PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Introducción a la lingüística computacional

César Antonio Aguilar

Facultad de Lenguas y Letras

12/09/2017

Cesar.Aguilar72@gmail.com

Corpus lingüísticos (1)

Para la lingüística contemporánea, y en particular para un área como PLN, el diseño y uso de corpus lingüísticos se ha convertido en una cuestión importante. Joaquim Llisterri, un fonetista catalán, ofrece el siguiente argumento:

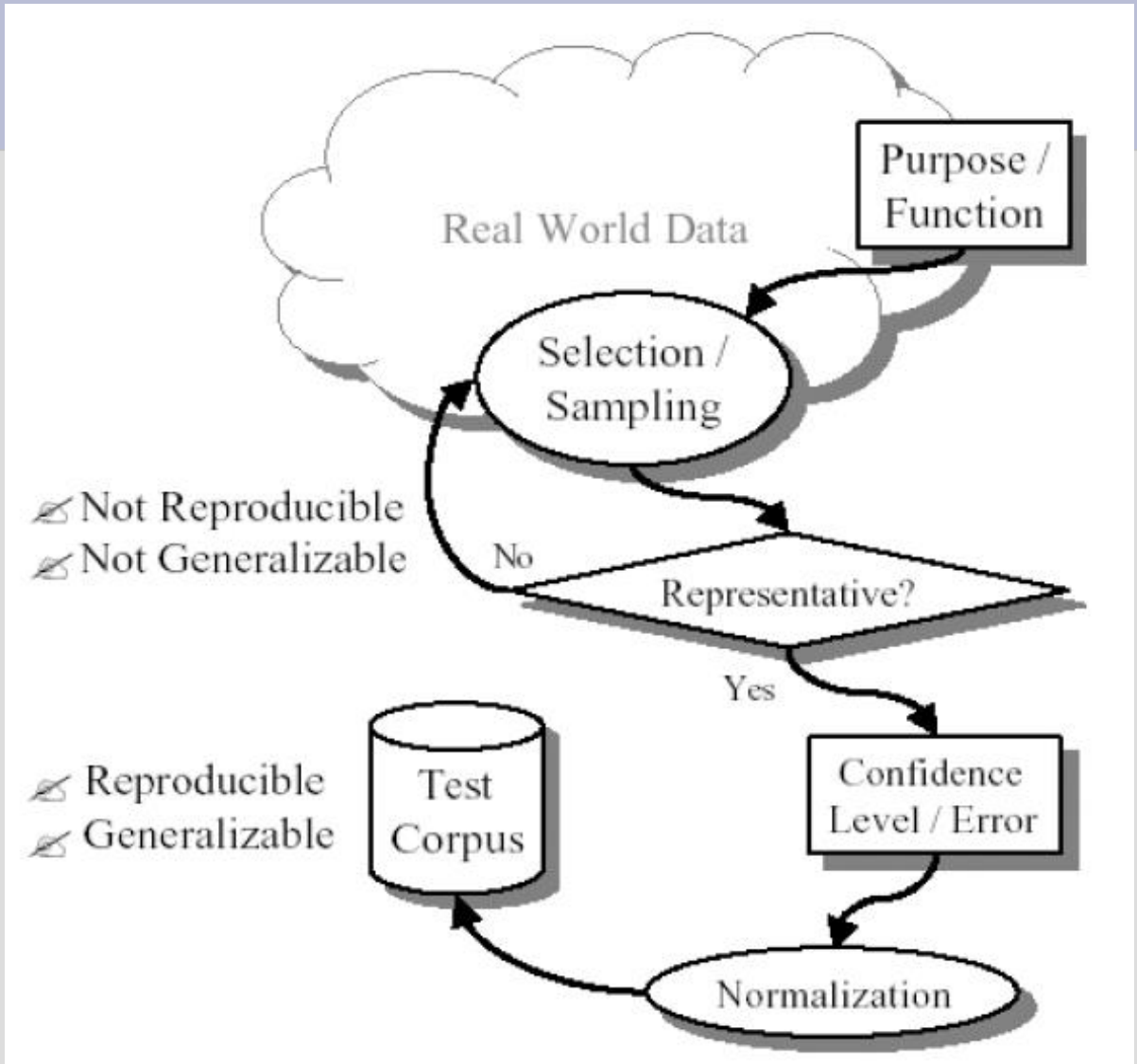
"La función central de los corpus lingüísticos (o CLs) es establecer la relación entre la teoría y los datos, lo que permite hacer hipótesis pertinentes respecto al funcionamiento de una lengua natural".

http://liceu.uab.es/~joaquim/home.html

Corpus lingüísticos (2)

Un corpus lingüístico es una recopilación de un conjunto de materiales escritos y/o hablados, la cual nos sirve para hacer análisis lingüísticos. Los corpus son representativos y se organizan bajo criterios específicos.

Regularmente se encuentran en soporte informático, pues su contenido llega a ser extenso (p. e., millones de palabras).



Corpus lingüísticos (3)

¿Qué cosa no es un corpus?:

Bases de datos	**No**	Como tales, lo que contienen son números, fórmulas o imágenes, que de entrada no aportan hechos lingüísticos.
Archivos bibliográficos, hemerotecas o bibliotecas digitales	**No**	Si bien contienen datos lingüísticos, su organización no está orientada hacia un análisis de esta naturaleza.
Acervos de audio, de video o algún otro material similar	**No**	Si bien existen corpus orales (e incluso pueden vincular videos), si no presentan algún tipo de codificación orientada con fines lingüísticos, no puede verse como un CL.
Internet	**No/Sí**	No es un corpus en un sentido estricto, pues no sigue los criterios para ser constituido como tal. Empero, en un sentido amplio, es una fuente de datos enorme y valiosa, ya que muestra la evolución de una lengua en diferentes registros, volviéndose así un objeto de estudio dinámico.

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 de un corpus lingüístico con anotado textual. Consulten la siguiente página:

www.iling.unam.mx/corpus/

Aquí pueden acceder a un **corpus de contextos definitorios** en el área de ingeniería, desarrollado por el **Grupo de Ingeniería Lingüística** de la **UNAM** (México).

Universidad Nacional Autónoma de México

INSTITUTO DE INGENIERÍA UNAM

Ingeniería lingüística

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

http://grupos.iingen.unam.mx/iling/es-mx/Paginas/default.aspx

Anotación textual (3)

Un **contexto definitorio** (o **CD**) se compone de los siguientes elementos:

Patrón discursivo **Término** **Predicación verbal**

<Matemáticamente>, <la Teoría Lineal> <se considera como> <una primera aproximación de una descripción teórica completa acerca del comportamiento del oleaje.> } **CD**

Definición

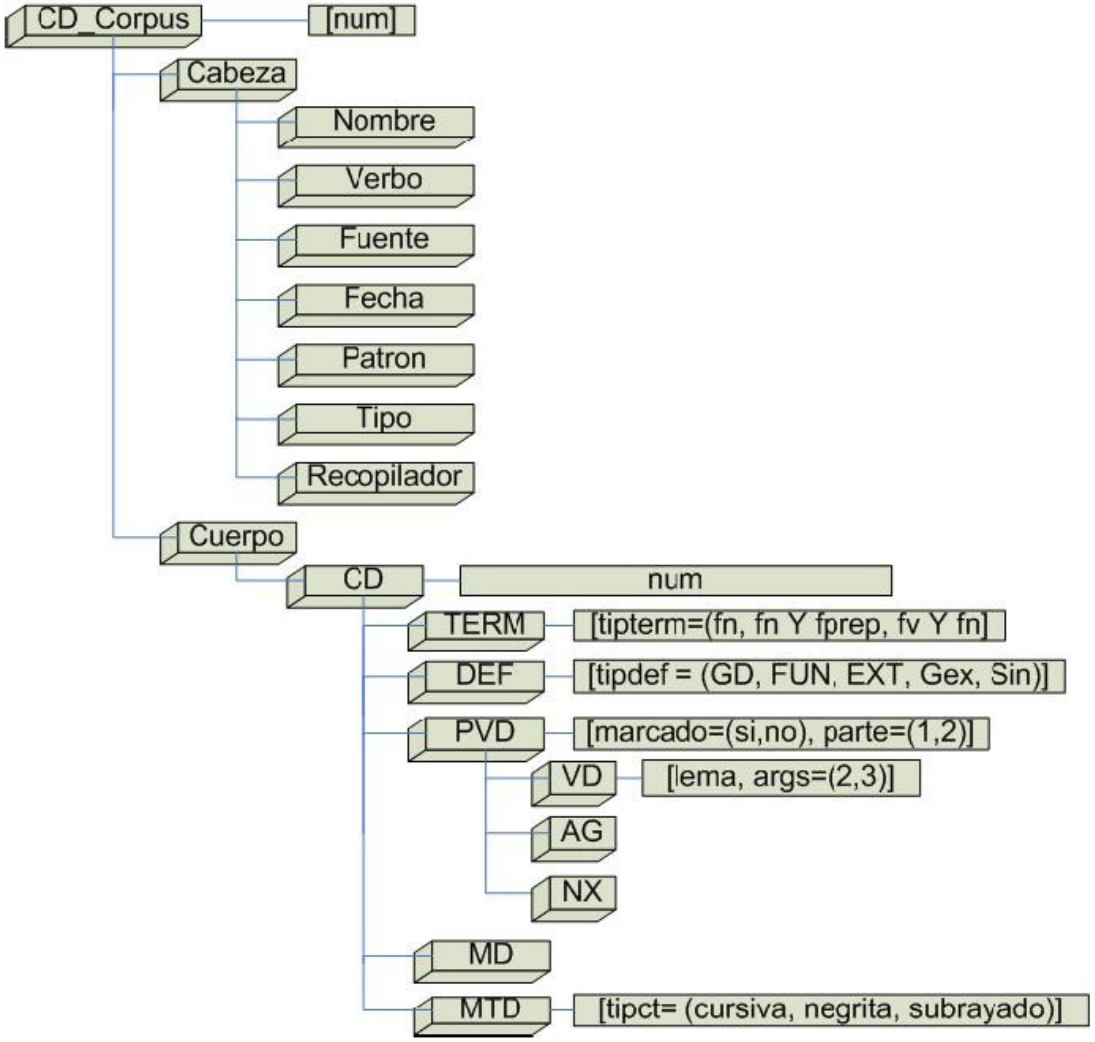
Anotación textual (4)

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)



Head/Body (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.

Head/Body (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"> 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 " </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>

Ordenar/clasificar (4)

El resultado final es:

Buscar

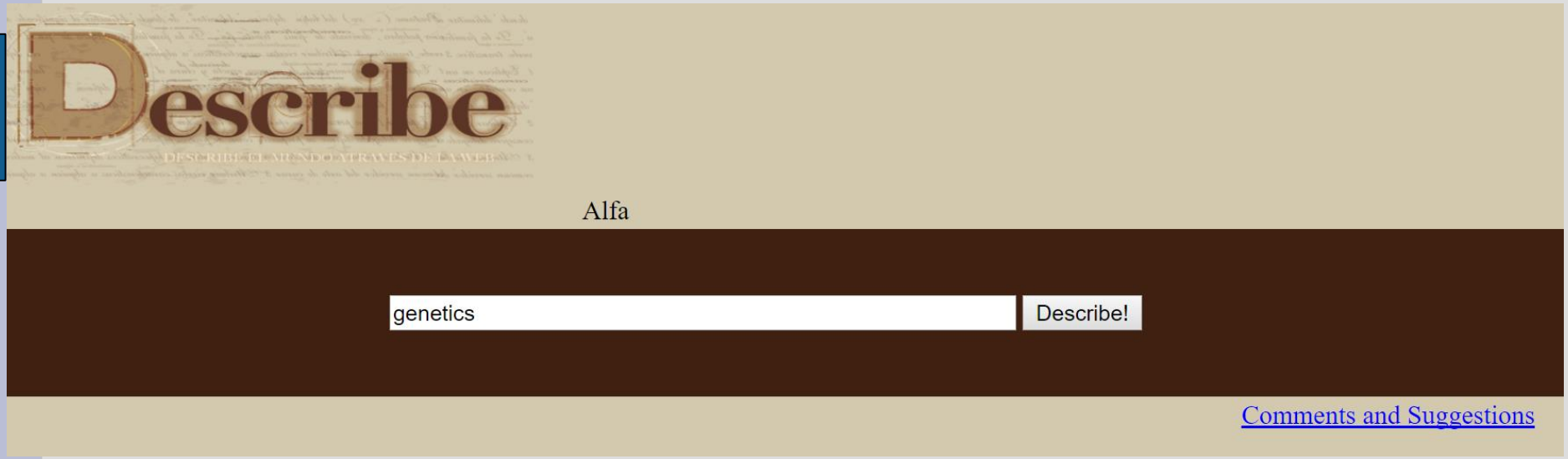
Los siguientes contextos definitorios cumplen su criterio de búsqueda "**Lema=definir. "**

Término
Predicación verbal definitoria
Definición

1	☑ Texto ☑ Atributos	**La longitud se define como el ángulo entre dos planos**; estos planos se llaman planos meridianos: uno contiene al eje de la Tierra y el punto en cuestión, y el otro contiene al eje de la Tierra y a un punto arbitrariamente definido que está localizado en el observatorio de Greenwich en Inglaterra. Tipo término: **lingüístico** Estructura término: **frase nominal** Tipo definición: **sinonímica** Tipo nexo: **adverbio** Tiene marcador SE: **sí** Tiene verbo auxiliar: **no** Tiene marcadores reformulativos: **no** Lema del verbo: **definir**
2	☑ Texto ☐ Atributos	**El ciclo de auditoria se define como "el periodo de tiempo durante el que se someten a auditoria todas las actividades de la organización"** y auditor persona o equipo, perteneciente al personal de la organización o exterior a ella, que actúa en nombre de la alta dirección de la organización y que posee, individual o colectivamente, las competencias indicadas en el punto 2.4 del Anexo II y sea lo suficientemente independiente de las actividades objeto de la auditoria para emitir un dictamen objetivo."
3	☑ Texto ☐ Atributos	**Una sustancia infecciosa es definida como una sustancia que contiene un microorganismo viable, tal como una bacteria, un virus, una rickettsia, un parásito o un hongo, que se sabe o se cree en forma razonable que causa enfermedad en humanos o animales**.

Ordenar/clasificar (5)

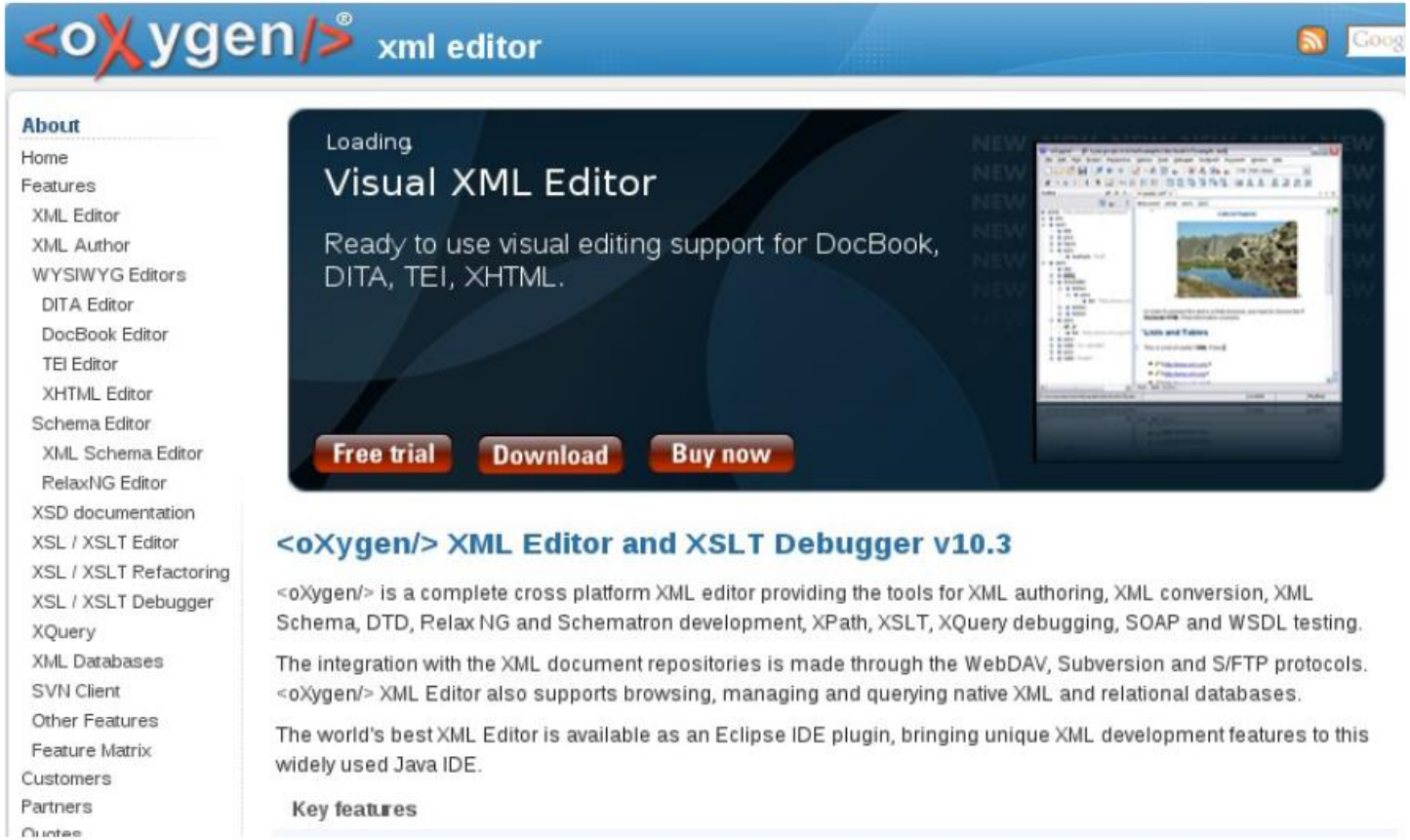
Este corpus puede servir como un **baseline** capaz de entrenar motores de búsqueda como el siguiente:



http://describe.com.mx:8080/_EN6/

Sistemas de etiquetado XML

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



www.oxygenxml.com

Etiquetado morfo-sintáctico (1)

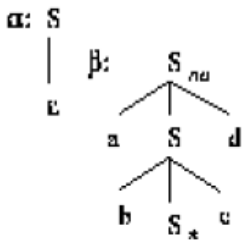
EAGLES: iniciativa de la Comisión Europea (1993-1996), propuesta para el desarrollo y generación de recursos para el procesamiento de lenguaje natural.

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

Etiquetado morfo-sintáctico (2)

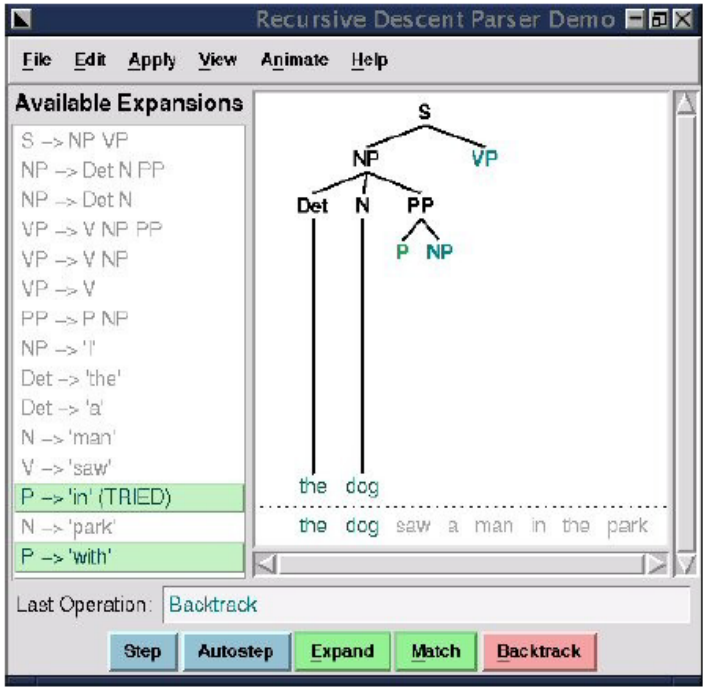
PennTreeBank: Desarrollado por la Universidad de Pensilvania, es una opción para la creación y explotación de análisis sintáctico computacional (*parsing*).



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



El corpus Brown (1)

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 "/" ./.
```

El corpus Brown (2)

El Brown es uno de los corpus electrónicos pioneros que cuenta con una anotación sintáctica. Para saber, vean esta entrada en Wikipedia:

http://en.wikipedia.org/wiki/Brown_Corpus

En la siguiente tabla, podemos ver un ejemplo de su contenido:

ID	File	Genre	Description
A16	ca16	news	Chicago Tribune: *Society Reportage*
B02	cb02	editorial	Christian Science Monitor: *Editorials*
C17	cc17	reviews	Time Magazine: *Reviews*
D12	cd12	religion	Underwood: *Probing the Ethics of Realtors*
E36	ce36	hobbies	Norling: *Renting a Car in Europe*
F25	cf25	lore	Boroff: *Jewish Teenage Culture*
G22	cg22	belles_lettres	Reiner: *Coping with Runaway Technology*
H15	ch15	government	US Office of Civil and Defence Mobilization: *The Family Fallout Shelter*
J17	cj19	learned	Mosteller: *Probability with Statistical Applications*
K04	ck04	fiction	W.E.B. Du Bois: *Worlds of Color*
L13	cl13	mystery	Hitchens: *Footsteps in the Night*
M01	cm01	science_fiction	Heinlein: *Stranger in a Strange Land*
N14	cn15	adventure	Field: *Rattlesnake Ridge*
P12	cp12	romance	Callaghan: *A Passion in Rome*
R06	cr06	humor	Thurber: *The Future, If Any, of Comedy*

El corpus Brown (3)

A través del intérprete de Python, podemos importar el corpus Brown de la librería de NLTK usando la siguiente instrucción:

```
from  nltk.corpus import brown
```

Para saber qué tenemos en el Brown, podemos usar la siguiente instrucción, que nos muestra sus géneros literarios:

```
brown.categories()
```

El corpus Brown (4)

Veamos el resultado:

```
>>> from nltk.corpus import brown
>>> brown.categories()
['adventure', 'belles_lettres', 'editorial', 'fiction', 'government', 'hobbies',
humor', 'learned', 'lore', 'mystery', 'news', 'religion', 'reviews', 'romance',
science fiction']
```

Ahora, probemos las siguientes instrucciones:

1. brown.words(categories='news')

2. brown.words(fileids=['cg22'])

3. brown.sents(categories=['news', 'editorial', 'reviews'])

El corpus Brown (5)

Veamos los resultados:

```
>>> brown.words(categories='news')
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
>>> brown.words(fileids=['cn15'])
['Early', 'in', 'November', 'the', 'clouds', 'lifted', ...]
>>> brown.sents(categories=['news', 'editorial', 'reviews'])
[['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', 'Friday', 'an', 'investiga
tion', 'of', "Atlanta's", 'recent', 'primary', 'election', 'produced', '``', 'no
', 'evidence', "''", 'that', 'any', 'irregularities', 'took', 'place', '.'], ['T
he', 'jury', 'further', 'said', 'in', 'term-end', 'presentments', 'that', 'the',
 'City', 'Executive', 'Committee', ',', 'which', 'had', 'over-all', 'charge', 'o
f', 'the', 'election', ',', '``', 'deserves', 'the', 'praise', 'and', 'thanks',
'of', 'the', 'City', 'of', 'Atlanta', "''", 'for', 'the', 'manner', 'in', 'which
', 'the', 'election', 'was', 'conducted', '.'], ...]
>>> brown.sents(categories=['fiction', 'mystery', 'humor'])
[['Thirty-three'], ['Scotty', 'did', 'not', 'go', 'back', 'to', 'school', '.'],
...]
>>> |
```

De acuerdo con esto, podemos consultar el contenido del Brown usando dos clases de objetos: palabras y oraciones, y mostrar en pantalla los resultados de nuestras consultas.

El corpus Brown (6)

Veamos un código un poco más complicado, con el cual podremos ver la distribución de verbos modales dentro del Brown. Para ello, necesitamos importar de la librería de NLTK la función Conditional FreqDist (o distribución de frecuencia condicional). Veamos:

```
cfd_brown01 = nltk.ConditionalFreqDist(
        (genre, word)
        for genre in brown.categories()
        for word in brown.words(categories=genre))
```

El corpus Brown (7)

Una vez que hemos especificado que queremos calcular frecuencias de palabras dentro de los textos del Brown (considerando su género literario), ahora lo que debemos haces es especificar los objetos que vamos a buscar y contar, esto es:

1. genres = ['news', 'religion', 'hobbies', 'science-fiction', 'romance', 'humor']

2. modals = ['can', 'could', 'may', 'might', 'must', 'will', 'shall']

El corpus Brown (8)

Y el resultado es:

```
>>> cfd_brown01.tabulate(conditions=genres, samples=modals)
                 can could  may might must will shall
            news  93   86   66    38   50  389     5
        religion  82   59   78    12   54   71    21
         hobbies 268   58  131    22   83  264     5
science_fiction  16   49    4    12    8   16     3
         romance  74  193   11    51   45   43     3
           humor  16   30    8     8    9   13     2
>>>
```

Gracias por su atención

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