



# Seminario de análisis del discurso

**Dr. César Antonio Aguilar**  
**Facultad de Lenguas y Letras**  
**06/09/2010**

**CAguilar@ingen.unam.mx**

# Introducción (1)

El diseño y empleo de **corpus lingüísticos** (o CLs) es uno de los métodos de trabajo que ha cobrado gran relevancia para la lingüística actual.

Joaquim Llisterri, un fonetista de la Universidad Autónoma de Barcelona (UAB), ofrece el siguiente argumento:

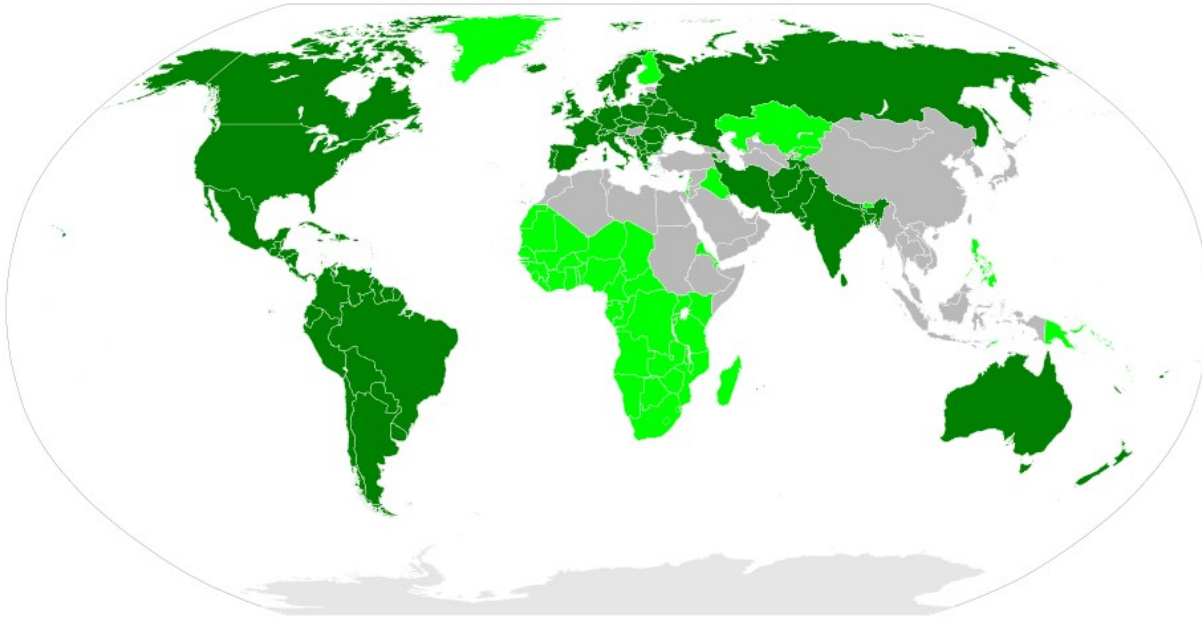
“La función central de los CLs es establecer la relación entre la teoría y los datos, lo que permite hacer hipótesis pertinentes respecto al funcionamiento de una lengua natural”.



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

## Introducción (2)

Si bien es cierto que en los últimos años el análisis de corpus ha cobrado una relevancia importante, esto no quiere decir que se trata de un área nueva en la lingüística.



**Nota:** en este mapa se muestra los países que cuentan con grandes (verde oscuro) y pequeñas (verde claro) poblaciones de hablantes de una lengua indoeuropea.

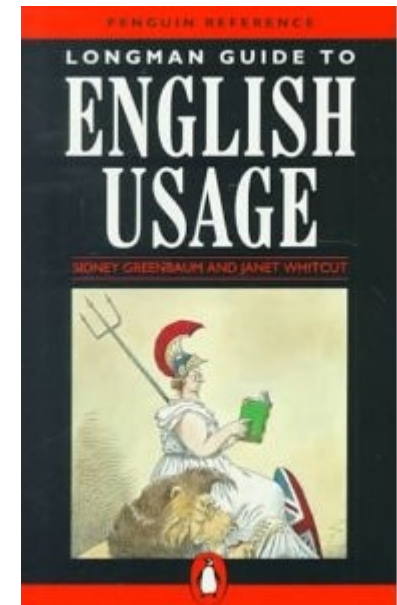
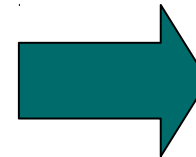
A lo largo del siglo XIX, y en particular a raíz del interés por la filología europea por el análisis de textos sánscritos, se ponderó la recopilación de datos lingüísticos que permitieran la reconstrucción de una “proto-lengua”, la cual sería la base de todas las lenguas pertenecientes a la familia indoeuropea.

## Introducción (3)

Si bien estos trabajos no lograron reconstruir esta “lengua original”, muchas de los métodos usados dieron lugar a toda una serie de estudios enfocados a la recopilación y documentación de datos lingüísticos para fines tales como la historiografía lingüística, la dialectología, la lexicografía, la sociolingüística, la estilística, y otras similares.

Un ejemplo es el trabajo de Sir Randolph Quirk (1920), enfocado en el estudio y la comparación de distintas variedades y usos del inglés británico.

En 1960 elaboró una base de datos de un millón de palabras, la cual le permitió hacer cálculos estadísticos para validar sus ideas sobre variación dialectal.

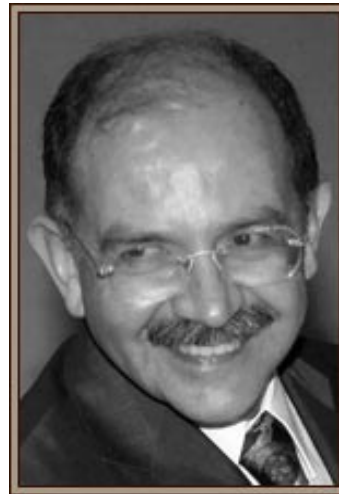
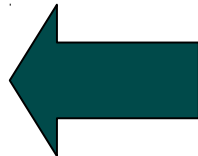
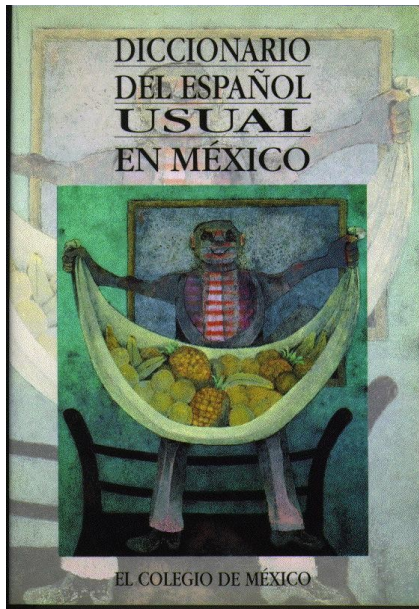


**Nota:** para mayores detalles, consulten la siguiente página WEB:

[www.ucl.ac.uk/english-usage/index.htm](http://www.ucl.ac.uk/english-usage/index.htm)

## Introducción (5)

Otro ejemplo relevante es el *Diccionario del español usual de México* (1973-a la fecha), dirigido por Luis Fernando Lara (1943), el cual es un diccionario cuyos vocablos y entradas léxicas fueron deducidas a partir de datos lingüísticos obtenidos del **Corpus del Español de México** (o **CEM**).



El CEM puede ser considerado como el primer corpus latinoamericano que se compiló para ser procesado a través de sistemas de cómputo, lo que le da a este diccionario un valor especial: sus definiciones se basan en un análisis sobre el uso real de los hablantes sobre su vocabulario.

**Nota:** para consultar este diccionario en línea, pueden acceder a esta liga:

<http://mezcal.colmex.mx/dem/>

# Introducción (5)

En los últimos años, el gran impulso que ha recibido el análisis de corpus ha venido del procesamiento del lenguaje natural (PLN), lo que ha permitido no sólo construir corpus mucho más grandes, sino desarrollar herramientas computacionales para hacer análisis en todos los niveles lingüísticos, incluido el discursivo.

Un ejemplo de esta evolución es **Corpus de Referencia del Español Actual**, elaborado por la Real Academia Española, el cual contiene 154 millones de palabras, así como una serie de herramientas para hacer consultas lingüísticas en textos escritos recopilados de 1975 al 2004.

Real Academia Española - Corpus de Referencia del Español Actual (CREA)

Consulta:

Criterios de selección:

Autor: <input type="text"/>	Obra: <input type="text"/>	
Cronológico: <input type="text"/> <input type="text"/>	Medio: (Todos) Libros Periódicos Revistas Miscelánea Oral	Geográfico: (Todos) Argentina Bolivia Chile Colombia Costa Rica
Tema: (Todos) 1.- Ciencias y Tecnología. 101.- Biología. 102.- Veterinaria. 103.- Ecología. 104.- Tecnología.		

[Consulta CORDE](#) [Nómina de autores y obras](#) [Lista de frecuencias](#) [Cómo citar el CORPUS](#) [Ayuda.](#)

<http://corpus.rae.es/creanet.html>



# Definiendo un CL

Una manera breve de definir lo que es un CL es la siguiente:

**AMERICAN NATIONAL CORPUS**

**B**RITISH **N**ATIONAL **C**ORPUS

**The Brown Corpus**

**CHEM** Corpus Histórico del Español de México

Child Language Data Exchange System

**TIGERCORPUS**

Recopilación de un conjunto de materiales escritos y/o hablados para realizar análisis lingüísticos.

Son representativos y organizados bajo criterios específicos.

Regularmente se encuentran en soporte informático, pues su contenido llega a ser extenso, incluso de varios millones de palabras.

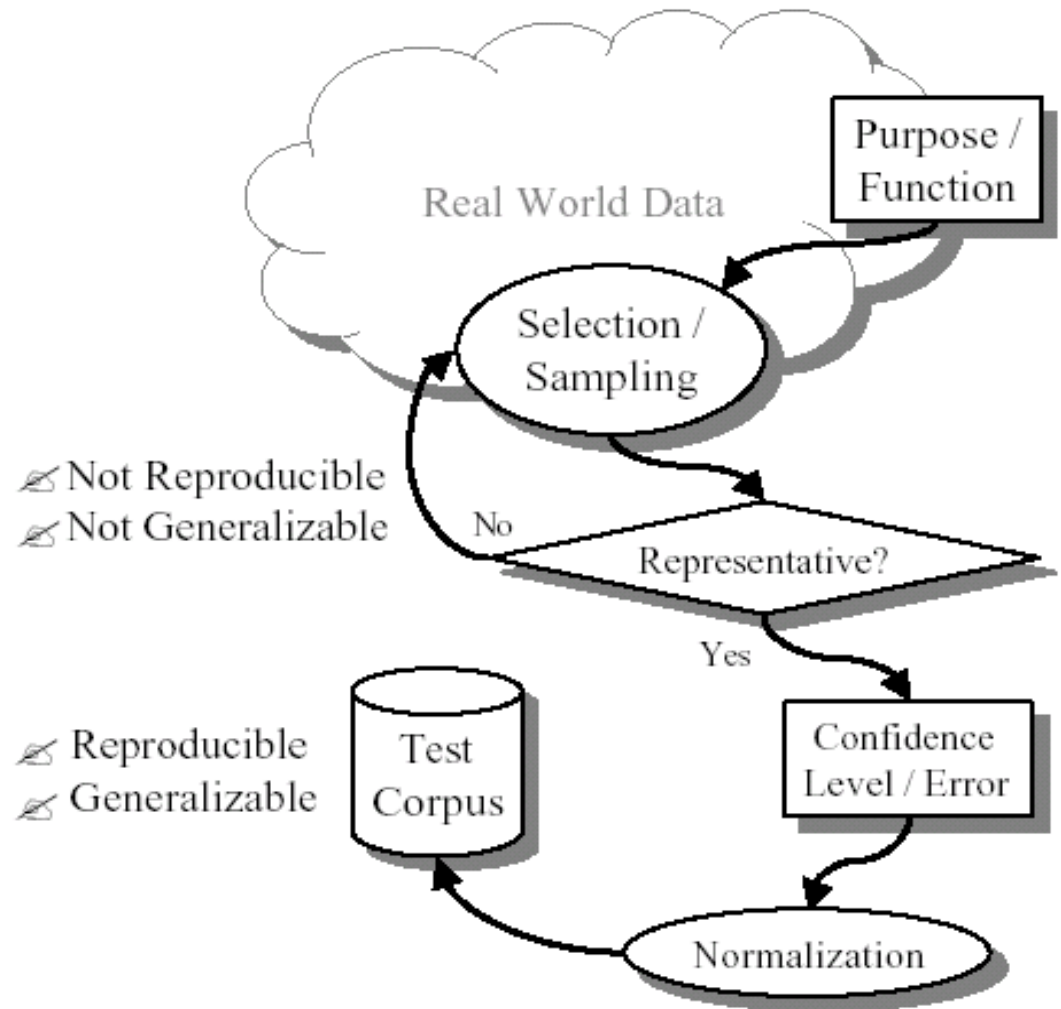
# Utilidad de un CL

Se trata de un modelo que representa una realidad lingüística.

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.

Finalmente, es un Instrumento reutilizable para distintos tipos de análisis.





## Justificación: ¿por qué son necesarios?

- Los CLs, al ofrecer una base empírica sólida sobre el funcionamiento de una lengua, permiten desarrollar modelos teóricos acordes con una realidad específica.
- Del mismo modo, son un valioso repositorio de datos que pueden ser consultados desde cualquier lugar, e igualmente mostrar la evolución de una lengua a lo largo de la historia.
- Lo anterior ayuda a la conservación de datos de lenguas en peligro de extinción, pues quedan almacenados en un soporte electrónico seguro.
- Apoya la interacción entre lingüistas de diferentes áreas, así como de otros investigadores (computólogos, antropólogos, historiadores, estadísticos, etc.), al compartir recursos y datos útiles para distintos análisis.

# Tipos de CLs (1)

<b>Origen de los textos</b>	Orales Textuales
<b>Especificidad de los textos</b>	Generales Específicos Genérico Canónico Cronológico
<b>Según la lengua</b>	Monolingüe multilingüe Paralelo Comparable
<b>Cantidad de texto recogido</b>	Textual Referencia Léxico
<b>Codificación y anotación</b>	Simple Anotado

## Tipos de CLs (2)

<b>Distribución de textos</b>	Equilibrado Desequilibrado Monitor
<b>Propósito</b>	Multipropósito Específico
<b>Disponibilidad</b>	Público Privado
<b>Documentación</b>	Documentado No documentado
<b>Variedad dialectal</b>	Mono-dialectal Multi-dialectal

# ¿Qué cosas no son CLs?

<b>Bases de datos</b>	<b>No</b>	Como tales, lo que contienen son números, fórmulas o imágenes, que de entrada no aportan hechos lingüísticos.
<b>Archivos bibliográficos, hemerotecas o bibliotecas digitales</b>	<b>No</b>	Si bien contienen datos lingüísticos, su organización no está orientada hacia un análisis de esta naturaleza.
<b>Acervos de audio, de video o algún otro material similar</b>	<b>No</b>	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.
<b>Internet</b>	<b>No/Sí</b>	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.

# Construyendo un CL

## 1. Objetivos

### 1.1. Propósitos

### 1.2. Límites

## 2. Selección de textos

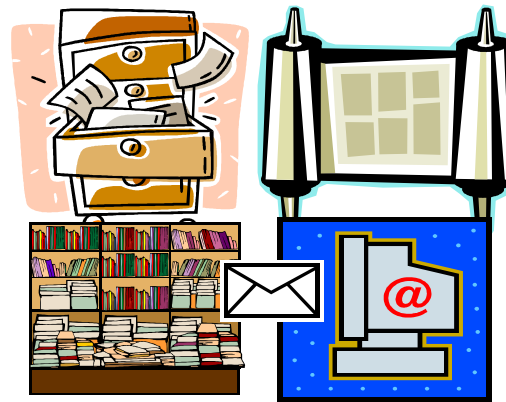
## 3. Obtención de textos

### 3.1. Digitalización

### 3.2. Documentos electrónicos

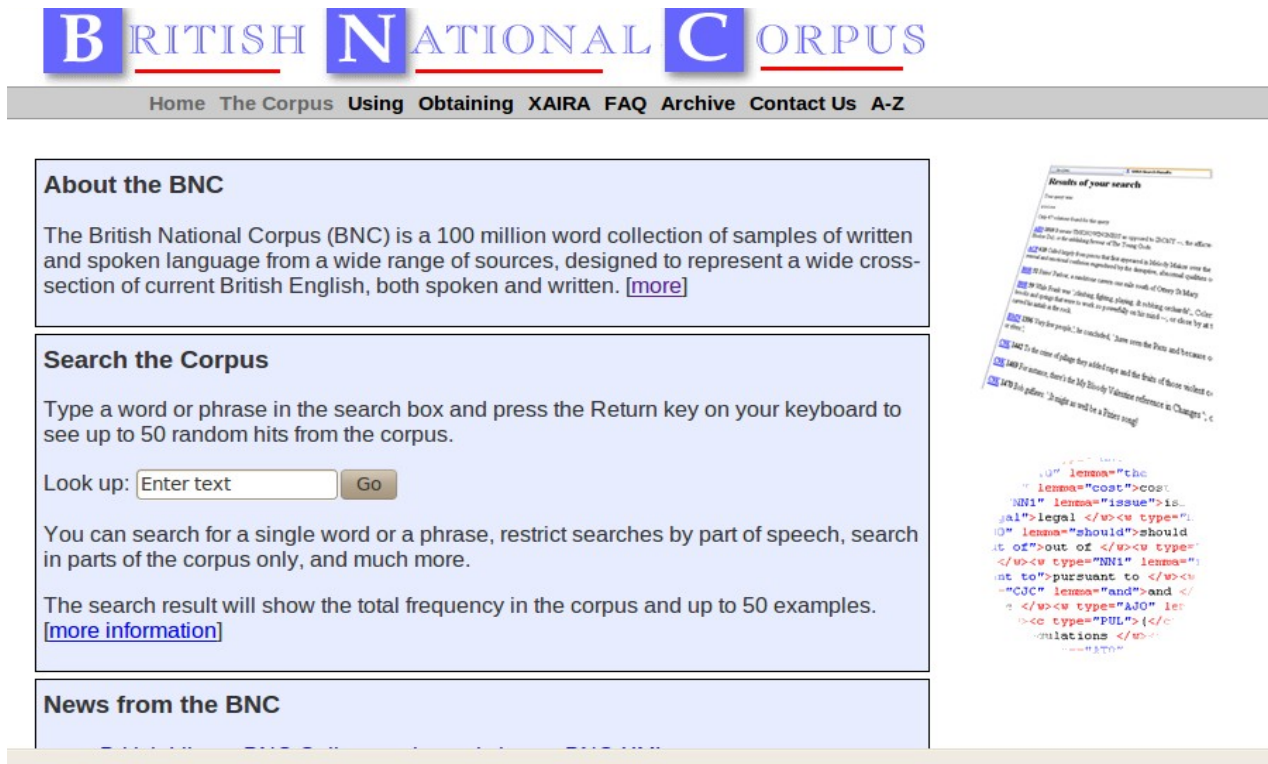
## 5. Administración

## 6. Sistemas de consulta



# Un ejemplo de corpus: BNC (1)

Veamos ahora un ejemplo de un CL bastante conocido: el **British National Corpus** (o **BCL**).



The screenshot shows the homepage of the British National Corpus. At the top, the logo reads "BRITISH NATIONAL CORPUS" with each word in a blue box. Below the logo is a navigation bar with links: Home, The Corpus, Using, Obtaining, XAIRA, FAQ, Archive, Contact Us, and A-Z. The main content area is divided into three sections:

- About the BNC:** A text box explaining that the BNC is a 100 million word collection of written and spoken language from various sources, designed to represent a wide cross-section of current British English. It includes a link to "[more]".
- Search the Corpus:** A section with a search box labeled "Look up:" containing the text "Enter text" and a "Go" button. Below the search box, it states: "Type a word or phrase in the search box and press the Return key on your keyboard to see up to 50 random hits from the corpus." It also mentions that users can search for single words or phrases, restrict searches by part of speech, and search in parts of the corpus only. A note says: "The search result will show the total frequency in the corpus and up to 50 examples." with a link to "[more information]".
- News from the BNC:** A section with a heading "News from the BNC" and a list of news items, though the text is partially obscured.

On the right side of the screenshot, there is a preview of search results titled "Results of your search". It shows a list of search results with their corresponding text snippets. Below this, there is a snippet of XML-like code showing word frequency and part-of-speech tags, such as "lemma='the'", "lemma='cost'", "lemma='issue'", "lemma='legal'", "lemma='should'", "lemma='out of'", "lemma='pursuant to'", "lemma='and'", "lemma='AJO'", "lemma='PUL'", and "lemma='ulations'".

**Nota:** si desean consultarlo posteriormente, accedan al sitio:

[www.natcorp.ox.ac.uk/](http://www.natcorp.ox.ac.uk/)



## Un ejemplo de corpus: BNC (2)

El BCL es un corpus que cuenta con 100 millones de palabras obtenidas de distintos textos escritos, junto con transcripciones orales, previamente seleccionados, el cual da una muestra bastante amplia de diversos registros léxicos, sintácticos, semánticos y discursivos del inglés británico contemporáneo.

*Table 3. Written Domain*

	<b>texts</b>	<b>w-units</b>	<b>%</b>	<b>s-units</b>	<b>%</b>
<b>Imaginative</b>	476	16496420	18.75	1352150	27.10
<b>Informative: natural &amp; pure science</b>	146	3821902	4.34	183384	3.67
<b>Informative: applied science</b>	370	7174152	8.15	356662	7.15
<b>Informative: social science</b>	526	14025537	15.94	698218	13.99
<b>Informative: world affairs</b>	483	17244534	19.60	798503	16.00
<b>Informative: commerce &amp; finance</b>	295	7341163	8.34	382374	7.66
<b>Informative: arts</b>	261	6574857	7.47	321140	6.43
<b>Informative: belief &amp; thought</b>	146	3037533	3.45	151283	3.03
<b>Informative: leisure</b>	438	12237834	13.91	744490	14.92

**Pregunta 1:** con esta información, si quisiéramos clasificar el BNC, ¿qué clase de corpus sería?

## Un ejemplo de corpus: BNC (3)

Un rasgo particular del BNC es que se trata de un corpus anotado, el cual permite hacer búsquedas a través de una interfaz, con miras a identificar grupos de palabras o frases. En este caso, se ocupan etiquetas **XML** (Extensible Markup Language), las cuales permiten identificar categorías de palabras dentro de cada texto.

La idea de utilizar XML responde a una necesidad: darle un formato adecuado a los datos lingüísticos que contiene el BNC, con miras a explotarlo de forma automática.

```
<div type="u">
  <head type="MAIN">
    <s n="835">
      <w c5="AJ0" hw="serious" pos="ADJ">Serious </w>
      <w c5="NN1" hw="fit" pos="SUBST">fit </w>
      <w c5="PRF" hw="of" pos="PREP">of </w>
      <w c5="NN2" hw="giggle" pos="SUBST">giggles</w>
    </s>
  </head>
  <p>
    <s n="836">
      <w c5="AT0" hw="a" pos="ART">A </w>
      <w c5="NN0" hw="pair" pos="SUBST">PAIR </w>
      <w c5="PRF" hw="of" pos="PREP">of </w>
      <w c5="NN1" hw="tv" pos="SUBST">TV </w>
      <w c5="NN2" hw="newsreader" pos="SUBST">newsreaders </w>
    </s>...</p> ... </div>
```

# Anotación lingüística

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. Básicamente, hay 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*).

# Etiquetado textual

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>
```

# Etiquetado lingüístico

## Fonético

0002 Participant: 4

Orthography **c o o k i e b r o k e n**

Target IPA **ˈ k u w k i ˈ b r o w k n**

Actual IPA **ˈ k u k i ˈ b o w k**

Phrases	Orthography	IPA Target	IPA Actual
cookie		ˈkuwki	ˈkuki
broken		ˈbrowkɹ	ˈbowk

◀ prev next ▶ 1

## Discursivo

(0) The state Supreme Court has refused to release (1[2 Rahway State Prison 2] inmate 1)) (1 James Scott 1) on bail .

(1 The fighter 1) is serving 30-40 years for a 1975 armed robbery conviction . (1 Scott 1) had asked for freedom while <1 he waits for an appeal decision. Meanwhile , [3 <1 his promoter 3] , {[3 Murad Muhammed 3] , said Wednesday <3 he netted only \$15,250 for (4 [1 Scott 1] 's nationally televised light heavyweight fight against (5 ranking contender 5)) (5 Yaqui Lopez 5) last Saturday 4) .

## Semántico (léxica)

```
>>> N['dog']
dog(n.)
>>> N['dog'].getSenses()
{'dog' in {noun: dog, domestic dog, Canis familiaris},
 'dog' in {noun: frump, dog}, 'dog' in {noun: dog},
 'dog' in {noun: cad, bounder, blackguard, dog, hound, heel},
 'dog' in {noun: pawl, detent, click, dog},
 'dog' in {noun: andiron, firedog, dog, dogiron}}
```

## Pragmático (diálogo)

Dialogue: d93-18.1

Number of utterances files: 65

Length of dialogue: 184.301076

Estimated number of turns: 59

utt1 : s: hello <sil> can I help you

utt2 : u: um okay I need to transport <sil>

four boxcars of oranges to Bath by two p.m.

# Etiquetado de partes de la oración (1)

Etiquetado **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 de partes de la oración (2)

Etiquetado **PennTreeBank**: Desarrollado por el Laboratorio de Lingüística Computacional de la Universidad de Pensilvania, se plantea como una opción para la creación y explotación de análisis sintáctico computacional (*parsing*).

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 <i>there</i>	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	<i>wh</i> -determiner
10. LS	List item marker	34. WP	<i>wh</i> -pronoun
11. MD	Modal	35. WP\$	Possessive <i>wh</i> -pronoun
12. NN	Noun, singular or mass	36. WRB	<i>wh</i> -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

The screenshot shows a window titled "Recursive Descent Parser Demo" with a menu bar (File, Edit, Apply, View, Animate, Help). The main area is divided into two panes. The left pane, titled "Available Expansions", lists various grammar rules such as "S -> NP VP", "NP -> Det N PP", and "VP -> V NP PP". The right pane displays a partial parse tree for the sentence "the dog saw a man in the park". The root node is "S", which branches into "NP" and "VP". The "NP" node branches into "Det" (labeled "the") and "N" (labeled "dog"). The "VP" node branches into "V" (labeled "saw") and "NP". This second "NP" node branches into "P" (labeled "in") and "NP". The third "NP" node branches into "P" (labeled "with") and "NP". Below the tree, the words "the dog saw a man in the park" are displayed, with "the dog" and "in the park" aligned under their respective nodes in the tree. At the bottom of the window, there is a "Last Operation:" field containing "Backtrack" and a row of five buttons: "Step", "Autostep", "Expand", "Match", and "Backtrack".

# Métodos de análisis

## Basados en reglas formales

((Takes NP SBar) (Type 2))

((Takes NP NP Inf) (Type 2 ORaising))

(or ((Takes NP NP NP) (Type 2 ORaising)))

((Takes NP NP AuxInf) (Type 2 ORaising)))

(or ((Takes NP NP AP) (Type 2 ORaising)))

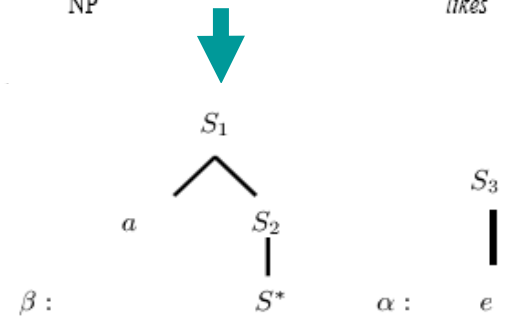
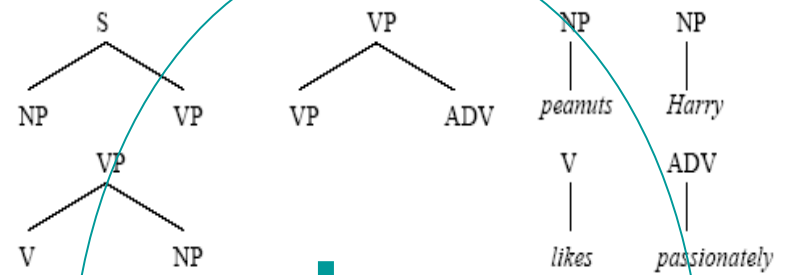
((Takes NP NP AuxInf) (Type 2 ORaising)))

## Estadísticos

$a_{ij} :: B_{ij}$	forms	$p(b_{kj}   a_{ij})$	$-p * \log(p)$
p::a	1365	0.189425	0.31518
p::b	1	0.000138773	0.00123268
p::c	2	0.000277546	0.00227297
p::d	1	0.000138773	0.00123262
p::e	1396	0.193727	0.317965

## Híbridos

CFG G    S → NP VP                    NP → Harry  
           VP → V NP                    NP → peanuts  
           VP → VP ADV                V → likes  
   ADV → passionately



$\phi(S_1 \rightarrow \beta) = 0.99$   
 $\phi(S_1 \rightarrow \epsilon) = 0.01$   
 $\phi(S_2 \rightarrow \beta) = 0.98$   
 $\phi(S_2 \rightarrow \epsilon) = 0.02$   
 $\phi(S_3 \rightarrow \beta) = 1.0$   
 $\phi(S_3 \rightarrow \epsilon) = 0.0$

# Técnicas

## Conteo de palabras

the	4194
and	2976
I	2847
of	2641
to	2094
my	1777
a	1391
in	1129
was	1021

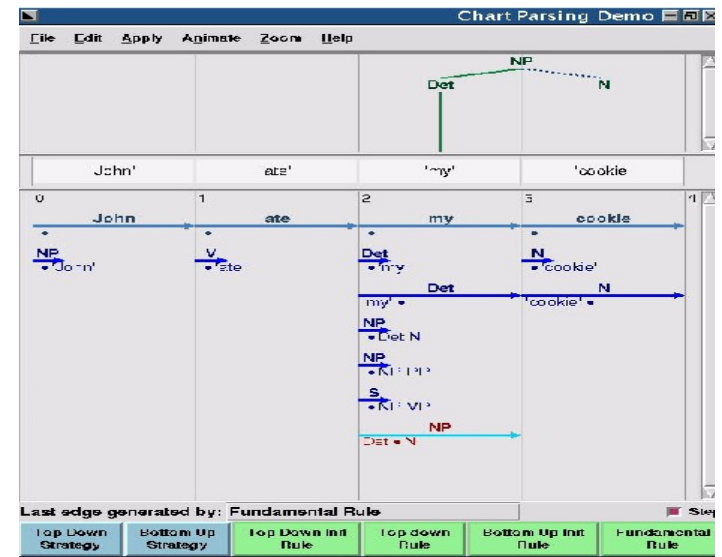
## Medidas de asociación

USED	7	3	2	0	0	device	54	4
CONSISTING	0	3	1	0	0	device	22	2
ELECTRICAL	3	3	9	3	27	device	0	0
MECHANICAL	2	1	5	4	35	device	0	1
PERSON	5	0	0	6	4	device	0	0
CALLED	6	9	6	0	0	device	0	2
SIMILAR	1	1	1	3	35	device	0	0
COMPUTER	6	8	9	5	1	device	0	0

## Concordancias

ocabulario terminal de la *gramática* o lo que es lo mismo, o a través de una *gramática* .</s>  
 ocabulario terminal de la *gramática* ).</s>  
 algunos restringiendo las *gramáticas* utilizadas y otras construidas a partir de la *gramática* , y que dirigen , e  
 tabla depende sólo de la *gramática* , se realiza sólo c  
 <s>Las reglas de la *gramática* son de l tipo de la

## Chunking y Parsing



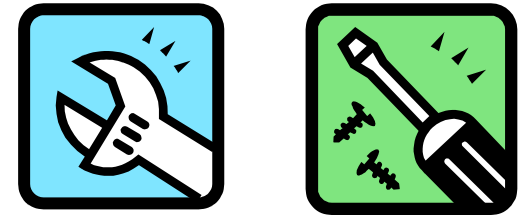
# Herramientas (1)

Otro aspecto importante es la creación y uso de herramientas computacionales.

Dependiendo de los fines de cada investigación, así como de los métodos y técnicas considerados es como se determina qué clase de herramienta se va a ocupar.

Existen dos vías:

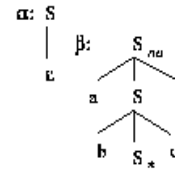
1. Uso de herramientas disponibles (sean de acceso libre o restringido).
2. Empleo de plataformas y lenguajes de cómputo para su desarrollo.



# Herramientas (2)

Etiquetado

XMLWriter



The XTAG Project

TIGER Corpus

Annotation

Plataformas



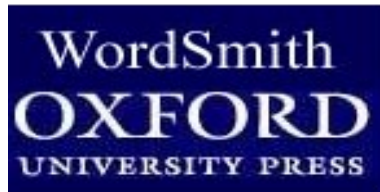
Lenguajes y desarrollos



perl

Natural Language Toolkit

Herramientas de análisis



WordNet

# Gracias por su atención

**Blog del curso:** <http://discurso-uaq.weebly.com/>