Recursos y herramientas al amparo de las teorías lingüísticas

Mercè Lorente

merce.lorente@upf.edu

Guión

1. Introducción
2. Recursos y herramientas
3. De teorías, modelos y mecanismos
4. El diseño de recursos y herramientas: adecuación y coherencia
5. Ejemplos

De recursos y herramientas

Recursos lingüísticos (1)

- Corpus textuales
  - Orales
  - Escritos
  - Multimedia
- Recursos léxicos
  - Diccionarios
  - Bancos de datos
  - Jerarquías léxicas

Recursos lingüísticos (2)

- Bancos de conocimiento
  - Enciclopedias
  - Ontologías
  - Portales integrados (corpus, léxicos, ontologías)
- Otros

Características

- Datos ordenados sistemáticamente
- Datos etiquetados (estándares)
- Comparables
- Reutilizables

Herramientas lingüísticas

- De búsqueda y clasificación de documentos
- De estructuración (preproceso)
- De procesamiento del lenguaje natural (PLN)
- De extracción de datos lingüísticos
- De recuperación de la información (RI)

Búsqueda y clasificación

- Buscadores y metabuscadores
- Indizadores de documentos
- Filtros lingüísticos
- Clasificadores temáticos de documentos

Estructuración y preproceso

- Metadatos de corpus y de documentos
- Segmentación de unidades (o tokenización)
- Etiquetaje estructural
- Identificación de nombres propios
- Identificación y estandarización de fechas y cantidades
- Identificación de unidades fraseológicas
- Identificación de préstamos de otras lenguas

Procesamiento del lenguaje natural

- Etiquetaje morfosintáctico
- Desambiguación lingüística
- Desambiguación estadística o estocástica
- Análisis sintáctico
- Etiquetaje semántico
- Etiquetaje pragmático-discursivo

Extracción

- Interfaces de consulta de corpus
- Frecuencias y concordancias
- Análisis lexicométrico
- Extracción automática de terminología
- Detección automática de neología
  - Neología formal y filtros lexicográficos
  - Neología semántica y estrategias formales

Recuperación de información

- Expansión de consultas
- Sistemas de pregunta-respuesta
- Sistemas de diálogo persona-máquina

De teorías, modelos y mecanismos

Paradigmas científicos(1)

- El progreso científico no es visto como la acumulación de observaciones, sino como "el repetido derrocamiento de teorías científicas y su reemplazo por otras mejores o más satisfactorias“ (carácter permanentemente revolucionario de la ciencia).

POPPER, Karl R. (1959) *La lógica de la investigación científica*. México: Rei, 1996. p. 16

Paradigmas científicos (2)

- Los paradigmas son *"realizaciones científicas universalmente reconocidas que, durante mucho tiempo, proporcionan modelos de problemas y soluciones a una comunidad científica*".

KUHN, Thomas S. (1962) *La estructura de las revoluciones científicas*. México: Fondo de Cultura Económica, 2001. p. 13.

¿Qué es una teoría?

Conjunto de principios y fundamentos básicos sobre un objeto científico. Se refiere a

- la delimitación del objeto,
- los objetivos científicos que se propone,
- y al método científico utilizado.

Ejemplo: El generativismo, respecto de la lingüística anterior, es una teoría mentalista del lenguaje, que desarrolla subteorías como la teoría de la adquisición del lenguaje, la teoría de los universales del lenguaje y la teoría formal del lenguaje.

¿Qué es un modelo?

Modelo: Representación ideal de un objeto

Modelo lingüístico: Representación ideal de la gramática del hablante

- Se inscribe en una teoría (marco teórico)
- Puede ser un modelo completo, parcial o simplificado.
- Puede evolucionar.
- Para cada teoría, puede haber varios modelos.

Ejemplo: Diferentes versiones de la gramática generativa

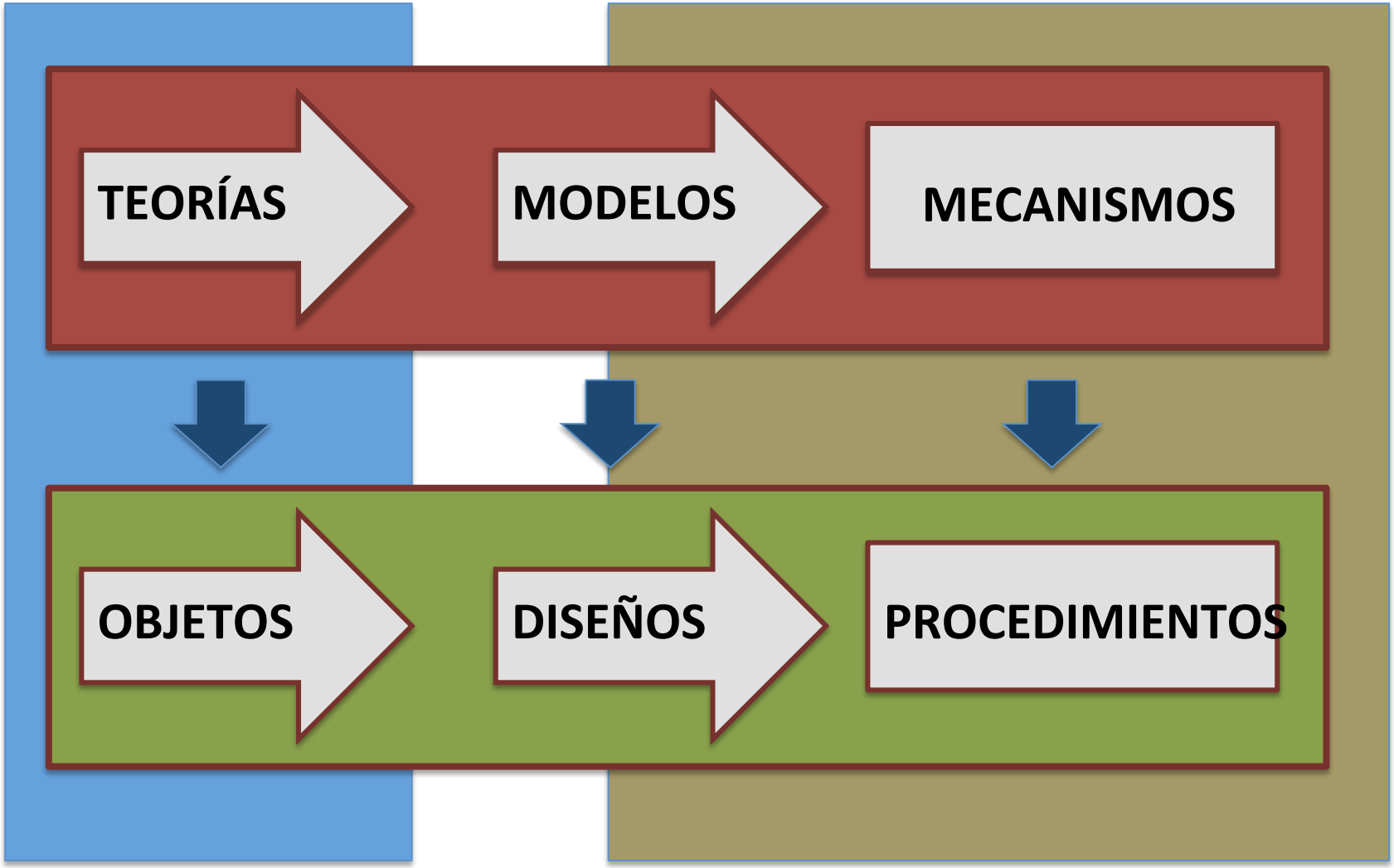
¿Qué son los mecanismos?

Lenguajes formales o matemáticos para el desarrollo de modelos gramaticales. Símbolos, reglas, relaciones, etc.

- Mecanismos descriptius o de representació
- Mecanismos de processament i validació
- Mecanismos de generalització

Ejemplos:

- Mecanismos de unificación, como los de LFG (Bresnan & Kaplan 1982) o HPSG (Pollard & Sag 1990)
- Mecanismos de generalización


TEORÍAS
MODELOS
MECANISMOS
OBJETOS
DISEÑOS
PROCEDIMIENTOS

Los paradigmas actuales de la lingüística

La historia de la lingüística (1)

- Los estudios lingüísticos antes de la lingüística
  - Las aplicaciones: Orientación prescriptiva.
  - La reflexión: Los antecedentes de la filosofía del lenguaje
  - La historia de la lengua. La gramática histórica

La historia de la lingüística (2)

- La lingüística, como disciplina científica
  - El estructuralismo europeo
  - El estructuralismo americano
  - El generativismo
  - El funcionalismo
  - El cognitivismo

El generativismo

- La teoría innatista y formal del lenguaje
- La gramática generativa
- La evolución del modelo
  - Principales hitos bibliográficos
  - Características comunes de las diferentes versiones
- Los mecanismos
  - De representación y de procesamiento

El generativismo, la teoría

- Teoría del lenguaje
  - Cambio de paradigma
  - Teoría de la adquisición del lenguaje
  - Teoría formal del lenguaje
- Innatismo y teoría de la adquisición
- Gramática universal
- Gramatica formal

Teoría del lenguaje

- Teoría formal del lenguaje
- Adecuación observacional, descriptiva y explicativa (noción de gramaticalidad)
- Método hipotético-deductivo
- Competencia y actuación
- Generación infinita
- Simplicidad (no redundancia)

Gramática formal

- Estructuras y categorías
- Reglas, principios, restricciones
- Gramática como hipótesis
- Universalidad de capacidad (procesos), no de contenidos

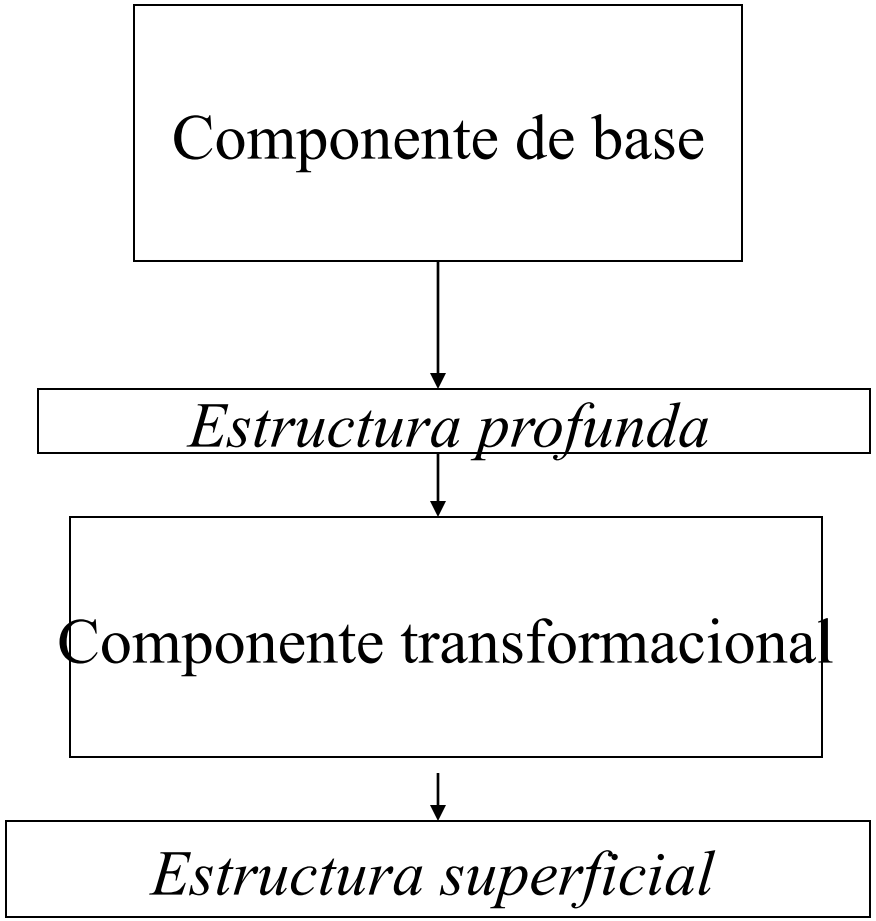
La gramática generativa, el modelo

- Gramática de la competencia
- Modelo explicativo
- Procesamiento secuencial
- Modular
  - Componentes de la gramática
  - Módulos teóricos
- Orientación sintactista
- La metáfora del ordenador: input/output

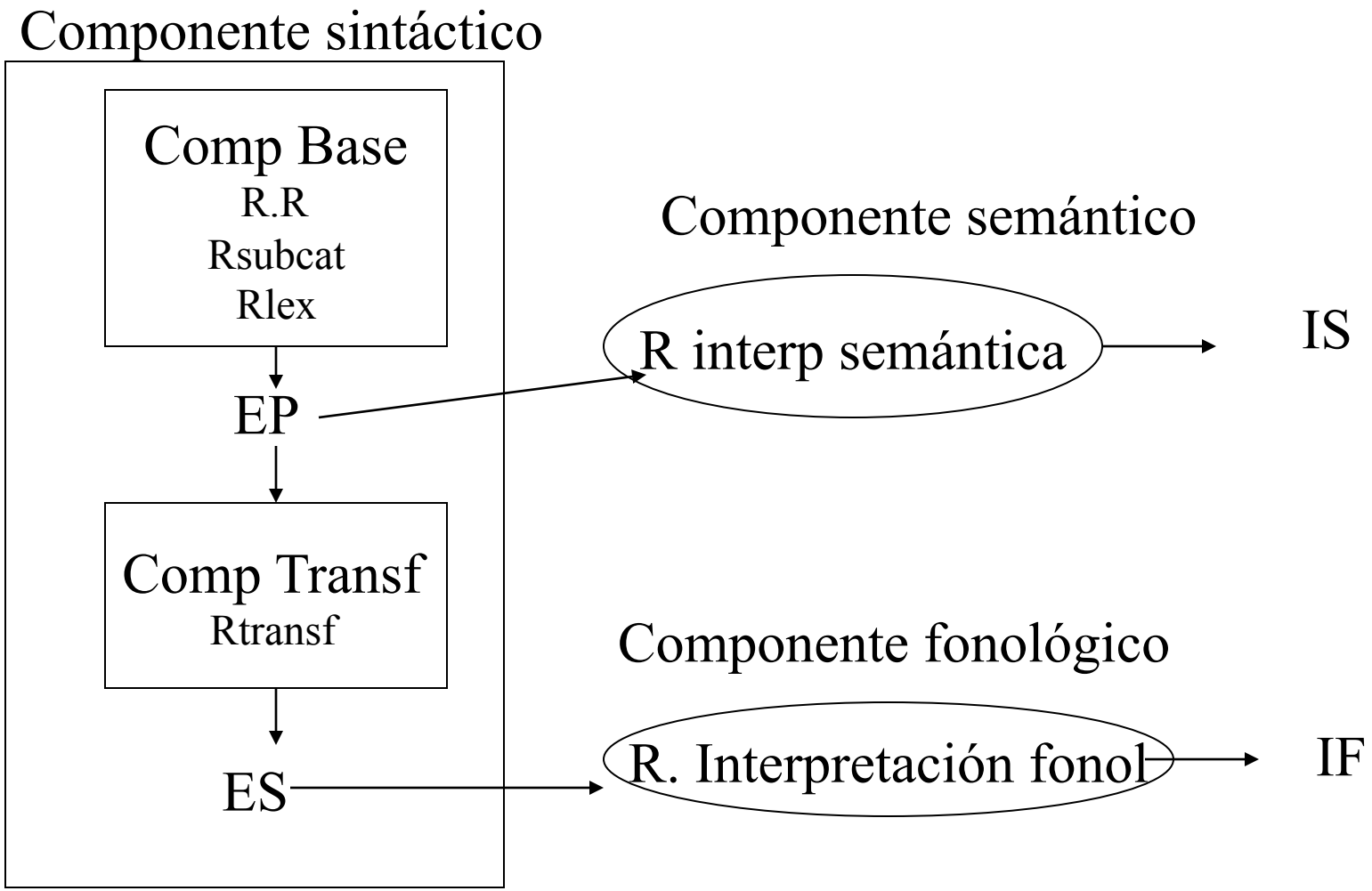
Evolución del modelo

- *Syntactic Structures* (1957)
- *Aspects of the Theory of Syntax* (1965)
- *Remarks on nominalizations* (1970)
- *Rules and Representations* (1980)
- *Lectures on Government and Binding* (1981)
- *The Knowledge of Language* (1986)
- *Barriers* (1986)
- *A Minimalist program for linguistic theory* (1992)
- *The Minimalist program* (1995)

Syntactic Structures (1957)



Aspects of the Theory of Syntax (1965)



El programa minimalista (1995)

- Programa que reduce las representaciones del modelo de P&P hacia un modelo económico, simplificado, no redundante.
- *A particular language L is an instantiation of the initial state of the cognitive system of the language faculty with options specified.* (Chomsky 1995: 219)
- El sistema cognitivo del lenguaje está formado por un componente computacional (derivacional) y por el lexicón.
- Los únicos niveles de representación son las interfaces hacia la FF y la FL.
- Reducción de categorías funcionales (T, C, D)
- En síntesis, el PM refuerza la hipótesis de la autonomía del lenguaje y incorpora mecanismos formales (*merge*) parecidos a los planteados por los FU y por la fonología de la optimidad.

EJEMPLOS DE APLICACIONES

- Lingüística de corpus
- Análisis sintáctico
- Gestión de la terminología

Qué es un corpus?

- *A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research*.

(Sinclair, 2005: 16)

No todos los corpus son corpus

- Archivo digital: Agrupación de textos en soporte informático sin relación.
- Biblioteca de textos electrónicos: Colección de textos en soporte informático , de formato estándar y guiados por normas de contenido, sin criterio de selección.
- Corpus informatizado: Colección de textos seleccionados por criterios lingüísticos (externos o internos), codificados de manera estándar y homogenea, para ser procesados informáticamente y para reflejar el comportamiento de una o más lenguas.

(Llisterri y Torruella, 1999: 50)

Uso de corpus

- La investigación en lingüística
- La edición de obras de referencia (diccionarios, gramáticas, libros de estilo, tesauros documentales...)
- La enseñanza de lenguas (manuales, materiales de aprendizaje, etc.)
- El desarrollo de herramientas

Contenidos

- Lengua oral: conferencies, mesas redondas, interacciones en áula, grabaciones TV o radio, cine y teatro, diálogos, entrevistas, llamadas (call centre), diarios de laboratorio, videoguía quirúrgica, etc.
- Lengua escrita: publicaciones, prensa, cartas, testamentos, leyes, pero también blogs, webs, publicidad, memorias de empresas y organismos, correo electrónico, Twitter, mensajes, ...


TIPOS DE CORPUS
(Torruella y Llisterri, 1999)
PORCENTAJE/DISTRIBUCIÓN
corpus grande
corpus equilibrado
corpus piramidal
corpus monitor
corpus paralelo
corpus comparables
corpus multilingües
corpus oportunista
ESPECIFICIDAD TEXTOS
corpus general
corpus especializado
corpus genérico
corpus periódico o cronológico
corpus diacrónico
CANTIDAD TEXTOS
corpus textual (whole text corpus)
corpus de referencia (reference corpus)
corpus léxico (samples corpus)
CODIFICACIÓN Y ANOTACIÓN
corpus simple
corpus codificado o anotado
DOCUMENTACIÓN
corpus documentado
corpus no documentado

La lingüística de corpus (LC)

- The study of language based on examples of 'real life' language use. (McEnery & Wilson 1996: 1)
- You know a word by the company it keeps. (Firth 1957)
- I'm interested in explaining what does occur, not what might occur. (Sinclair 1991)

Teorías y Lingüística de Corpus

- Estructuralismo
- Funcionalismo
- Lingüística textual
- Variación lingüística


Concordance - Mozilla Firefox
Eitxer Edita Visualitza Historial Adreces d'interès Eines Ajuda
:: CampusGlobal :: Gmail - Recibidos... 324. El portal inf... Concordance Herramientas de ... Corpus Linguistic... WordSmith Tools Corpus Linguistic...
http://the.sketchengine.co.uk/bonito/run.cgi/first?corpname=preloaded%2Fbnc2&queryselector=phraserow&iquery=&lemma=&lpos=&phr
kwic corpus
ria Bel
sh National Corpus
Search
in British National Corpus
Corpus: British National Corpus
Hits: 318 (2.8 per million)
Page 1 of 16 Go Next | Last
J2X scientists have called attention to the large amount of halon gas, hundreds of times more damaging
J0V header allows for the provision of a very large amount of structured or unstructured information
J0V Archive are nevertheless very naive and need a large amount of support and assistance from Archive staff
J56 several dialects of Arabic. It claimed that a large amount of money would be the immediate reward if
J56 their way to Egypt. Thus, in spite of the large amount of flying experience I had accumulated overseas
J54 her blue sports car, its boot piled with a large amount of luggage. Pulling off her bright headscar
J52 light, but the process seems to consume a large amount of energy. Fireflies use their light for attracting
JXG Reserved Memory 3.2.4 In the example below, a large amount of memory is initially reserved. To begin
JXG space/speed compromise for files holding a large amount of data. As far as the programmer is
JT5 help. I, I'm not looking to put a I put a large amount of money away but that's I'm interested in
J75 unions will be placing with their lawyers a large amount of employment-related personal injury work
JS9 estate, for which he probably asks for a large amount of money, since it involves an enormous amount
JY8 stretched the gap to reveal an embarrassingly large amount of cleavage. His eyes were still fixed
JJG Bassett magistrate's court, If we have such a large amount of money that we can spend I think the questions
J18 fruited during the dry season, producing a large amount of seeds, which were simultaneously dispersed
J18 dispersing some species, howlers waste a large amount of fruit and seed material. Figs are
J15 amounts of cash, they nevertheless hold a large amount of other relatively liquid assets which act
J15 maturity date and then suddenly releasing a large amount of liquidity into the economy. When
HRH systematic collection and analysis of a large amount of quantitative and qualitative data. It also

FREELING
http://nlp.lsi.upc.edu/freeling/demo/demo.php

FreeLing 3.0

AN OPEN-SOURCE SUITE OF LANGUAGE ANALYZERS

Write your sentences

Rubalcaba se inspira en Francia y Suecia para lanzar una propuesta de carácter retroactivo que establece como tasación del inmueble la que tenía en el momento de la

Analysis options

- [x] Multiword detection
- [x] Number recognition
- [x] Date/Time recognition
- [x] Quantities, ratios, and percentages
- [x] Named Entity detection
- [ ] Named Entity classification
- [ ] Phonetic encoding
- (•) No sense annotation
- ( ) WN sense annotation: Frequency sorted (MFS disambiguation)
- ( ) WN sense annotation: PageRank sorted (UKB disambiguation)

Select language: Spanish

Select output: PoS Tagging

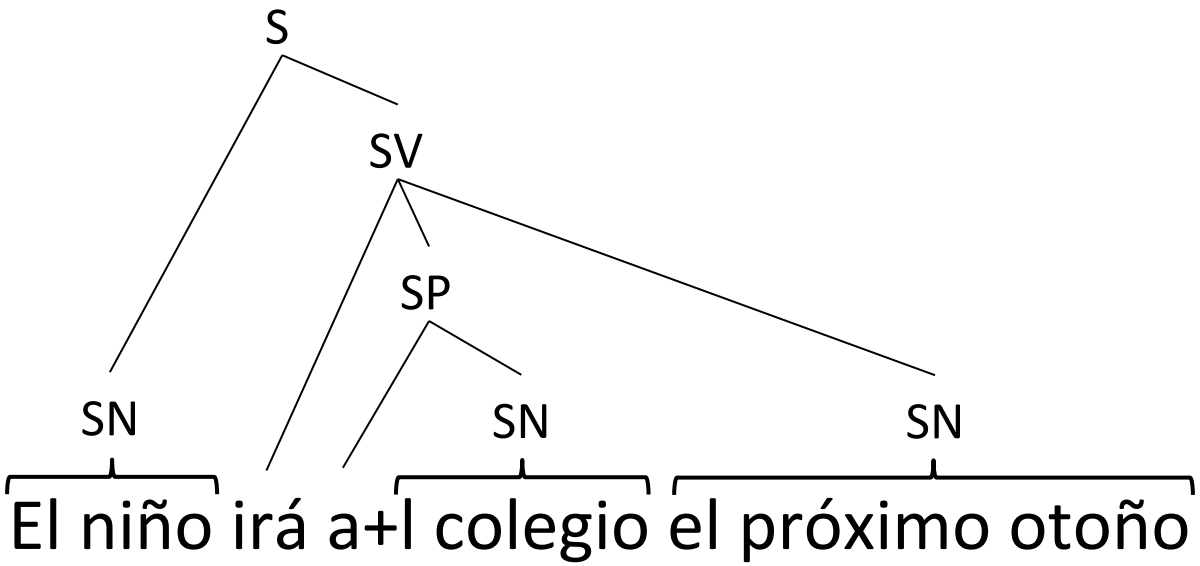
Submit

Analysis Results

Sentence #1

Rubalcaba	se	inspira	en	Francia	y	Suecia	para	lanzar	una	propuesta	de	carácter	retroactivo	que	establece
rubalcaba	*se*	*inspirar*	*en*	*francia*	*y*	*suecia*	*para*	*lanzar*	*uno*	*propuesta*	*de*	*carácter*	*retroactivo*	*que*	*establecer*
NP00000	P00CN000	VMIP3S0	SPS00	NP00000	CC	NP00000	SPS00	VMN0000	DI0FS0	NCFS000	SPS00	NCMS000	AQ0MS0	PR0CN000	VMIP3S0

Análisis sintáctico



(El niño) irá (al colegio) (el próximo otoño)

((El niño) (irá (a(l colegio)) (el próximo otoño)))

Sintaxis

- Chunking (identificación de constituyentes o sintagmas)
- Full parsing (análisis de constituyentes)
- Constraint grammar (análisis de dependencias)

- HERRAMIENTAS: IULA, Freeling, MaltParser

terminus 2.0

Gestió de corpus i terminologia en línia

Llengua: Català

Nom d'accés:

Contrasenya:

d'acord

Us agrada la nova versió del programmari?
Envieu la vostra opinió

Accés a la versió antiga de Terminus

TERMINUS és una estació de treball per a la terminologia. Integra la gestió de corpus i de terminologia. Permet crear i gestionar grups de treball i modelar les categories de dades. Inclou la cadena completa del treball terminogràfic individual i en equip: cerca, constitució i exploració de corpus textuals, extracció de termes, gestió de glossaris i projectes, creació i manteniment de bases de dades i edició de diccionaris. Terminus consta de diversos mòduls articulats:

- **Projectes:** permet crear un projecte terminològic.
- **Fonts:** permet gestionar les fonts utilitzades en un projecte terminològic.
- **Estructuració conceptual:** permet crear un arbre de camp per estructurar els termes d'un glossari.
- **Documents:** permet incloure arxius de text que després constituiran el corpus de treball.
- **Corpus:** permet agrupar documents en corpus.
- **Anàlisi:** permet explorar corpus mitjançant freqüències, concordances, n-grames i càlcul d'associació entre formes, i també permet l'extracció de termes a partir de corpus textuals especialitzats.
- **Glossaris:** permet declarar els glossaris que formen part d'un projecte.
- **Termes:** permet entrar les dades terminològiques al glossari i consultar-les.