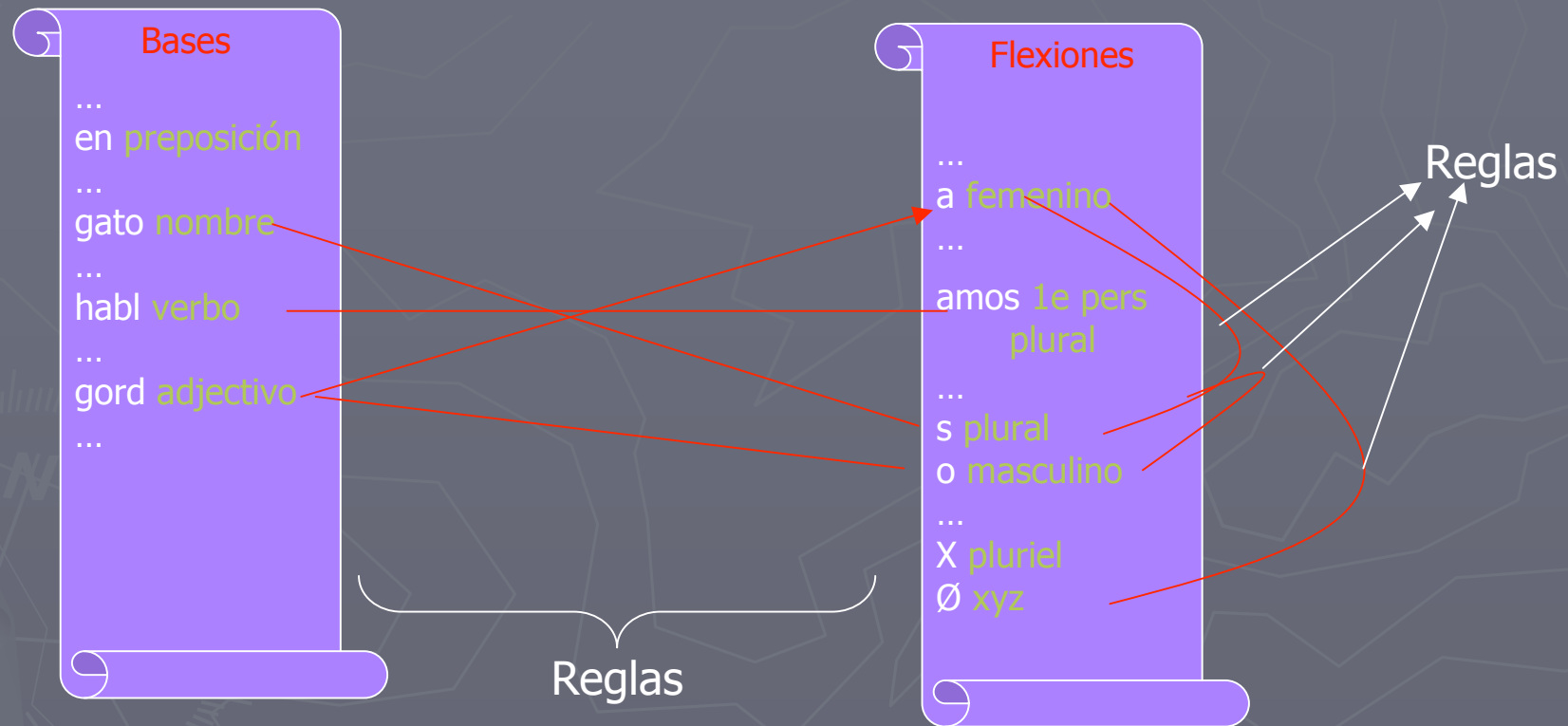
Análisis morfológico de textos

- ▶ Objetivo : asociar a cada forma su categoría gramatical y informaciones morfológicas (lema, genero, numero, tiempo, persona...)
- ▶ Necesidad y dificultad para definir la lista de las categorías (modelo morfológico)
 - Categoría de : si, no ?
- ▶ Vínculo entre numero de categorías , ambigüedad y precisión
 - Numero mínimo de categorías ? máximo?
 - más categorías, más precisión pero más ambigüedad

Métodos de análisis morfológico

- ▶ Con diccionario completo de (formas, categorías)
 - Problemas de construcción del diccionario (tamaño, elección de las informaciones)
- ▶ Trabajando con pedazos predefinidos de formas y reglas de combinación de estos pedazos
 - Base + flexiones : análisis morfológico flexional
 - Prefijos + raíces + postfijos : análisis morfológico derivacional

Análisis morfológico flexional



Análisis morfológico flexional : calcul

Formas analizadas : autoras

Se consideran todos los recortes Base + flexion posibles :

autoras, autoras, autoras, autoras, autoras, autoras

Solución cuando cada pedazo esta parte de los diccionarios y de las reglas permitiendo la concatenación.

Autoras, nombre, femenino, plural, autor

Otro ejemplo (en frances):

marches, marches, marches, marches, marches (*verbo, 2e
persona, presente, marcher*), marches (*nombre, femenino,
plural, marche*), marches

Análisis morfológico derivacional : calculo

Forma analizada : determinaciones

∅

determinaciones, determinaciones, determinaciones, determinaciones,
determinaciones, determinaciones...

d

eterminaciones, eterminaciones, eterminaciones, eterminaciones,
eterminaciones, eterminaciones...

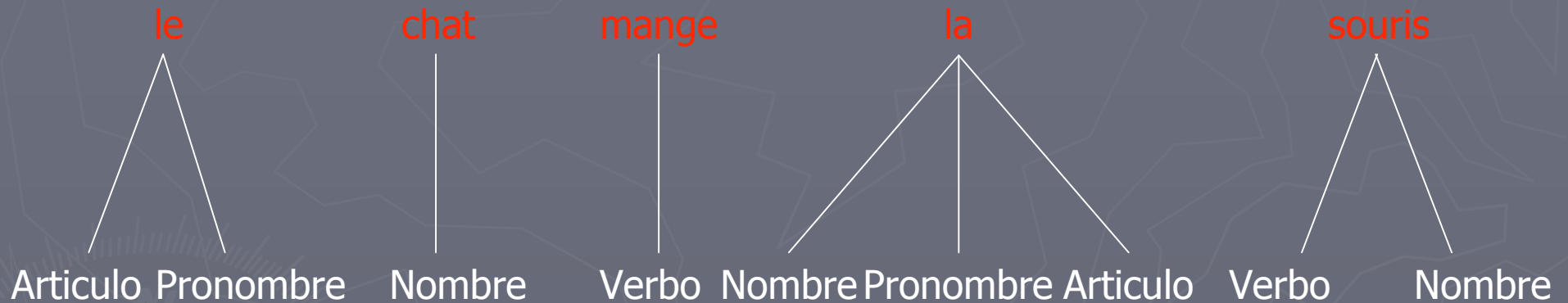
de

terminaciones, terminaciones, terminaciones, terminaciones,
terminaciones, terminaciones...

det ...

Resolución de las ambigüedades morfológicas

El resultado del análisis morfológico es ambiguo :



Para resolver las ambigüedades :

Utilización de reglas lingüísticas basadas en el contexto

Utilización de procedimientos estadísticos

Resolución de las ambigüedades morfológicas

Métodos lingüísticos

Reglas de determinación :

$$\text{Cont}_g + \textcolor{red}{X} + \text{Cont}_d \longrightarrow \text{Cont}_g + \textcolor{red}{C}_k + \text{Cont}_d$$

Con : $\text{Cont}_g = C_1C_2C_3\dots C_i$ et $\text{Cont}_d = C_1C_2C_3\dots C_j$

Estas reglas son contadas y difíciles para crear

Resolución de las ambigüedades morfológicas

Métodos lingüísticos

Reglas de exclusión :

$$\text{Cont}_g + \begin{matrix} C_k \\ C_l \\ C_m \\ C_n \\ \dots \end{matrix} + \text{Cont}_d \longrightarrow \text{Cont}_g + \begin{matrix} C_k \\ C_l \\ C_n \\ \dots \end{matrix} + \text{Cont}_d$$

Con : $\text{Cont}_g = C_1 C_2 C_3 \dots C_i$ y $\text{Cont}_d = C_1 C_2 C_3 \dots C_j$

Estas reglas son más fáciles para crear

A priori no se sabe si el proceso converge hacia la resolución total de las ambigüedades

Resolución de las ambigüedades morfológicas

Métodos estadísticos

Primera etapa : construcción de la tabla de transiciones de las categorías

1. Se considera cómo representativo (de la lengua, de un corpus) un texto dado y se lo analiza automáticamente
2. Se resuelven manualmente las ambigüedades
3. Se cuentan todos los dobletes (tripletes, cuádrupletes...) de categorías del texto
4. Se transforman las frecuencias de los dobletes (tripletes, cuádrupletes...) en probabilidades
5. Así se obtiene la probabilidad para cualquier doblete (triplete, cuádruplete...) de categorías morfológicas

Resolución de las ambigüedades morfológicas

Métodos estadísticos

Secunda etapa : aplicación de la tabla de transiciones al texto con ambigüedades



12 soluciones (caminos) posibles

Se calcula la probabilidad de cada camino y se escoge el camino lo más probable

Método basado en las probabilidades condicionales, conocido como *Cadenas de Markov*

Muy buenos resultados : 95 à 98% ; porque ? Naturaleza y consecuencias de los errores del método

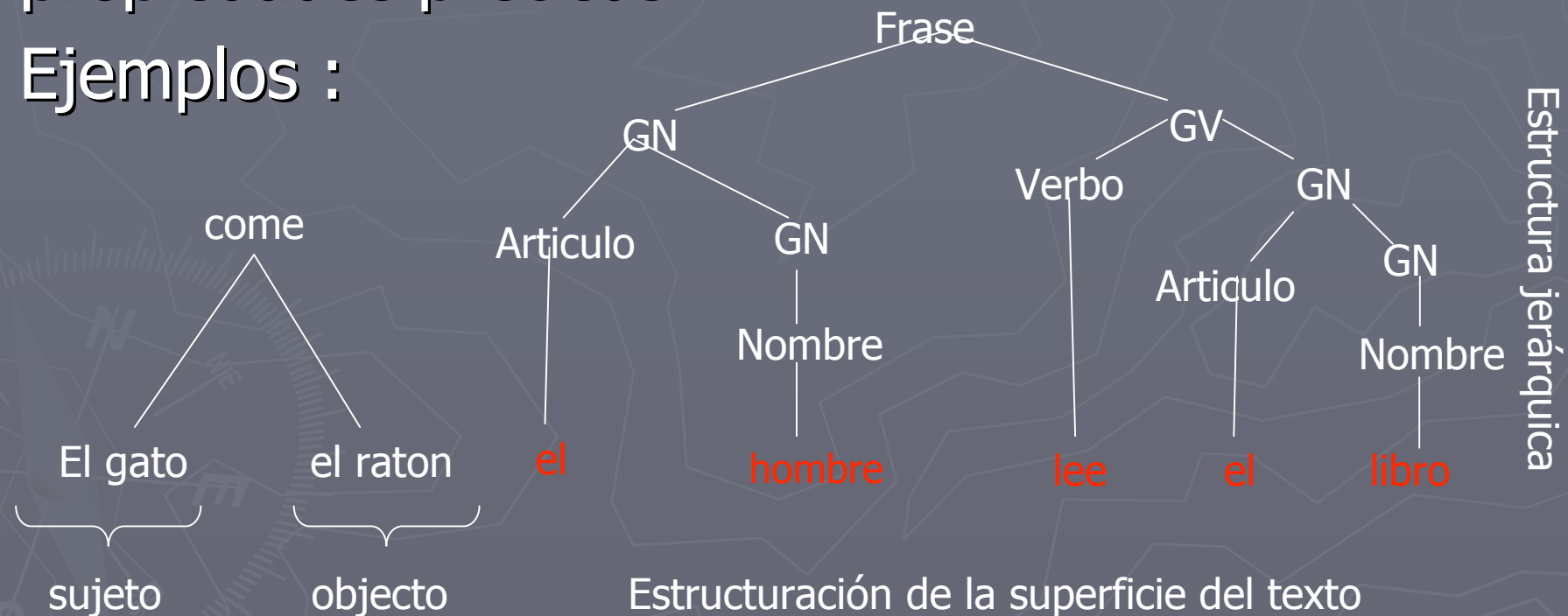
Análisis morfológico, aplicaciones

- ▶ Estudio de textos
 - Tenida en cuenta de los lemas
- ▶ Detección de los errores de ortografía y de gramática
 - Método usado actualmente por las herramientas de tratamiento de textos
- ▶ Búsqueda de informaciones
 - Tenida en cuenta de todas las variaciones morfológicas del término dado como criterio de búsqueda
- ▶ Aprendizaje de los idiomas
 - Creación automática de ejercicios a partir de cualquier texto
 - Indexación pedagógica de recursos
- ▶ Conjugadores
- ▶ Diccionarios de las formas de un idioma, diccionarios para el aprendizaje de un idioma
 - Ejemplo : existe un diccionario dando la significación de : está, trenes, diciéndome, tuviste, conducía... ?

Análisis sintáctico

Objetivo : Agrupar las palabras de una frase para constituir unidades homogéneas que tienen propiedades precisas

Ejemplos :



Análisis sintáctico

No se puede trabajar con diccionarios, se necesita trabajar con gramáticas

Una gramática es un conjunto de reglas que describe todas las combinaciones posibles para un idioma

La construcción de una gramática es un trabajo lingüístico especializado

Nos limitamos a la construcción de frases porque no sabemos como trabajar más allá

Una gramática utilice símbolos terminales (categorías morfológicas en general) y símbolos intermediarios (dichos non terminales, definidos en respecto al tipo de estructura que se quiere crear).

Le objetivo, la estructura de la frase, se llama axioma.

El análisis sintáctico genera muchas ambigüedades : una frase de 17 palabras produce 572 estructuras diferentes

Análisis sintáctico

Ejemplo de gramática

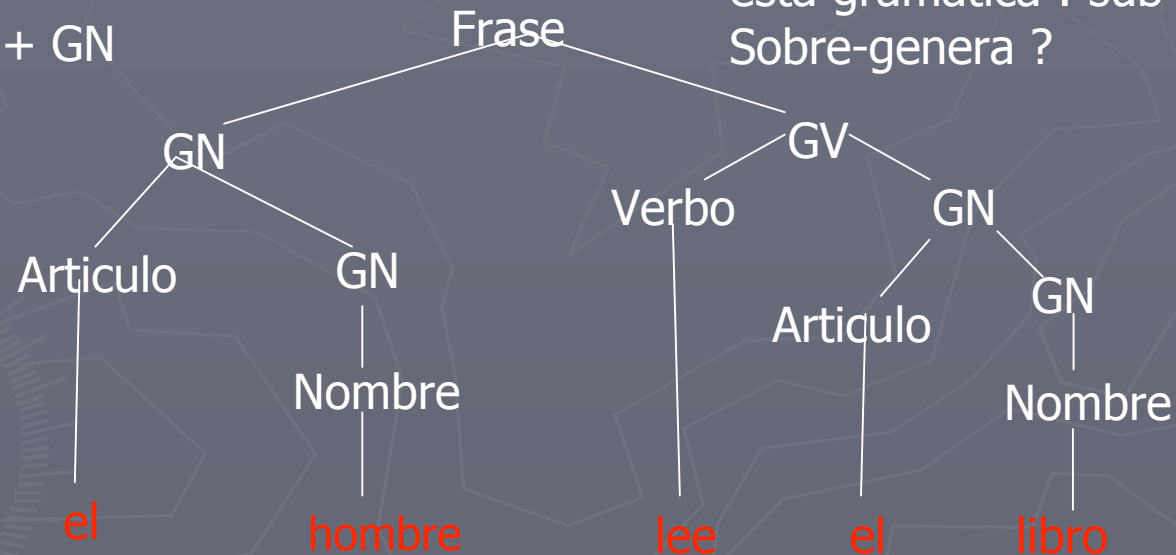
Frase $\longrightarrow$ GN + GV

GN $\longrightarrow$ Nombre / Artículo + GN

GV $\longrightarrow$ V + GN

La recursividad juega un papel importante

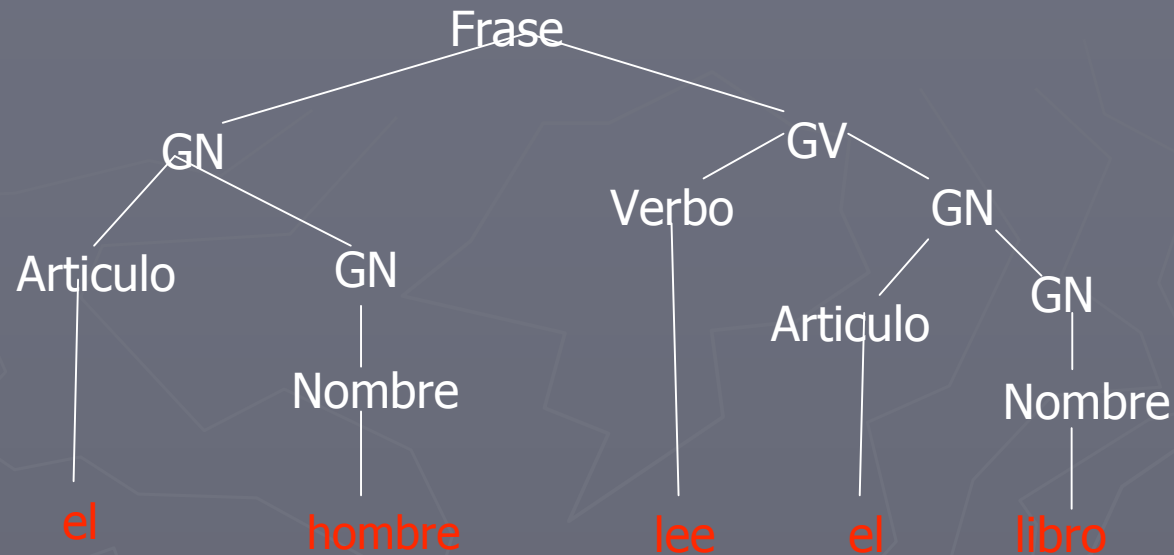
En respecto a la lengua española esta gramática : sub-genera ?
Sobre-genera ?



El turista come una guanábana; la chica mira la montaña; el pájaro sobrevuela la ciudad; ...

Análisis sintáctico

Ventajas / inconvenientes de la estructura arborescente



Ventajas :

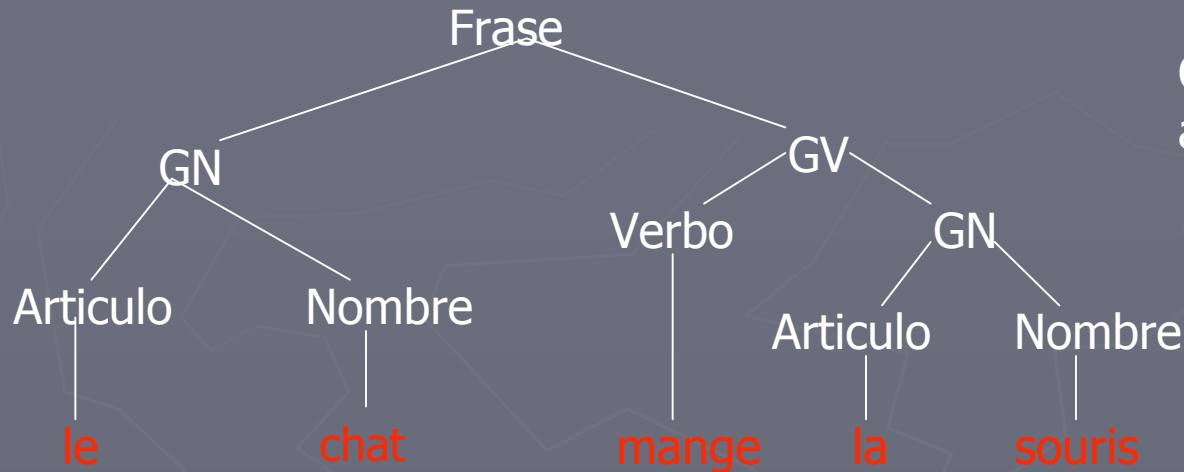
Se conserva el orden de la superficie del texto, mientras disponiendo des los grupos
La informática manipula bien los árboles

Inconvenientes :

No siempre esta posible conservar el orden de la superficie del texto sin transformaciones

El papel de los constituyentes no esta indicado

Análisis sintáctico dificultades



Cual es la gramática asociada a esta estructura ?

Cual es el árbol sintáctico de :

Le chat ne mange pas la souris

Le chat a bien mangé la souris

Le chat mange le fromage, la souris aussi

Le chat et la souris mangent le fromage vs le chat mange le fromage et la souris mange le fromage

Je l'ai entendu chanter ce matin (double rattachement)

Estrategias de análisis sintáctico

► Descendente

- Se empieza por el axioma de la gramática y aplicando reglas se intenta a llegar al texto

► Ascendente

- Se empieza por el texto y se trata de llegar al axioma de la gramática

Estrategias de análisis sintáctico

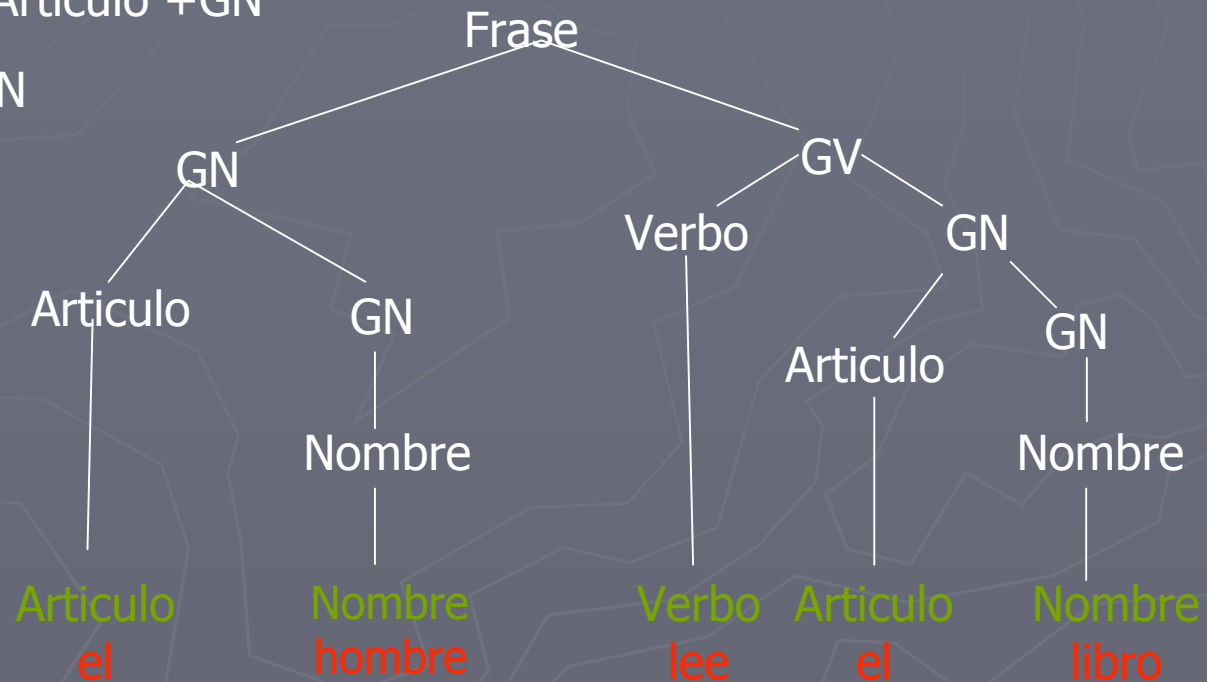
Análisis descendente

Frase $\longrightarrow$ GN + GV

GN $\longrightarrow$ Nombre / Artículo +GN

GV $\longrightarrow$ Verbo + GN

Problema de la recursividad



Se lo puede hacer también de la izquierda a la derecha

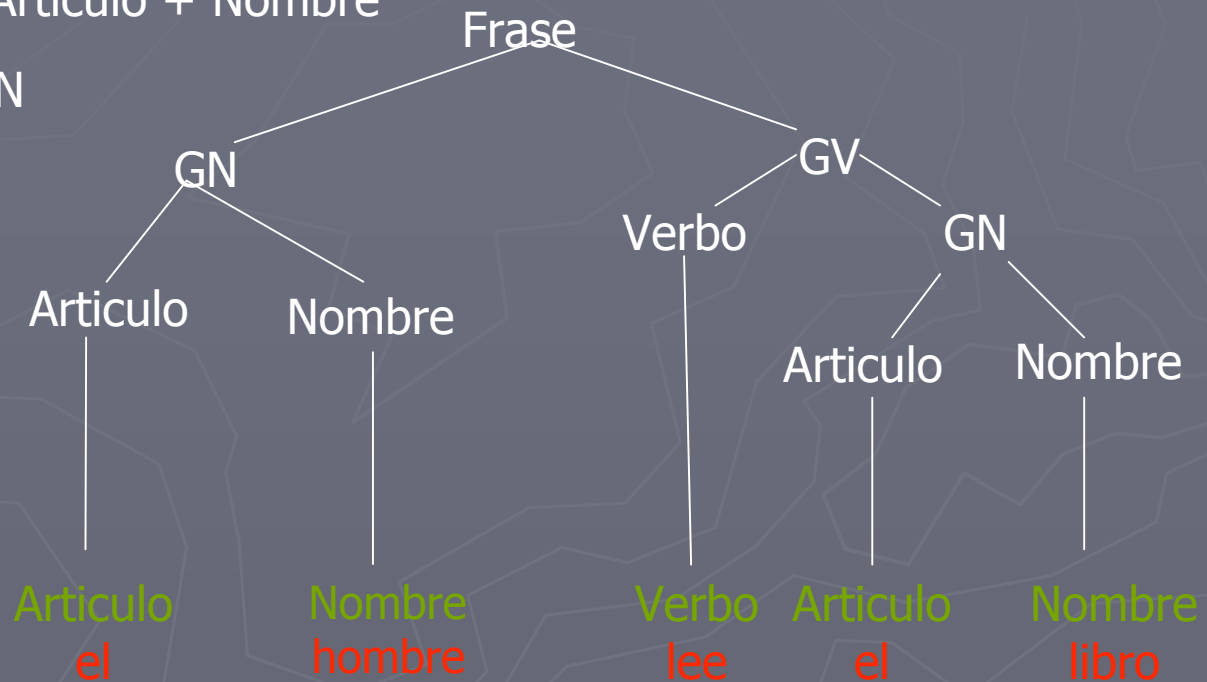
Estrategias de análisis sintáctico

Análisis ascendente

Frase → GN + GV

GN → Nombre / Artículo + Nombre

GV → Verbo + GN



Se lo puede hacer también de la izquierda a la derecha

Estrategias de análisis sintáctico

Ascendente

Descendente

De la izquierda a la
derecha

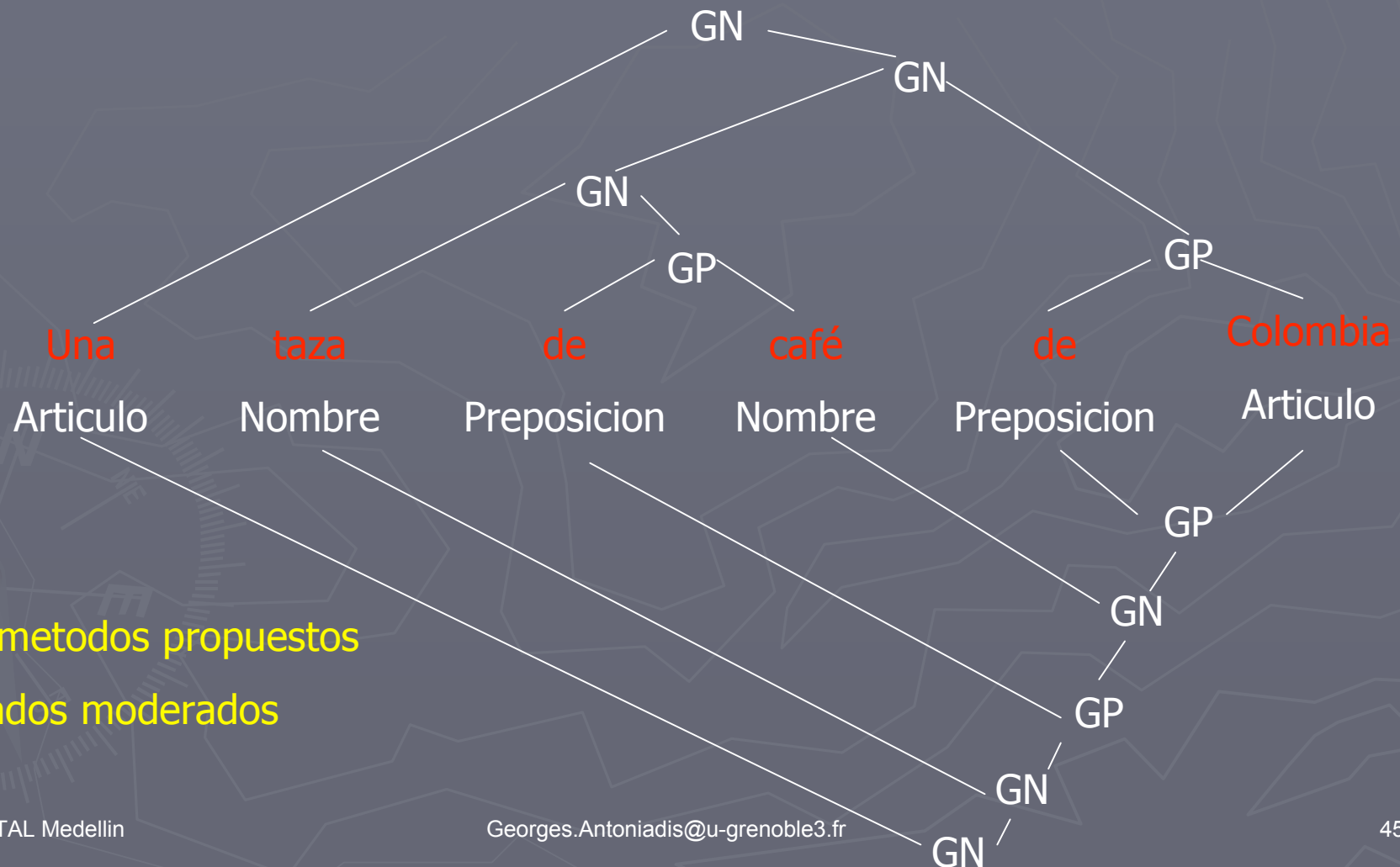
De la derecha a la
izquierda

4 estrategias posibles

La cual elegir ?

Análisis Sintáctico

Ambigüedades



Varios metodos propuestos

Resultados moderados

Análisis Sintáctico

Resultados

Actualmente, para textos cualquiera, los resultados (por lo menos una solución) están al nivel de 60 - 70%, máximo

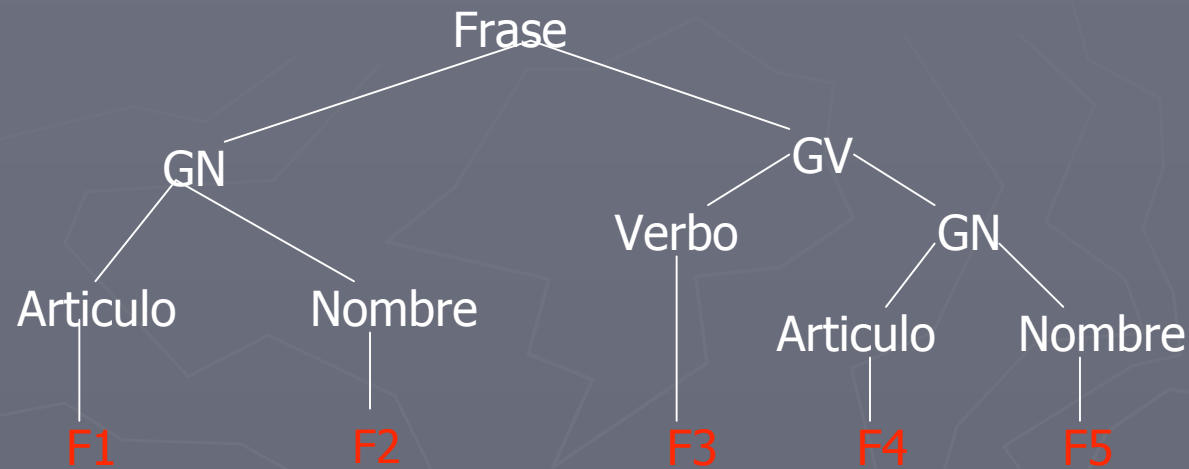
- Todas las soluciones no son exactas
- Numerosas ambigüedades

Resultados mucho mejores cuando :

- Los textos pertenecen a un campo cerrado : meteorología, ley, medicina...
- Diccionarios et gramáticas especializados
 - ▶ Resolución de ambigüedad a priori
 - ▶ Ejemplo : verbo volar

Análisis Sintáctico

Resultados



Como “entender” más ?

Conocer propiedades de elementos

¿ « Entender » frase por frase es suficiente para entender un texto ?

No, el caso típico es el del dialogo

Establecer conexiones entre los elementos del texto

Análisis semántico

Objetivo :

- A. Asociar a cada elemento de una frase las propiedades que lo caracterizan
- B. Establecer y explotar las conexiones entre los elementos de una frase / de un texto

A. Necesita diccionarios que proporcionan las propiedades

B. Necesita reglas :

1. Definiendo las conexiones
2. Explotando las conexiones

Análisis semántico

Semántica lexical

Definición del “sentido” de un elemento por características llamadas rasgos

- Rasgos binarios. Valores posibles : si/no
 - ▶ Ejemplo : animal, humano, animado, silbar, etc.
- Rasgos n-arios. más de dos valores posibles
 - ▶ Ejemplo : color (blanco, negro, amarillo, verde), peso(1, 2, 3, ... n gramas)

Análisis semántico

Semántica lexical

Mirlo animado: si; humano: no; animal: si; color: negro; peso: 80

Avocat1 (avocado) animado: si; humano: si; animal: si; color: blanco, negro, amarillo; peso: >45 kg

Avocat2 (aguacate) animado: no; humano: non; animal: no; color: verde; peso: <200 g

Quién elige los rasgos ? Cómo ? Con arreglo a que ?

Factabilidad de estos diccionarios ?

Análisis semántico

Conexiones entre palabras

Definir y representar relaciones que pueden existir entre dos terms

Ejemplo de relación : hiperonimia, hiponimia, parte de, tipo de, sinonimia, consecuencia de, etc.

Ejemplos : rueda *parte de* carro, tabla *hiponimo de* mueble, mirlo *tipo de* pájaro, etc.

Creación de redes semánticas

Análisis semántico

Redes semánticas

Red semántica la más antigua

Siglo 3 antes de Cristo, Porphyre (Porfirio), filósofo griego, comentario de las categorías de Aristote

Método de definición de categorías

- Genus (genero, tipo general)
- Differentiae (diferencial)

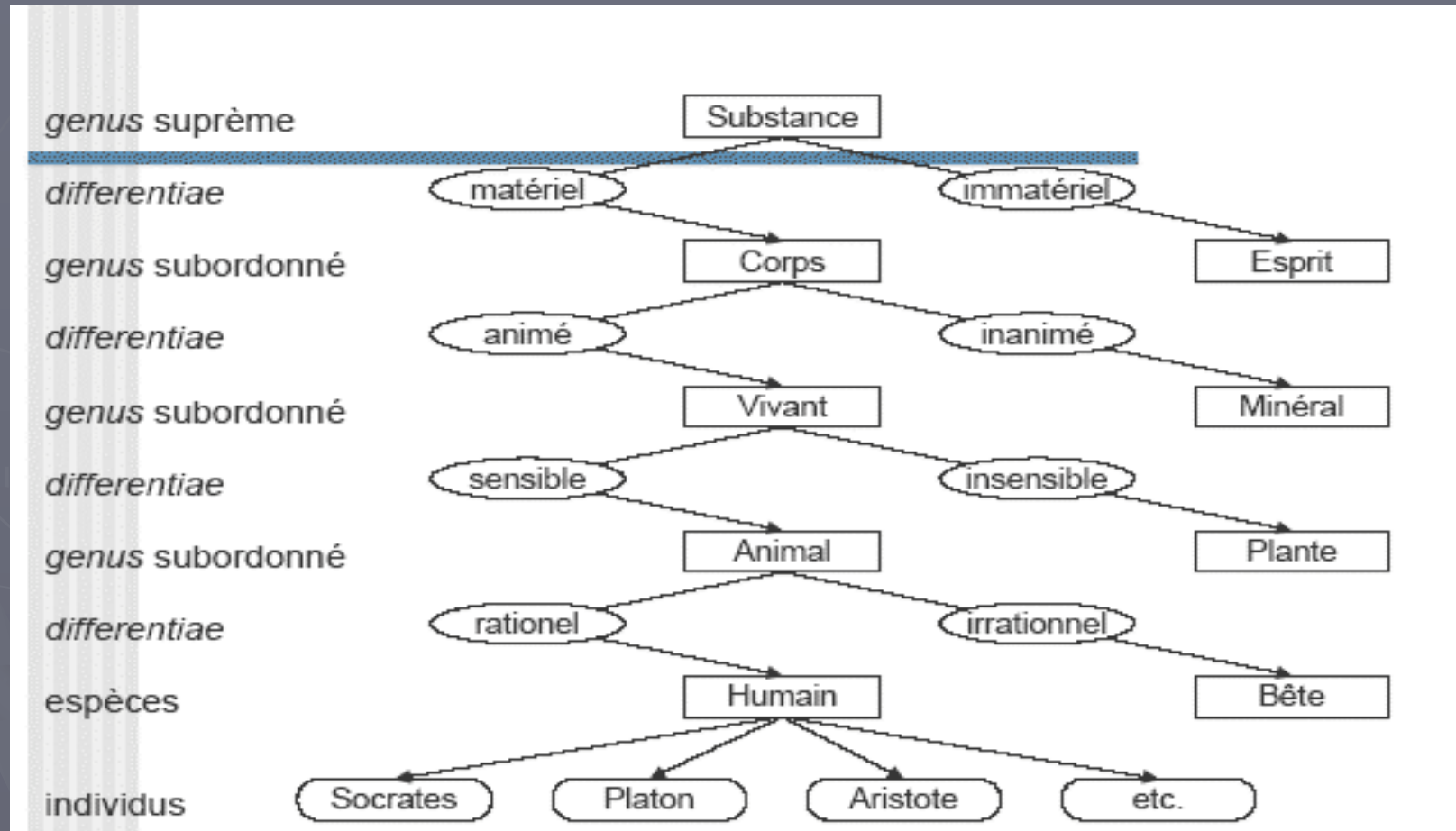
Reglas de herencia resultante del silogismo Aristotelico : subtipos heredan las propiedades de los supertipos

Distinción entre categorías (rasgos) et individuales (elementos)

Base de varias redes actuales y de la programación orientada objeto

Análisis semántico

Redes semánticas

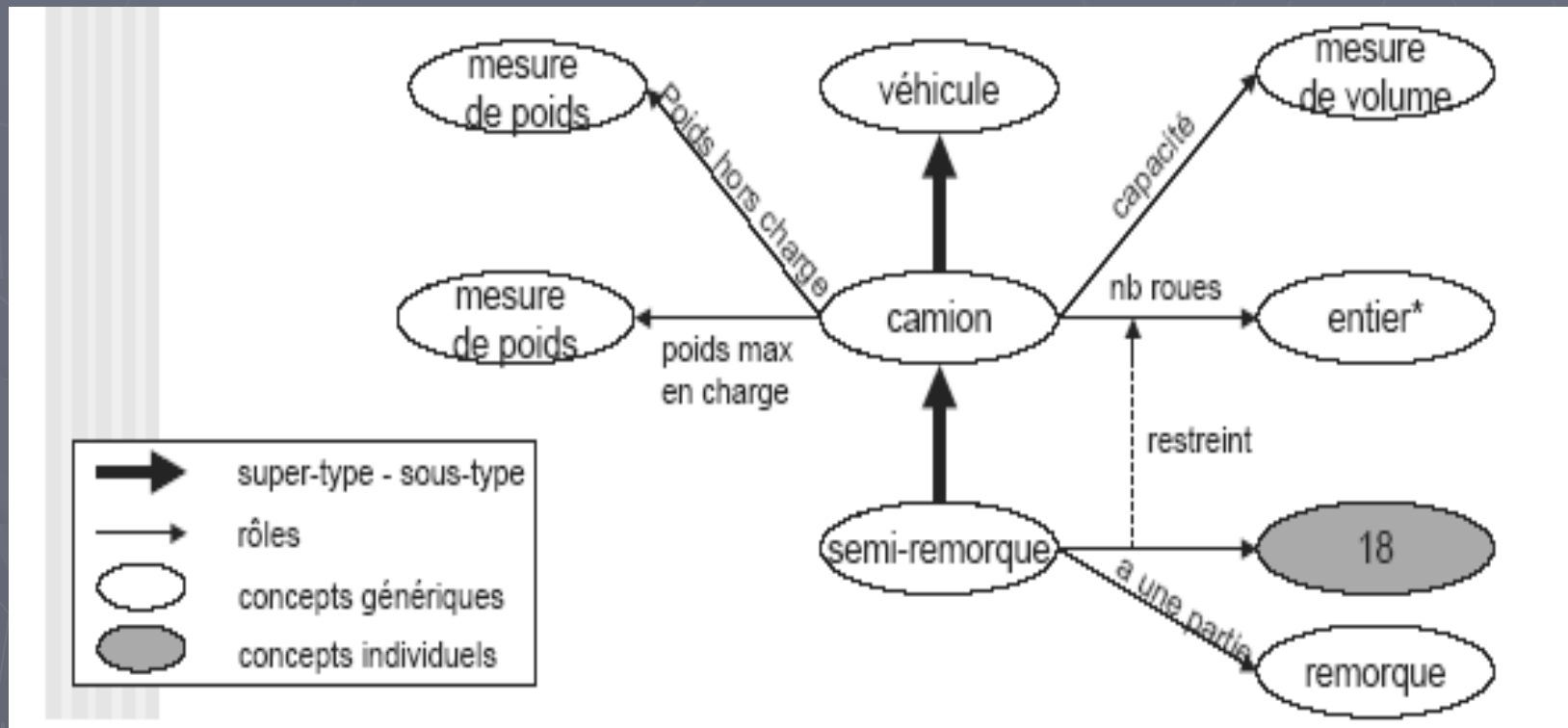


« Arbol de Porfirio » dibujado por Pierre d'Espagne (1329)

Análisis semántico

Redes semánticas

Reutilización mejorada del árbol de Porfirio por Woods (1975) y Brachman (1979, Knowledge Language One)



Análisis semántico

Redes semánticas

Aparte las redes descritas (definicionales) otros tipos de redes existen : implicacionales, ejecutables, aprendantes, etc.

Quién elige los rasgos ? Cómo ? Con arreglo a que ?

Factabilidad de estas redes ?

Muy pocos sont operacionales, principalmente redes definicionales.

Análisis semántico

Interpretación de un texto depende mucho de las informaciones contenidas en los diccionarios y las redes semánticas

- más informaciones, problemas de factabilidad
- Menos informaciones, problemas de interpretación

No existen herramientas semánticas generales, o son muy pobres.

No existen analizadores semánticos de textos cualquiera.
Tiene sentido ?

Herramientas o analizadores semánticos para problemas muy precisos o campos muy cerrados existen y funcionan.

Extracción y puesta en relación de objetos de un texto (tesauros, ontologías)

Verificación de las propiedades definidas para los objetos de un textos : diagnosis medicales, textos de derecho, etc. a veces asociados a sistemas de inferencia

Generación automática de textos - GAT

Definición

Proceso de elaboración,
a partir de una representación
formal de un contenido
informativo,
de un texto restituyendo a cada
uno de sus destinatarios (lectores)
este contenido informativo, **con un
objetivo dado**

Cours TAL Medellín

Georges.Antoniadis@u-grenoble3.fr

Representación formal
de un contenido
informativo (sentido)

Tratamientos sucesivos
Generación automática

Texto de superficie
(respeto todas las
propiedades de un
texto)

Lo entendí
todo

Generación automática de textos

Ejemplos de aplicaciones

- ▶ Producción automática o semi-automática de informes, correos, resumen, artículos de agencias de prensa...
- ▶ Traducción automática
- ▶ Interrogación de bases de datos (formulación de respuestas)
- ▶ Aprendizaje de idiomas
- ▶ Ayuda a la redacción
- ▶ Simulación del proceso cognitivo
- ▶
- ▶ Todo programa produciendo resultados de los que el destinatario es humano

Generación automática de textos

Ejemplos de aplicaciones

Pero también :

Comme l'homme multiple le peuple divers
l'insecte varié le peuple divers dans toute la ville
l'enfant noir la moire changeante l'enfant jaune
dans nos villages la vie changeante
je t'aime dans les forêts comme partout
dans les forêts le temps changeant
dans nos rues l'amitié la femme
je t'aime sous tous les cieux eau des fleuves
enfant jaune dans nos rues
comme tu es je t'aime comme tu es comme

Cómo el hombre multiple el pueblo diverso
el insecto variado el pueblo diverso en toda la ciudad
el niño negro el muaré cambia el niño amarillo
dentro nuestros pueblos la vida cambia
te quiero el las selva cómo en todas partes
en las selvas el tiempo cambia
en nuestras calles el amistad la mujer el niño rojo
bajo todos los cielos, agua de ríos
niño amarillo en las calles
cómo eres te quiero cómo eres cómo

Generación automática de textos

Ejemplos de aplicaciones

O :

La campagne est harmonieuse, c'est ainsi. Parce qu'elle paraît
être mondialement harmonieuse, se déplaient des ar invisibles
qui influent des forces invisibles qui influent lentement sur les
êtres.

Pas-en-Artois parecía una pequeña ciudad armoniosa. Sin
embargo, Artois semblait une petite ville harmonieuse. Pourtant...

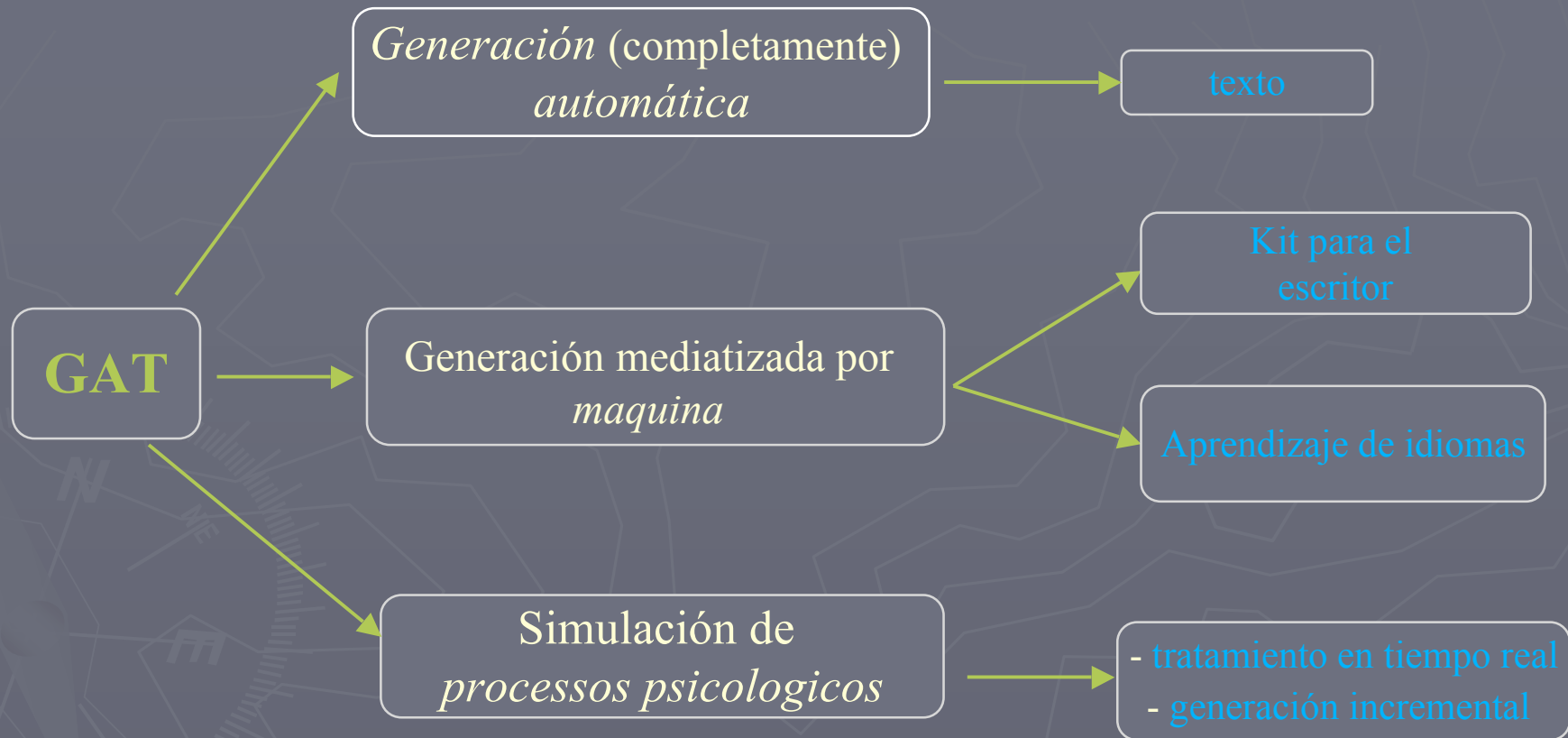
Soudain, cette soirée là, le printemps fut particulièrement feja,
les pinsons gazouillaient partout.

.....

A veces más fácil escribir poemas o historias : no objetivo
preciso, no información precisa a comunicar

Generación automática de textos

Aplicaciones



Generación automática de textos

Histórico

Principio de los años 60 hasta 1975

- expresión de una **forma** correcta
- la lengua (y no la comunicación),
- gramática formal (gramática generativa)
- generación aleatoria de frases (desprovistas de sentido)

De 1975 hasta 1980

- expresión de un contenido informativo
- texto de una frase

De 1980 hasta 1990

- expresión de un contenido informativo con un objetivo
- unas cuantas frases

Desde 1990

- se intenta a hacerlo mejor

Generación automática de textos

Problemática

4 problemas esenciales :

- Definición del (de los) formalismo(s) y representaciones formales para la entrada del sistema
- Elección de una versión
- Tenida en cuenta del destinatario
- Definición y arquitectura de los sistemas de generación

Generación automática de textos

Entrada del sistema

Búsqueda del Grial

A falta de Grial, se contentara de vasos de plástico o de copas de vidrio

Datos brutos : sistemas ANA, RAREAS, comunicado de l'AFP (Danlos, 84 et après)...

Tratamiento de datos, representaciones formales

Generación automática de textos

Entrada del sistema

domaine : **bourse**, ANA (Kukich, 1983)

► DOW JONES INDUSTRIALS AVERAGE - 01/12/83

► 01/12	CLOSE	30	INDUS	1083.61
► 01/12	330PM	30	INDUS	1089.40
► 01/12	3PM	30	INDUS	1093.44
► 01/12	230PM	30	INDUS	1100.07
► 01/12	2PM	30	INDUS	1095.38
► 01/12	130PM	30	INDUS	1095.75
► 01/12	1PM	30	INDUS	1095.84
► 01/12	1230PM	30	INDUS	1095.75
► 01/12	NOON	30	INDUS	1092.35
► 01/12	1130AM	30	INDUS	1089.40
► 01/12	11AM	30	INDUS	1085.08
► 01/12	1030AM	30	INDUS	1085.36
► 01/12	CLOSE	30	INDUS	1083.79

► CLOSING AVERAGE 1083.61 DOWN 0.18

Generación automática de textos

Entrada del sistema

domaine : *bourse*, ANA (Kukich, 1983)

After climbing steadily through most of the morning, the Stock Market was pushed downhill late in the day. Stock prices posted a small loss, with the indexes turning in a mixed showing yesterday in brisk trading.

The Dow Jones average of 30 industrials surrendered a 16.28 gain at 4pm and declined slightly, finishing the day at 1083.61, off 0.18 points.

Generación automática de textos

Entrada del sistema

domaine : *météo*, RAREAS (Kittredge et al., 1986)

- ▶ *fecha*
- ▶ *lugar*
- ▶ *velocidad y dirección del viento*
- ▶ *temperatura*
 - etc.
- ▶ Un conjunto de operadores lógicos
 - "." quiere decir "luego"
 - "-" quiere decir "o"
 - etc.

Generación automática de textos

Entrada del sistema

domaine : météo, RAREAS (Kittredge et al., 1986)

Weather forecast for frobisher bay issued at 10 o' clock mdt
monday evening 16 october 1989 winds 30 mps strong
breeze south-westerly strengthening to near gale force
veering to north-westerly later strengthening to gale force 8
veering to northerly tomorrow weather rain then heavy
becoming sleet or snow temperatures 0 falling to -3

Generación automática de textos

Entrada del sistema

Varios tipos de formalismos utilizados

- Redes
- Lógica, predicado / argumentos; $X \text{ comer } Y \rightarrow \text{comer}(X, Y)$
- Árboles (Danlos, 84), Árboles adjuntos (TAG : Tree Adjoining Grammars)
- Grafos : grafos conceptuales (Sowa), definición 1984 y después, utilización fin años 80 et años 90 (Nogier 93, Ponton 96)

Pregunta : ¿ Se usa el mismo formalismo para el sistema entero ?

Generación automática de textos

Elección de una versión

Características de una buena versión

- Correcta en respecto a la ortografía, la sintaxis, la semántica y las referencias al mundo
- Coherente (hilo conductor del relato)
- Cohesivo
- Non ambiguo para el lector

Generación automática de textos

Elección de una versión, coherencia



Generación automática de textos

Elección de una versión, cohesión

Question : ¿ Donde se ubica Uruguay ?

- (1) The location of Uruguay is south america.
- (2) The latitude ranges from -30 to -35 degrees.
- (3) The longitude ranges from -53 to -58 degrees.
- (4) The northern and eastern bordering country is Brazil.
- (5) The western bordering country is Argentina.
- (6) The boundary is the uruguay river.
- (7) The southeastern coast is the atlantic ocean.
- (8) The southern coast is the Rio de la Plata.

Uruguay lies in **South America** between -30 to -35 degrees **latitude** and -53 to 58 degrees **longitude**. **Its neighbour** countries are Brazil in the **north** and the east, and Argentina in the **west**. **The boundary** between the two of them is the Uruguay river. Uruguay's **natural borders** are the atlantic ocean in the **south-east** and the Rio de la Plata in the **south**.

- (1-3) **posición** (latitud & longitud)
(4-5) **país próximo** (nordeste & oeste)
(6-8) **fronteras naturales** (ríos y costas)

Generación automática de textos

Tenida en cuenta del destinatario

“Complazcame *y* vayase.”

“Me preguntaba cual era el sentido de este “y”, si unía los términos de una *equivalencia* o una *cronología*, elección que implicaría, por mi parte, dos comportamientos netamente diferentes. Suzanne se acercó y me quitó la ropa y las dudas”

Michel Rio, Alizès

Generación automática de textos

Tenida en cuenta del destinatario

Objetivo : Comprensión completa y sin ambigüedades

Entender todo el contenido informativo y solo el contenido informativo

Tenida en cuenta de los conocimientos y de las competencias del (de los) destinatario(s)

- Conocimientos y competencias lingüísticos
 - ▶ Palabras y estructuras más o menos simples
- Conocimientos y competencias en relato con el dominio de aplicación
 - ▶ Términos o expresiones más o menos especializados, necesidad de explicaciones
- Enlace entre conocimientos / competencias y idioma ?
 - ▶ El problema no se plantea si el caso no es en la lengua materna

Destinatario que sabe todo (experto) vs destinatario que no sabe todo (en respecto con la aplicación posible, pero con la lengua ?)

Generación automática de textos

Tenida en cuenta del destinatario, ejemplo

Shakespeare escribió Hamlet.

Hamlet fue escrito por Shakespeare.

El autor de Hamlet es Shakespeare.

Shakespeare es el autor de Hamlet.

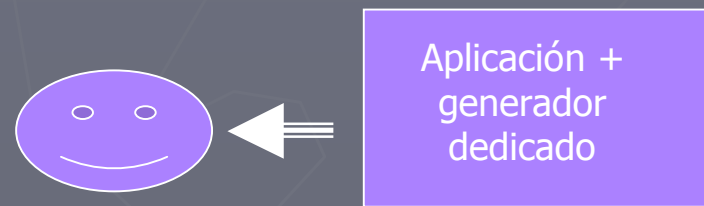
¿ Son las cuatro frases equivalentes :

- para todo lector que sea su nivel de conocimiento de la lengua ?
- para todo lector que sea su nivel de conocimiento de la literatura ?

Generación automática de textos

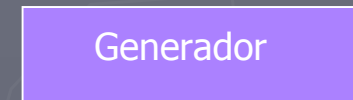
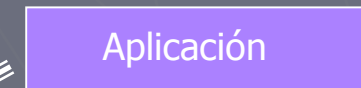
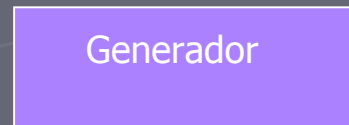
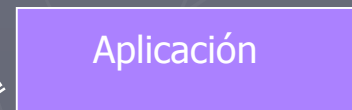
Tenida en cuenta del destinatario

Sistemas actuales



Sistemas mono-aplicación

Llegar a :



Sistemas no dedicados

Sistemas núcleos

Generación automática de textos

Definición y arquitectura de los sistemas

Que decir ?

plan global (texto)
plan local (frase)

Generación
profunda

Cómo decirlo ?

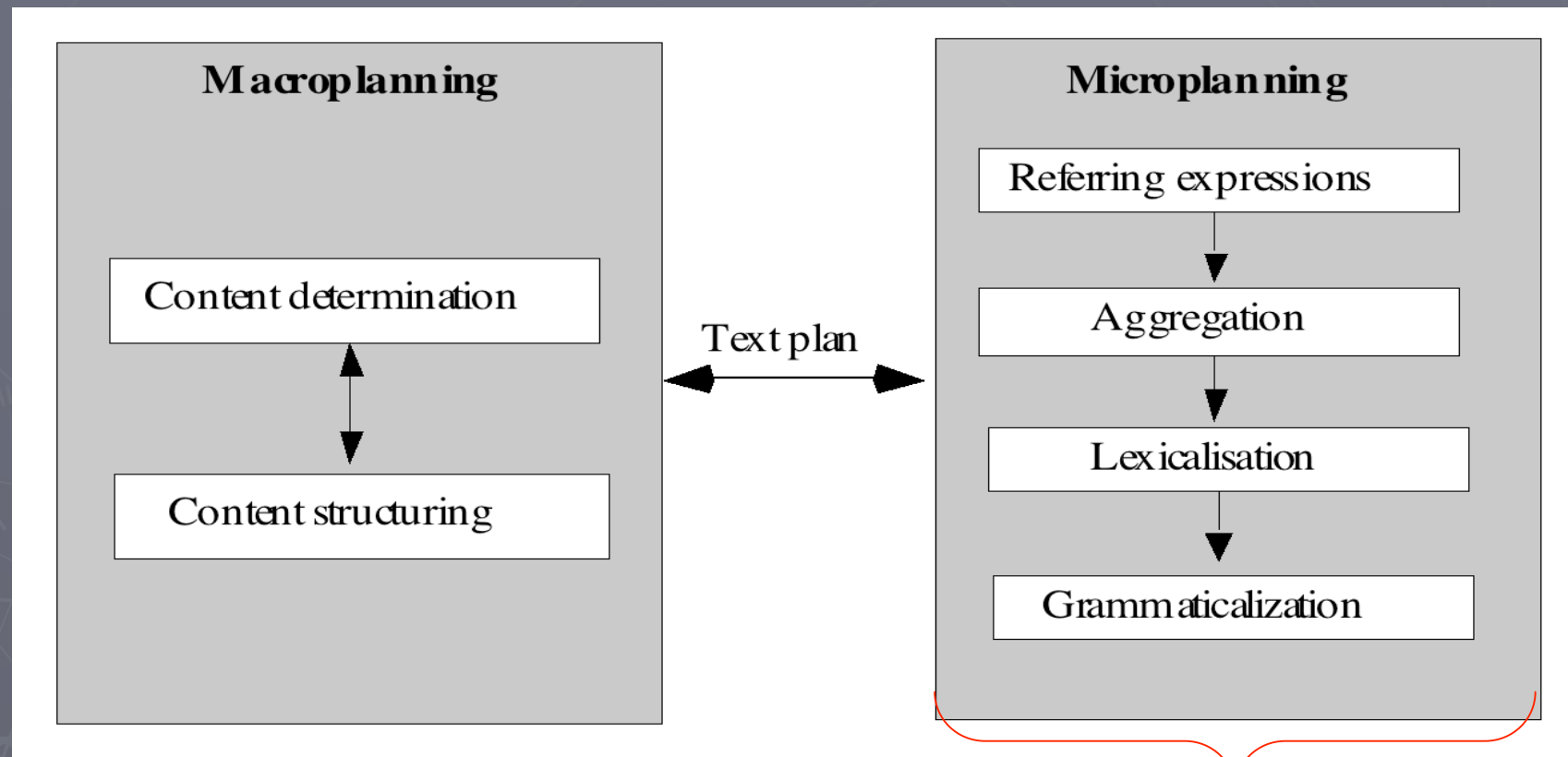
Realización lingüística
Presentación física

Generación
de superficie

Generación automática de textos

Definición y arquitectura de los sistemas

Planificaciones global y local

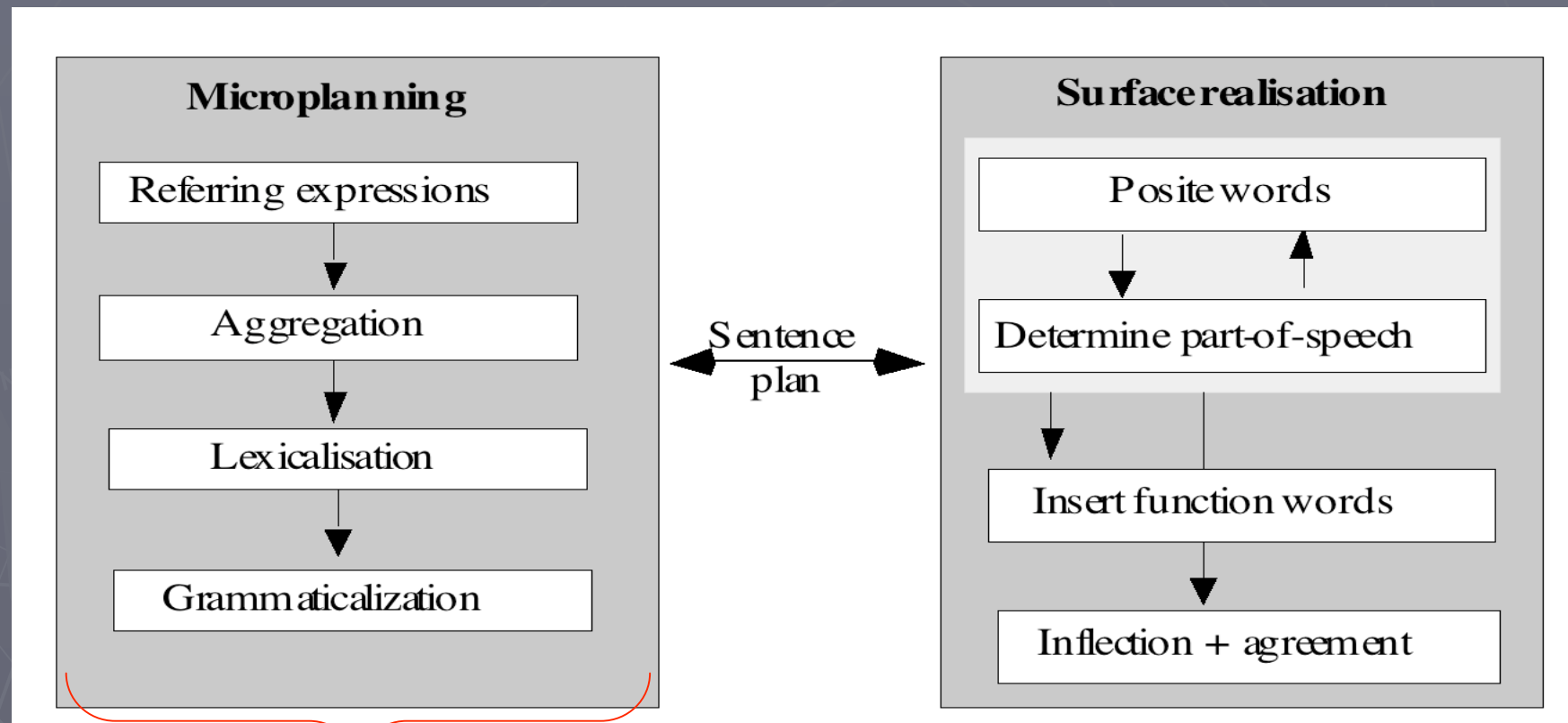


Determinar la elección

Generación automática de textos

Definición y arquitectura de los sistemas

Realización de la forma de superficie

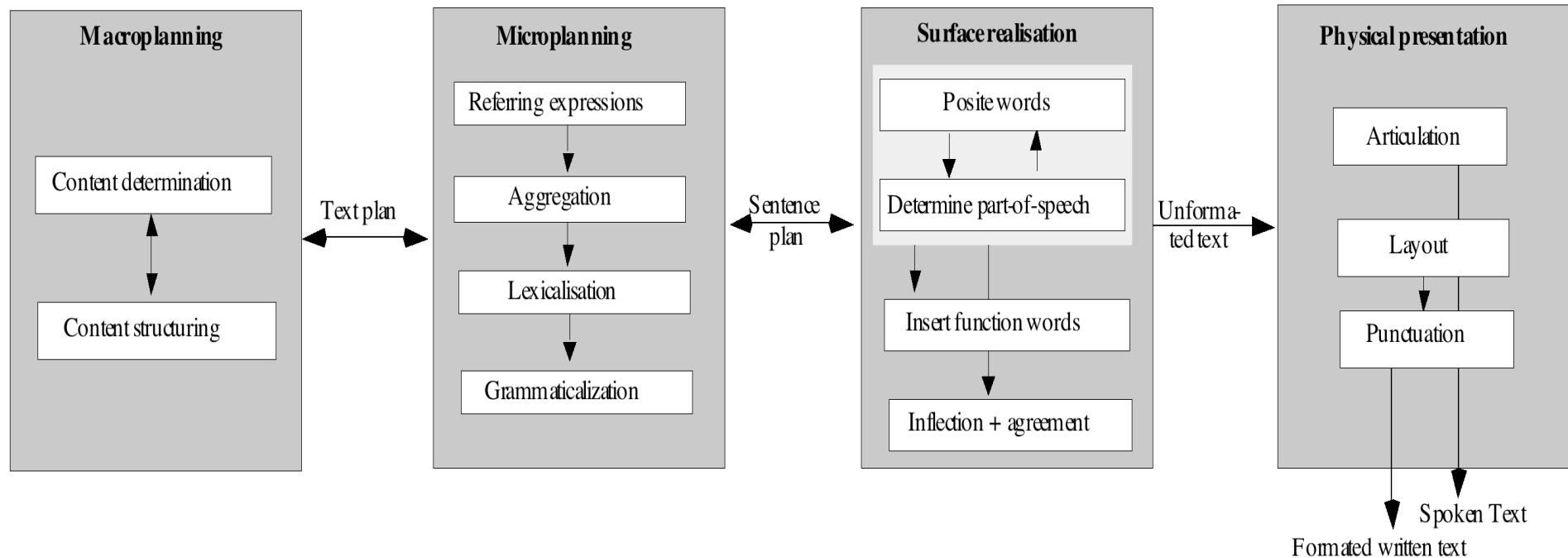


Realizar las elecciones

Generación automática de textos

Definición y arquitectura de los sistemas

[goals + discourse situation] + [knowledge base + plan library + ontology + user model] + [dictionary + grammar] + [discourse history]



Generación automática de textos

Definición y arquitectura de los sistemas

El numero de componentes esta maximal; en general los sistemas solo tienen una parte de estos.

La composición supone que se obtiene una sola versión posible; eso es teoría, en practica a menudo hay varias, con elección aleatoria o definida, la primera que esta construida.

Observación importante

En generación solo se puede implementar tratamientos decididos a priori y bien controlados; contrariamente al análisis no hay imprevisto, todo puede ser previsto al inicio.

En un sistema de generación, todo esta unido, volver sobre una elección a menudo supone la modificación de elecciones hechas antes.

Generación automática de textos

Definición y arquitectura de los sistemas

Nuestro enfoque en el LIDILEM

Generación núcleo

- Generador utilizable para varias aplicaciones
- Generador tomando como parámetros el modelo del usuario y del aplicación

Micro-planificación para cada párrafo con ayuda de operaciones lingüísticas

El modelo del usuario y del aplicación sirven cómo filtro para eliminar elecciones o versiones.

Generación automática de textos

Para ir más allá

<http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.htm>



Welcome to the **fourth** major release of John Bateman and Michael Zock's list of Natural Language Generation Systems...

Languages covered
[Sorted lists...](#)
[Systems by name](#)
[Systems chronologically](#)

Application domain:
[Systems by author](#)
[Systems reverse chronologically](#)

Most Recent Updates ? unclaimed!

Many thanks to those of you who continue to improve, correct and add to the information in this list. We are still checking and refining entries. We are aiming to make this list not only **COMPLETE** but also a **repository for accessible information concerning all the systems it contains...** Information concerning over 340 Natural Language Generation systems (somewhat loosely defined) is currently available. Additions/corrections are therefore **welcome** as ever.

We encourage contributors to provide informative short paragraphs concerning their (or any one else's) systems, which may also contain further hyperlinks, images, etc. Online bibliographical entries are also appropriately linked in to provide instant access. This release also contains some overviews of the data, which also reveal some gaps in the information that is available. Two overviews are given (clickable above): the **languages** generated and the **application domains** in which systems are active. If your system is not properly represented, please let us know!

Three basic instances of the list are maintained:

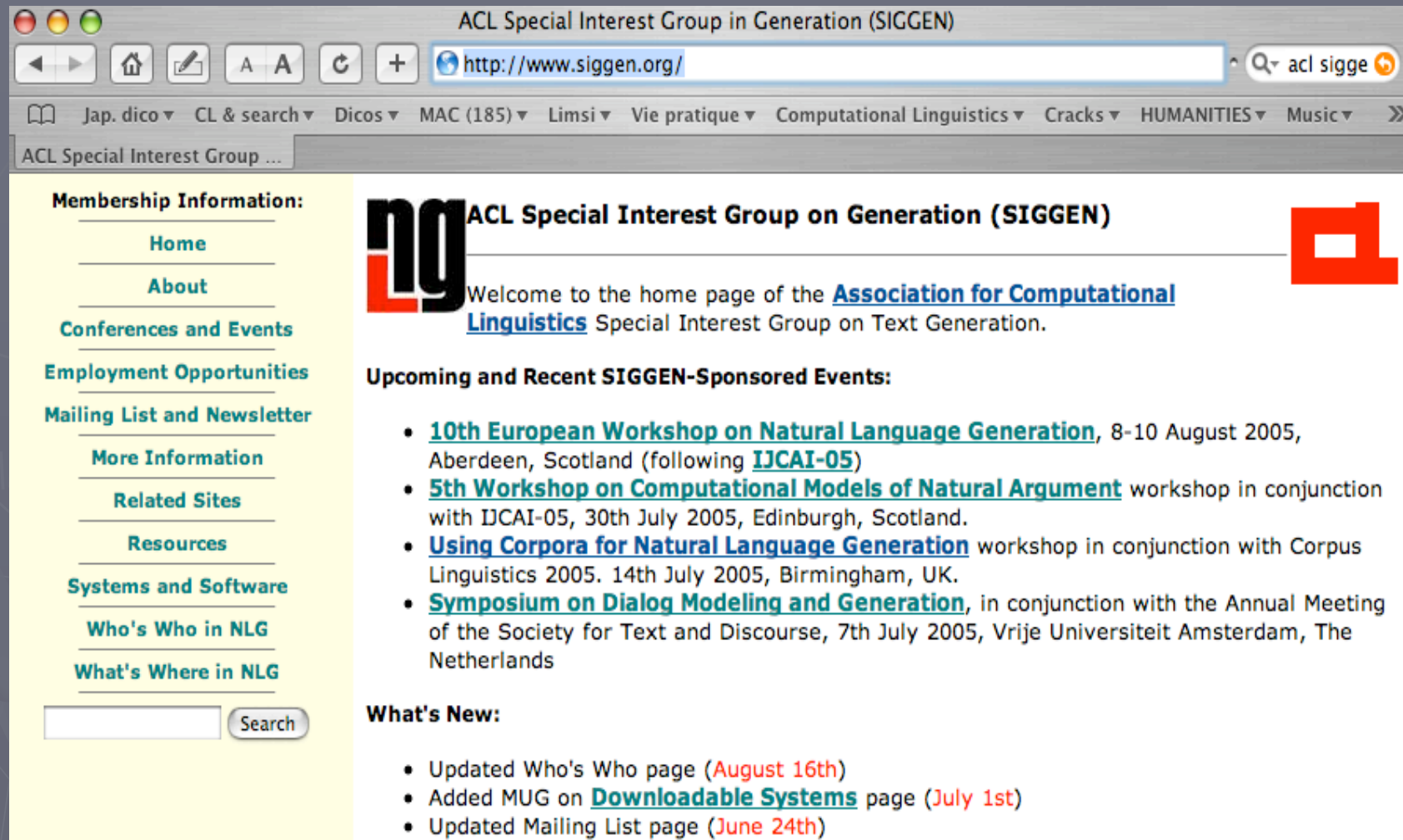
- [one sorted by system name](#), and
- [one sorted by date chronologically](#),
- [one sorted by date reverse chronologically](#).

There is also an index of the names of system builders mentioned in the tables available [here](#).

Sitio que contiene informaciones sobre los sistemas o publicaciones

Generación automática de textos

Para ir más allá



ACL Special Interest Group in Generation (SIGGEN)

http://www.siggen.org/

ACL Special Interest Group on Generation (SIGGEN)

Welcome to the home page of the [Association for Computational Linguistics](#) Special Interest Group on Text Generation.

Upcoming and Recent SIGGEN-Sponsored Events:

- [10th European Workshop on Natural Language Generation](#), 8-10 August 2005, Aberdeen, Scotland (following [IJCAI-05](#))
- [5th Workshop on Computational Models of Natural Argument](#) workshop in conjunction with IJCAI-05, 30th July 2005, Edinburgh, Scotland.
- [Using Corpora for Natural Language Generation](#) workshop in conjunction with Corpus Linguistics 2005. 14th July 2005, Birmingham, UK.
- [Symposium on Dialog Modeling and Generation](#), in conjunction with the Annual Meeting of the Society for Text and Discourse, 7th July 2005, Vrije Universiteit Amsterdam, The Netherlands

What's New:

- Updated Who's Who page ([August 16th](#))
- Added MUG on [Downloadable Systems](#) page ([July 1st](#))
- Updated Mailing List page ([June 24th](#))

Sitio del grupo SIGGEN en el contexto de la ACL

Generación automática de textos

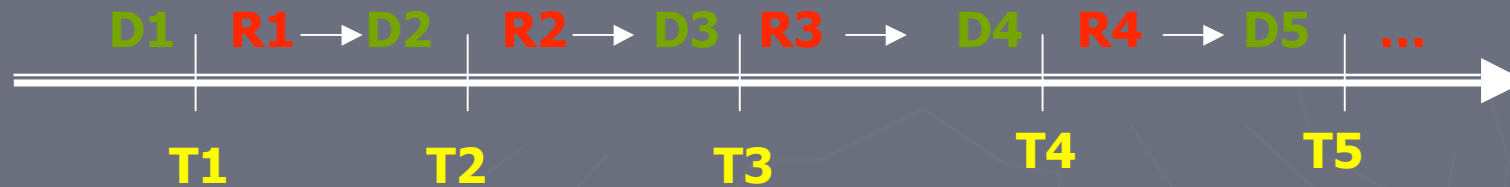
Para ir más allá

Bateman & Zock : (2003) **Natural Language Generation**. In R. Mitkov (Ed.) *Handbook of Computational Linguistics*, Oxford University Press, pp. 284-304

E. Reiter & R. Dale **Building Natural Language Generation Systems** (2000), Cambridge University Press

W. Levelt **Speaking : from Intention to Articulation**, MIT Press, 1989

TAL : observaciones informáticas



Tratamiento secuencial

Cuando se hace el tratamiento T1, todavía se conoce R3.
El tratamiento T3, no conoce R1.

Consecuencias

Ejemplo 1 : Si se sabía en T2 (análisis morfológico) que una frase tiene que contener por lo menos un verbo, no se hesitaría antes de asociar la categoría V a una palabra si ninguna otra en la frase la tiene la categoría (resultado conocido en T3, análisis sintáctico)

Ejemplo 2 : "del" se transforma en "de el". Si se sabía en T2 (análisis morfológico) que el "el" proviene de la transformación de "del" no se crearía una ambigüedad sobre "el"

TAL : observaciones informáticas

Modelo distribuido, multi-agentes

