PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Métodos de Investigación en Letras

César Antonio Aguilar

Facultad de Lenguas y Letras

16/10/2017

caguilara@uc.cl

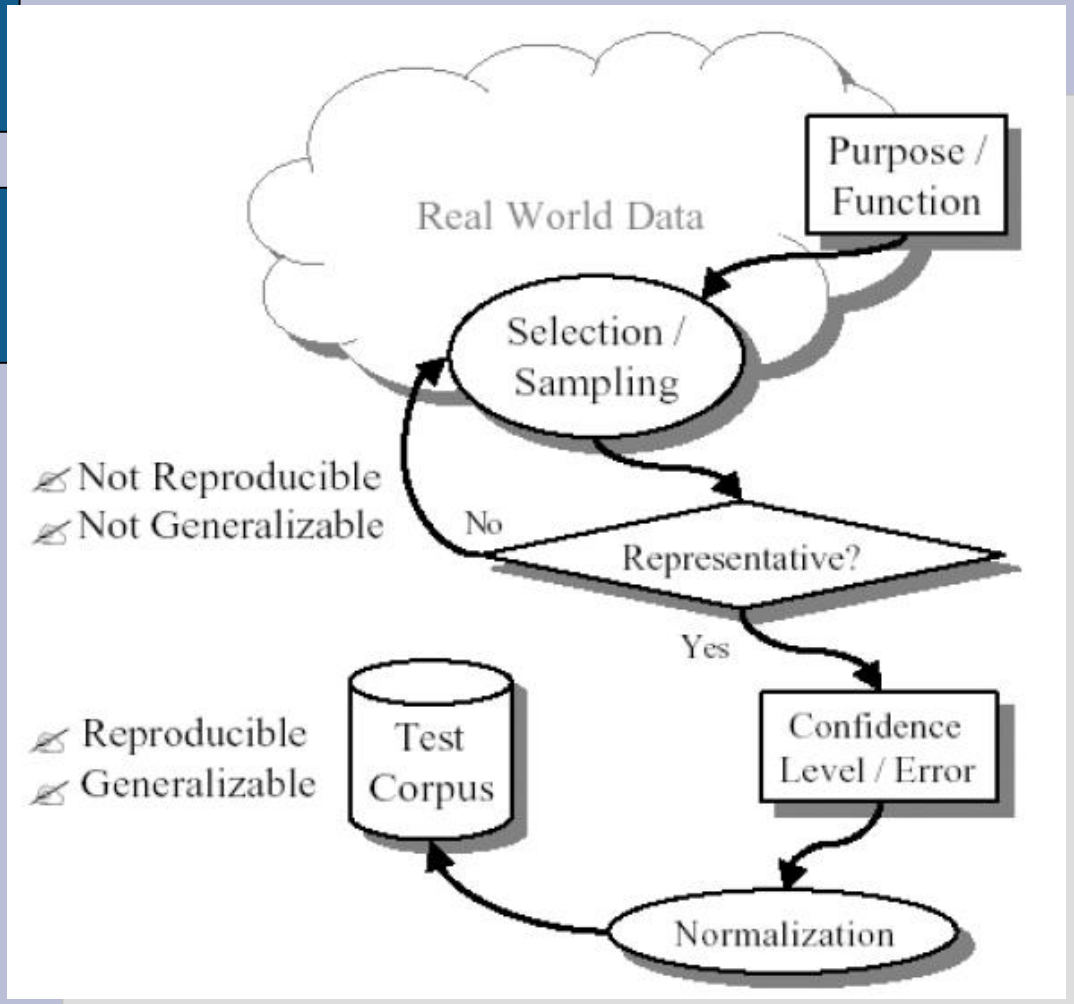
Síntesis de la clase pasada

En la clase pasada hicimos una breve descripción sobre cómo funcionan los métodos cualitativos, destacando dos de sus características:

1. Su énfasis en desarrollar herramientas precisas para la colecta de datos, tales como las entrevistas con formatos libres o con micrófono abierto, la generación de relatos de vida, y otras similares.

2. Su interés por construir **corpus lingüísticos**, los cuales permiten analizar datos previamente colectados y organizados por investigadores interesados en compartirlos a la comunidad.

¿Qué es un corpus lingüístico? (1)



Entrando en detalles, un corpus se concibe como un modelo que representa una realidad lingüística, por lo que ofrece una base empírica que muestra el funcionamiento de una lengua natural.

En un plano estadístico debe ser neutral, esto es, proporcional respecto a las muestras que se tomen.

Por ello, es un instrumento reutilizable para distintos tipos de análisis.

¿Qué es un corpus lingüístico? (2)

¿Cómo se construye un corpus?:

1. Objetivos
1.1. Propósitos
1.2. Límites

2. Selección de textos

3. Obtención de textos
3.1. Digitalización
3.2. Documentos electrónicos

5. Administración

6. Sistemas de consulta

¿Qué es un corpus lingüístico? (3)

- Algo que marca una diferencia fundamental entre tener una colección de textos o fragmentos y un CL, es su anotación o etiquetado.

- El etiquetado es una plantilla que ayuda a manipular electrónicamente el contenido textual de un CL.

- En este caso, se consideran dos tipos de etiquetados:

1. **Etiquetado textual:** se emplea para compilar y organizar los documentos que van a formar parte del CL.
2. **Etiquetado lingüística:** es aquel que se usa para simbolizar los hechos lingüísticos particulares que se van a analizar. Si bien se considera el etiquetado fonético, prosódico, pragmático y discursivo, regularmente se asocia el etiquetado morfológico y sintáctico bajo el nombre de **POST** (*Part-of-Speech-Tagging*).

Anotación textual (1)

1. Ayuda a los procesos de búsqueda y recuperación por medio de un sistemas computaciones.
2. Se relacionan con lenguajes de compilación como SGM, HTML, XML u otros similares.
3. Facilita su almacenamiento en repositorios o bases de datos.
4. Permite una mejor visualización vía el empleo de interfaces o Internet.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2//EN">
<html>
<head>
<!-- owner_name="Anoop Sarkar" -->
<title>The Xtag Project</title>

<link rev="made" href="mailto:anoop@linc.cis.upenn.edu">

<!-- HTML 3.2 tags -->
<meta http-equiv="keywords" content="computational linguistics">
<meta http-equiv="keywords" content="computer science">
<meta http-equiv="keywords" content="linguistics">
<meta http-equiv="keywords" content="natural language">
<meta http-equiv="keywords" content="natural language processing">
<meta http-equiv="keywords" content="syntax">
<meta http-equiv="keywords" content="semantics">
<meta http-equiv="keywords" content="research">
<meta http-equiv="keywords" content="cognitive science">
<meta http-equiv="keywords" content="technical reports">
<meta http-equiv="keywords" content="tech reports">
<meta http-equiv="keywords" content="research papers">
<meta http-equiv="keywords" content="source code">
<meta http-equiv="reply-to" content="anoop@linc.cis.upenn.edu (Anoop Sarkar)">

</head>
```

Anotación textual (2)

Veamos un ejemplo real para entender cómo funciona esta clase de anotado:

www.corpus.unam.mx:8080/corcode/

Se trata de un buscador de contextos definitorios para el inglés, desarrollado por el **Grupo de Ingeniería Lingüística**, afiliado a la UNAM (México)

Universidad Nacional Autónoma de México

INSTITUTO DE INGENIERÍA UNAM

Ingeniería lingüística

"In necessariis unitas, in dubiis libertas..."
San Agustín

www.iling.unam.mx/corpus/

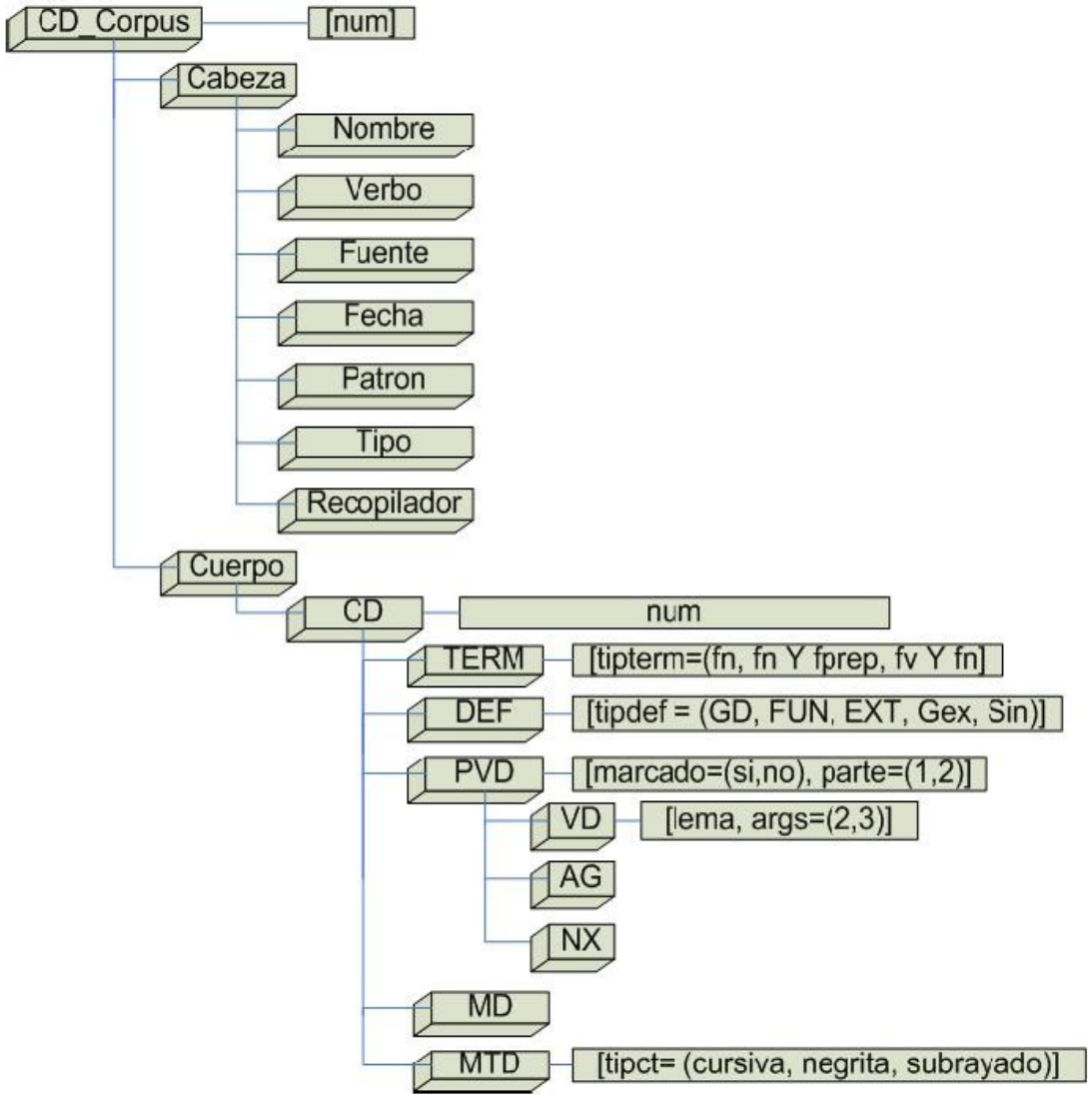
Anotación textual (4)

3. Metodología: un etiquetado del tipo XML se basa en el uso de dos elementos básicos:

Etiquetas: es el nombre o la variable que ocupo para reconocer a un objeto.

- Atributos: es el conjunto de rasgos mínimos que me sirven para identificar a dicho objeto.

Nota importante: antes de hacer cualquier cosa, hay que diseñar un esquema de etiquetado (p. e., usando papel y lápiz)



Encabezado y Cuerpo (1)

La *cabeza* (o *Head*), es un conjunto de descriptores que resumen la información contenida en una página WEB:

ETIQUETAS	FUNCIÓN
Cabeza	Dentro de este bloque se encuentra la información del documento como nombre, de qué verbo se trata, la fuente, fecha, recopilador, etc.
Fuente	Indica el nombre del corpus que se está etiquetando. Es muy importante tener localizada la fuente de origen de los documentos.
Fecha	Fecha en la que fue recopilado y etiquetado el documento.
Nombre	Contiene el nombre de la recopilación hallada en el documento, por ejemplo, puede contener “verbo ser”.
Verbo	Cuando en el documento únicamente es analizado un verbo definitorio en cualquiera de sus predicaciones tiene que señalarse el nombre del verbo.
Tipo	Existen varios tipos de definiciones: analítica, funcional, etc. En el caso en que el criterio de clasificación del documento sea el tipo de *definición* se tendrá que indicar.
Recopilador	Nombre de la persona que recopiló el documento y lo consignó al Corpus de CD.

Encabezado y Cuerpo (2)

En cambio, el *Body* ya es la estructuración de los contenidos de tal página:

Etiqueta	Función
CD	Indica los elementos que constituyen al CD dentro de ellos se encuentran el término, su definición, la predicación verbal y relaciones de correferencia.
TERM	En sus atributos se marca si se trata de un término lingüístico o de uno no lingüístico (cifras, símbolos). Se toman en cuenta tres tipos de frases: *fn* (frase nominal), *fn Y fprep* (frase nominal seguida de frase prepositiva) y *fv Y fn* (frase verbal seguida de frase nominal).
DEF	En ella se debe omitir cualquier texto complementario que de manera estricta no forme parte de dicha definición. Existen cinco tipos: *GD* (Género próximo/Diferencia específica), *FUN* (Funcional), *EXT* (Meronímia/Extensional), *Gex* (Género exclusivo) y *Sin* (Sinonímica) que se marcan en los atributos.
PVD	Contiene todos los componentes de una PVD: VD, clítico *se*, verbo auxiliar, verbo definitorio y nexo.
VD	Cuenta con los atributos *lema, args* (marca los argumentos del verbo); *mod* (indica el modo verbal: infinitivo *inf*, gerundio *ger*, participio *part*, formas finitas o verbo conjugado *fin*).
SEmarc	Se indica su posición respecto al verbo. El atributo distingue entre *enclítico* (*enc*) cuado *se* es parte de la morfología verbal y está en posición final y *preclítico* (*prec*) cuando el clítico está en posición preverbal.
Vaux	Contiene cualquier verbo auxiliar dentro de la PVD (p. e., *se* ***puede*** *considerar como*, se ***ha*** definido, se ***debe*** concebir como…)
NX	Señala la función que cumple un adverbio o preposición entre el verbo y la definición.

Ordenar/clasificar (1)

CD sin etiquetas

M. Godron y G. Merriam entre otros, quienes consideran a la Ecología del Paisaje como: "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 "

Ordenar/clasificar (2)

Insertando etiquetas...

<CD num= “1"> <PP tippp= "Aut"> M. Godron y G. Merriam </PP> entre otros, quienes consideran a la Ecología del Paisaje como: "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 " </CD>

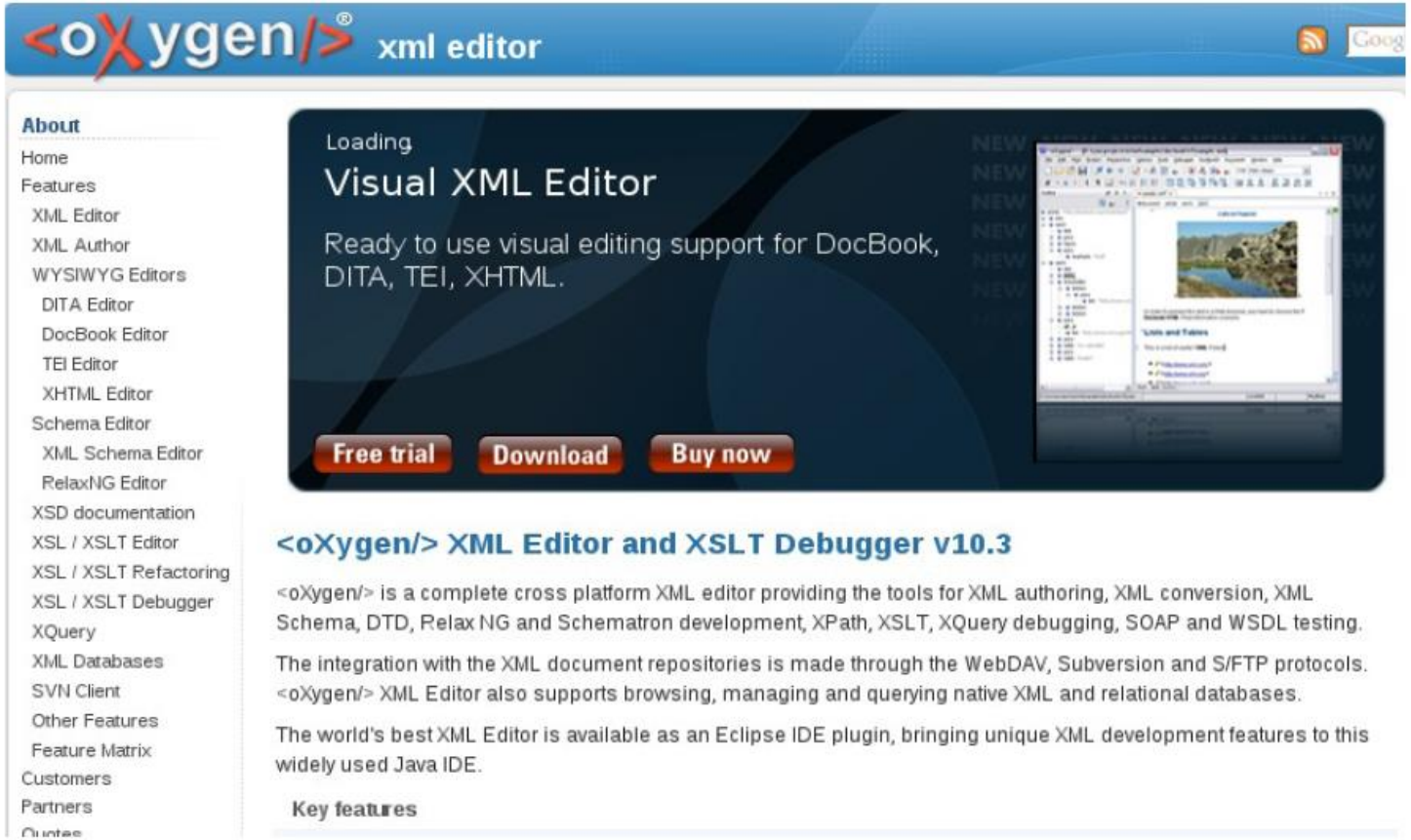
Ordenar/clasificar (3)

Resultado

<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>

Sistemas de etiquetado XML

¿Cómo construyo un corpus usando etiquetas XML? Pueden usar un editor llamado **oXygen**:

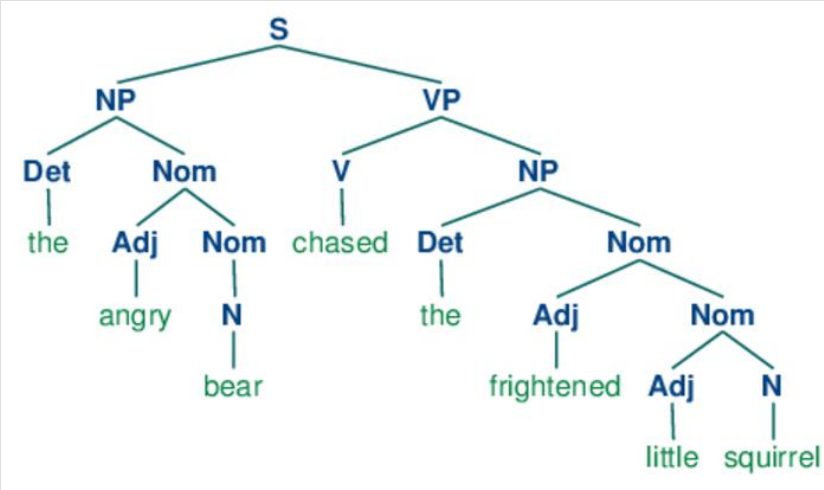


www.oxygenxml.com

Corpus con anotado lingüístico (1)

Vamos a ver ahora un tipo de anotado que se enfoca en el reconocimiento de unidades gramaticales, y que vamos a denominar **anotado morfo-sintáctico**. Este etiquetado nos permite hacer, sobre todo, análisis gramaticales usando programas de cómputo específicos para esta clase de tareas.

```
grammar2 = nltk.CFG.fromstring("""
  S  -> NP VP
  NP -> Det Nom | PropN
  Nom -> Adj Nom | N
  VP -> V Adj | V NP | V S | V NP PP
  PP -> P NP
  PropN -> 'Buster' | 'Chatterer' | 'Joe'
  Det -> 'the' | 'a'
  N -> 'bear' | 'squirrel' | 'tree' | 'fish' | 'log'
  Adj  -> 'angry' | 'frightened' |  'little' | 'tall'
  V ->  'chased'  | 'saw' | 'said' | 'thought' | 'was' | 'put'
  P -> 'on'
  """)
```



Corpus con anotado lingüístico (2)

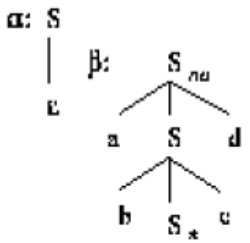
EAGLES: iniciativa de la Comisión Europea (1993-1996), propuesta para el desarrollo y generación de recursos para el procesamiento de lenguaje natural.

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	-	0
6	Género Semántico	-	0
7	Grado	Apreciativo	A

Forma	Lema	Etiqueta
chico	chico	NCMS000
chicos	chico	NCMP000
chica	chica	NCFS000
chicas	chica	NCFP000
oyente	oyente	NCCS000
oyentes	oyente	NCCP000
cortapapeles	cortapapeles	NCMN000
tesis	tesis	NCFN000
Antonio	antonio	NP00000

Corpus con anotado lingüístico (3)

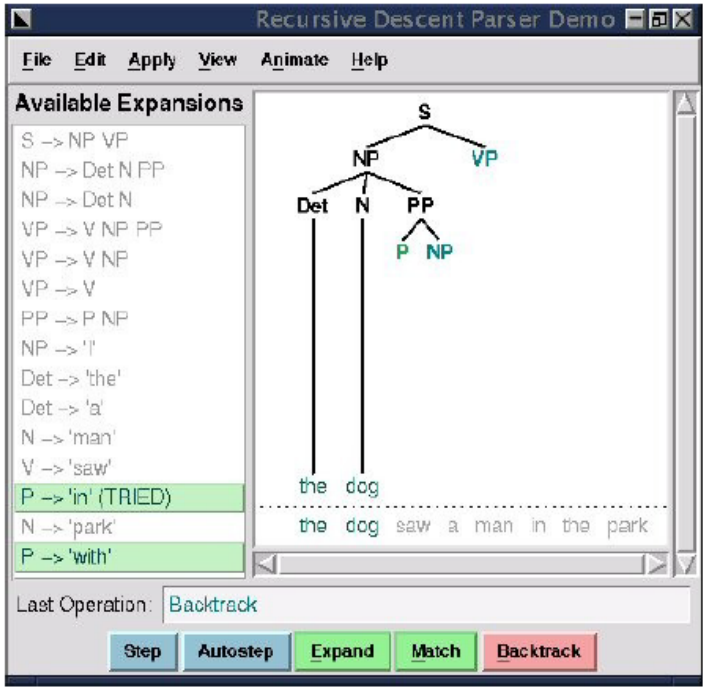
PennTreeBank: Desarrollado por la Universidad de Pensilvania, es una opción para la creación y explotación de análisis sintáctico computacional (*parsing*).



The XTAG Project

The Penn Treebank POS tagset

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subord. conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote



Corpus con anotado lingüístico (4)

1. Contar con una fuente de datos gramaticales fiables, la cual pueda consultar. Es el caso del *Corpus Brown*.

BROWN CORPUS MANUAL

MANUAL OF INFORMATION
to accompany
A Standard Corpus of Present-Day
Edited American English, for use
with Digital Computers.

by
W. N. Francis
H. Kucera
Brown University

Providence, Rhode Island
Department of Linguistics
Brown University
1964

Revised 1971

Revised and Amplified
1979

Figure 1. An Example of Tagged Text (excerpted from the Brown Corpus)

```
The/at grand/jj jury/nn commented/vbd on/in a/at number/nn of/in
other/ap topics/nns ,/, among/in them/ppo the/at Atlanta/np and/cc
Fulton/np-tl County/nn-tl purchasing/vbg departments/nns which/wdt
it/pps said/vbd "/" are/ber well/ql operated/vbn and/cc follow/vb
generally/rb accepted/vbn practices/nns which/wdt inure/vb to/in
the/at best/jjt interest/nn of/in both/abx governments/nns "/" ./.
```

Tareas relacionadas con corpus anotados gramaticalmente (1)

Veamos un ejemplo sobre cómo podemos usar un corpus con etiquetado gramatical. Para ello, vamos a usar un corpus de documentos técnicos llamado **Bwananet**, desarrollado por el Instituto Universitario de Lingüística Aplicada, afiliado a la Universidad Pompeu Fabra (Barcelona, Cataluña, España).

upf. Universitat Pompeu Fabra Barcelona

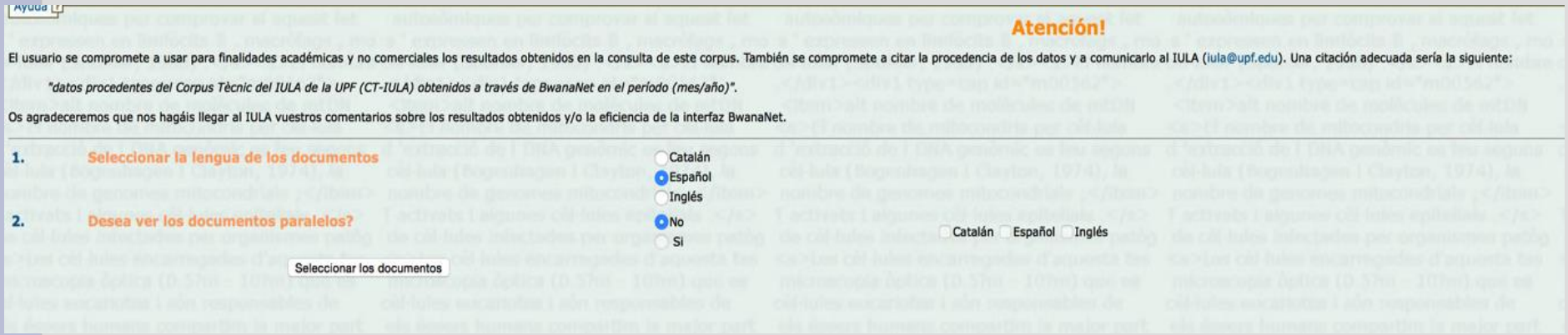
IULA Institut Universitari de Lingüística Aplicada

BWANANET Programa d'explotació del corpus Tècnic de l'IULA

http://bwananet.iula.upf.edu

Tareas relacionadas con corpus anotados gramaticalmente (2)

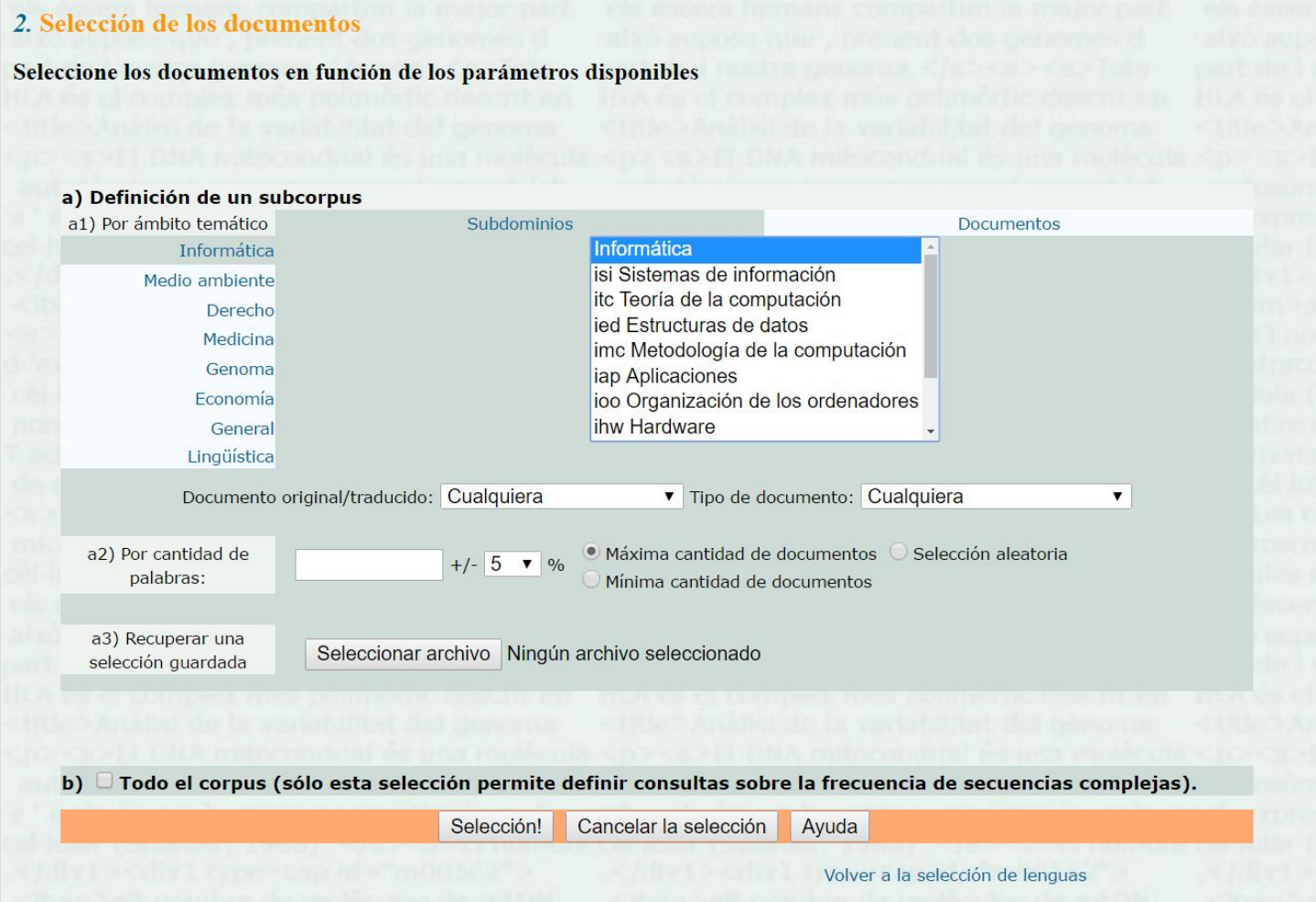
Bwananet es un sistema que tiene almacenado distintos corpus de áreas especializadas en 3 lenguas: inglés, español y catalán. Igualmente, cuenta con un motor de búsqueda que permite realizar varias tareas terminológicas. Así, empecemos seleccionado un idioma:



Pregunta: Vamos a plantearnos una que sea sencilla: la gente que trabaja en ingeniería y ciencias de la computación, ¿usan términos derivados del concepto de *gramática*, igual que los lingüistas? ¿Cómo podemos saberlo?

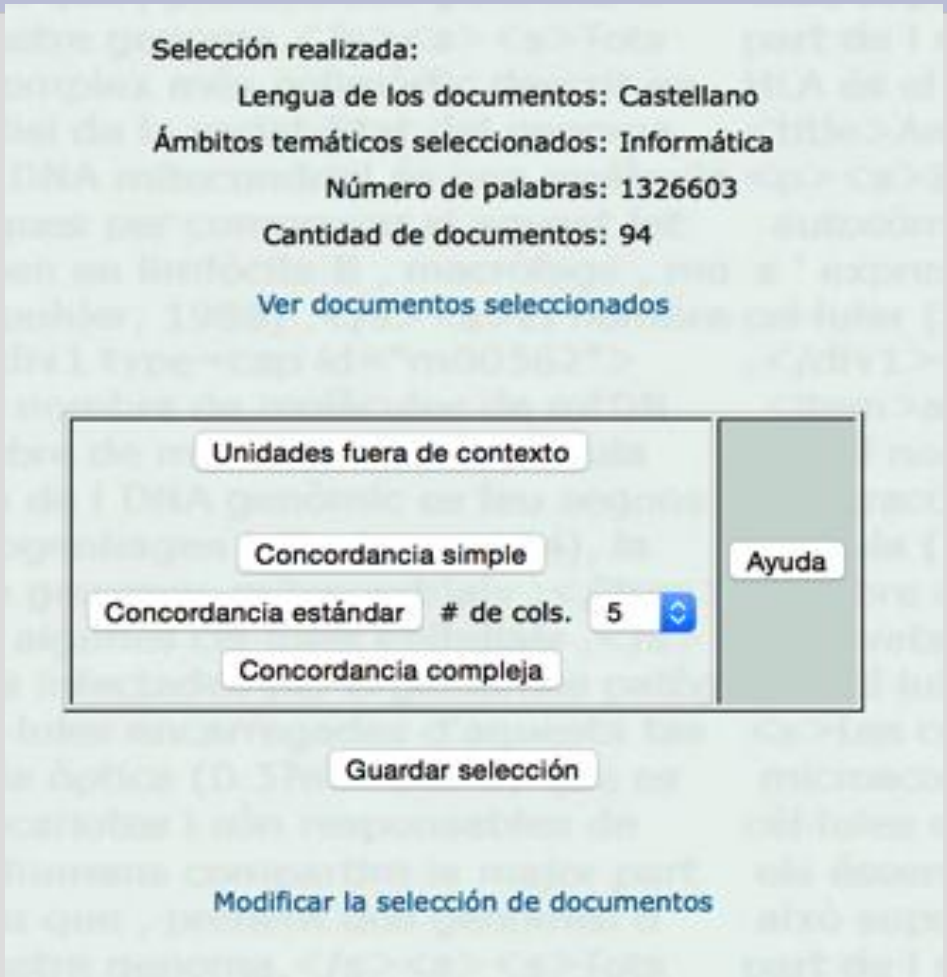
Tareas relacionadas con corpus anotados gramaticalmente (3)

Ahora, seleccionamos un dominio de conocimiento, p.e., **informática**:



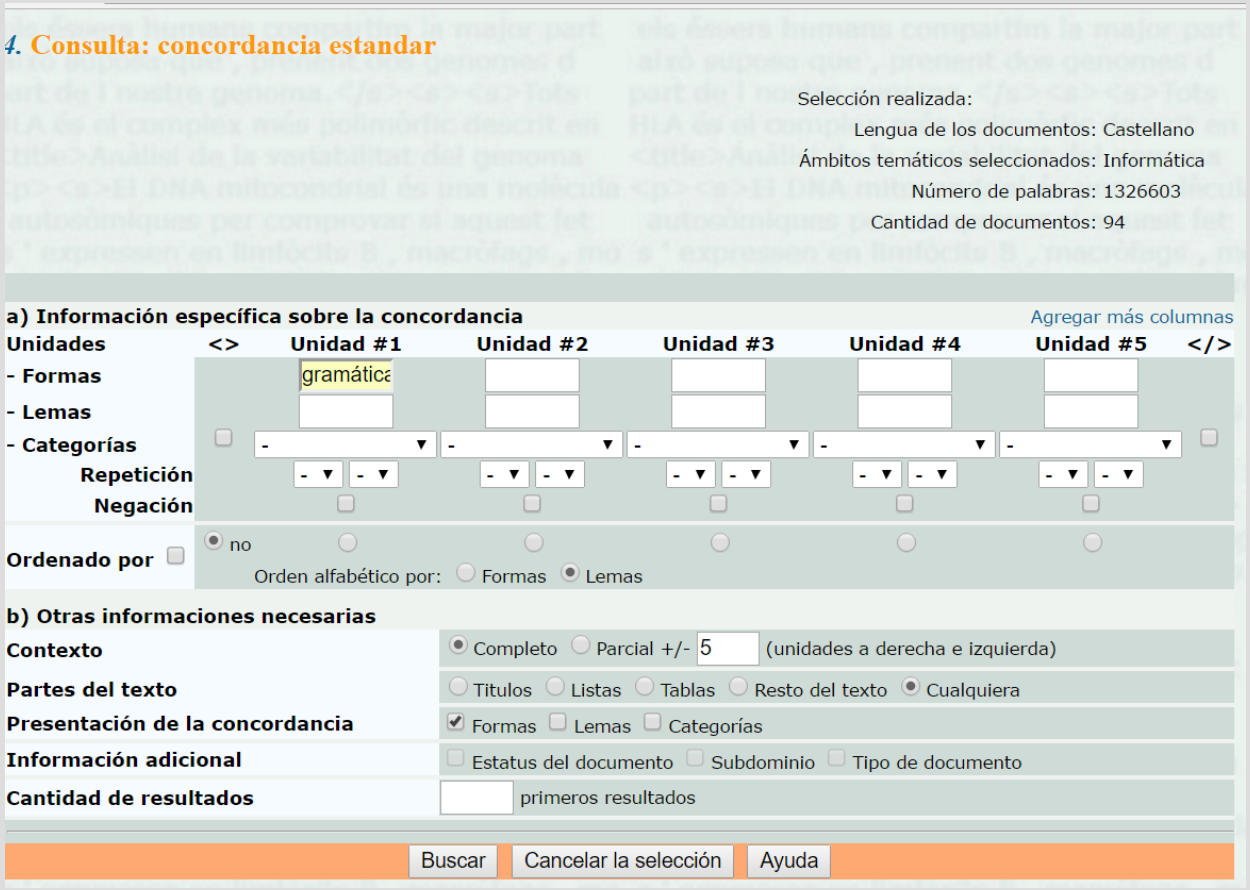
Tareas relacionadas con corpus anotados gramaticalmente (4)

A continuación, viene una de las partes más difíciles del proceso: fijar una *ventana de palabras,* que no es otra cosa más que una secuencia de lexemas organizada según el número de unidades que consideramos contiene el patrón que buscamos.



Tareas relacionadas con corpus anotados gramaticalmente (5)

Finalmente, tenemos una ventana que nos permite establecer los criterios de búsqueda que tenemos en mente para nuestros patrones.



Tareas relacionadas con corpus anotados gramaticalmente (6)

Veamos los resultados:

<s>Ahora definimos una	***gramática***	regular como aquella gramática cuyas reglas de reescritura se adhieren a las siguientes restricciones .</s>
<s>Ahora definimos una gramática regular como aquella	***gramática***	cuyas reglas de reescritura se adhieren a las siguientes restricciones .</s>
<s>El lado izquierdo de cualquier regla de reescritura de una	***gramática***	regular debe consistir en un solo no terminal , mientras que el lado derecho debe ser un terminal seguido por un
<s>Así , las reglas de reescritura de la forma xxx estarían permitidas en una	***gramática***	regular , mientras que las reglas de la forma xxx no lo estarían .</s>
<s>La figura 1.23 presenta una	***gramática***	regular que genera cadenas que representan números racionales en notación decimal , y la figura 1.24 presenta
esentan números racionales en notación decimal , y la figura 1.24 presenta una	***gramática***	regular que genera cadenas consistentes en una o más copias de l patrón xy .</s>
<s>Aunque lo permite la definición de una	***gramática***	regular , en_realidad no se requieren reglas de este tipo .</s>
<s>De hecho , en una	***gramática***	regular cualquier regla de la forma xxx podría reemplazar se con el par de reglas xxx donde X es un no terminal
reglas xxx donde X es un no terminal que no aparece en ningún otro lugar de la	***gramática***	, sin alterar el conjunto de cadenas que podría generar la gramática .</s>
lugar de la gramática , sin alterar el conjunto de cadenas que podría generar la	***gramática***	.</s>
<s>Para cada alfabeto xxx : G es una	***gramática***	regular de xxx : M es un autómata finito de xxx .</s>
<s>Si G es una	***gramática***	regular de xxx , podemos convertir la en una gramática regular G' que genera el mismo lenguaje pero que no co
<s>Si G es una gramática regular de xxx , podemos convertir la en una	***gramática***	regular G' que genera el mismo lenguaje pero que no contiene reglas de reescritura cuyo lado derecho consiste e
i M es el autómata finito no determinista , entonces podemos definir G' como la	***gramática***	regular de xxx para la cual los no terminales son los estados de S , el símbolo inicial es xxx y las reglas de reesc
si Q está en F._En ambos casos , xxx ya que la derivación de una cadena de la	***gramática***	G' corresponde directamente a una ruta en el diagrama de transiciones de M que conduce de l estado inicial a un
, en esencia , otra forma de representación de las reglas de reescritura de una	***gramática***	.</s>
<s>para describir un lenguaje por medio de una	***gramática***	, se requiere una lista de reglas de reescritura ;</s>
<s>El concepto de	***gramática***	presentado en este capítulo es importante ya que la sintaxis de la mayoría de los lenguajes de programación est
procesador para un lenguaje natural), nos gustaría basar nuestro diseño en la	***gramática***	más sencilla posible .</s>
econocidos por los autómatas mejorados , utilizaremos de nuevo el concepto de	***gramática***	.</s>
l permitir mayor complejidad en la estructura de las reglas de reescritura de la	***gramática***	, seremos capaces de identificar las gramáticas que generan los lenguajes reconocidos por nuestras máquinas m
os lenguajes que reconocen los autómatas de pila , presentamos el concepto de	***gramática***	independiente de l contexto .</s>

¿Qué nos muestra Bwananet?

Tareas relacionadas con corpus anotados gramaticalmente (7)

Y ahora, vamos a hacer búsquedas con *lemas* :

4. Consulta: concordancia estandar

Selección realizada:
Lengua de los documentos: Castellano
Ámbitos temáticos seleccionados: Informática
Número de palabras: 1326603
Cantidad de documentos: 94

a) Información específica sobre la concordancia Agregar más columnas

Unidades	<>	Unidad #1	Unidad #2	Unidad #3	Unidad #4	Unidad #5	</>
- Formas							
- Lemas		gramática	regular				
- Categorías		-	-	-	-	-	
Repetición		- -	- -	- -	- -	- -	
Negación							

Ordenado por no

Orden alfabético por: Formas Lemas

b) Otras informaciones necesarias

Contexto	Completo Parcial +/- 5 (unidades a derecha e izquierda)
Partes del texto	Titulos Listas Tablas Resto del texto Cualquiera
Presentación de la concordancia	Formas Lemas Categorías
Información adicional	Estatus del documento Subdominio Tipo de documento
Cantidad de resultados	primeros resultados

Buscar | Cancelar la selección | Ayuda

Tareas relacionadas con corpus anotados gramaticalmente (8)

Y los resultados son:

<s>Ahora definimos una	***gramática regular***	como aquella gramática cuyas reglas de reescritura se adhieren a las siguientes restricciones .</s>
<s>El lado izquierdo de cualquier regla de reescritura de una	***gramática regular***	debe consistir en un solo no terminal , mientras que el lado derecho debe ser un terminal seguido por
reglas de reescritura de la forma xxx estarían permitidas en una	***gramática regular***	, mientras que las reglas de la forma xxx no lo estarían .</s>
<s>La figura 1.23 presenta una	***gramática regular***	que genera cadenas que representan números racionales en notación decimal , y la figura 1.24 preser
os racionales en notación decimal , y la figura 1.24 presenta una	***gramática regular***	que genera cadenas consistentes en una o más copias de l patrón xy .</s>
<s>Aunque lo permite la definición de una	***gramática regular***	, en_realidad no se requieren reglas de este tipo .</s>
<s>De hecho , en una	***gramática regular***	cualquier regla de la forma xxx podría reemplazar se con el par de reglas xxx donde X es un no termi
<s>La importancia de las	***gramáticas regulares***	reside en que los lenguajes generados por ellas son exactamente aquellos que reconocen los autómat
<s>Para cada alfabeto xxx : G es una	***gramática regular***	de xxx : M es un autómata finito de xxx .</s>
<s>Si G es una	***gramática regular***	de xxx , podemos convertir la en una gramática regular G' que genera el mismo lenguaje pero que no
G es una gramática regular de xxx , podemos convertir la en una	***gramática regular***	G' que genera el mismo lenguaje pero que no contiene reglas de reescritura cuyo lado derecho consist
ata finito no determinista , entonces podemos definir G' como la	***gramática regular***	de xxx para la cual los no terminales son los estados de S , el símbolo inicial es xxx y las reglas de re
<s>A diferencia de las	***gramáticas regulares***	, estas gramáticas no tienen restricciones con_respecto_a la forma de l lado derecho de sus reglas de
<s>A l igual que las	***gramáticas regulares***	, las gramáticas independientes de l contexto generan cadenas por medio de derivaciones .</s>
de l contexto generan una mayor colección de lenguajes que las	***gramáticas regulares***	.</s>
<s>A diferencia de las	***gramáticas regulares***	, estas gramáticas no tienen restricciones con_respecto_a la forma de l lado derecho de sus reglas de
<s>A l igual que las	***gramáticas regulares***	, las gramáticas independientes de contexto generan cadenas por medio de derivaciones .</s>
de l contexto generan una mayor colección de lenguajes que las	***gramáticas regulares***	.</s>

¿Cuál es la diferencia, y qué clase de búsqueda les parece que es más eficiente?

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/presentacioacuten.html