UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

Seminario de análisis del discurso

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

CAguilar@iingen.unam.mx

Anotación lingüística (1)

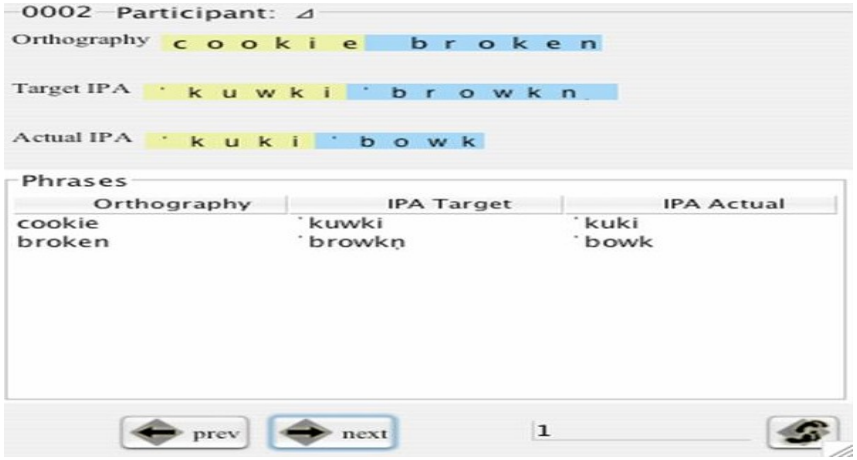
Algo que marca una diferencia fundamental entre tener una colección de textos o fragmentos y un CL, es su anotación o etiquetado.

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*).

Anotación lingüística (2)

Fonético



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 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>
```

Etiquetado textual (2)

Veamos un caso de la vida real: supongamos que queremos identificar en una colección de documentos técnicos términos y definiciones como las del ejemplo:

50	define	volver (12) El Minvu **define** Gobernabilidad Urbana como : identificar tareas a realizar y quien las hará , las funciones que deben descentralizarse y las que deben centralizarse .

1. Proceso de extracción: supongamos que tenemos un buscador que explora en textos predicaciones que ligan un término con una definición, siguiendo un patrón más o menos “normal”. Empero:

8	conoce	Conexión bloque o unidad Cuando los generadores se encuentran conectados al bus con un transformdor de por medio , entonces se dice , que cada generador forma con cada transformador una unidad o bloque , por lo que a esta conexión se le **conoce** como bloque o unidad

No siempre es fácil.

Interludio (1): contexto definitorio

Patrón pragmático

Término

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

CD

Predicación verbal

Definición

Interludio (2): contexto definitorio

¿Qué es una *definición*?

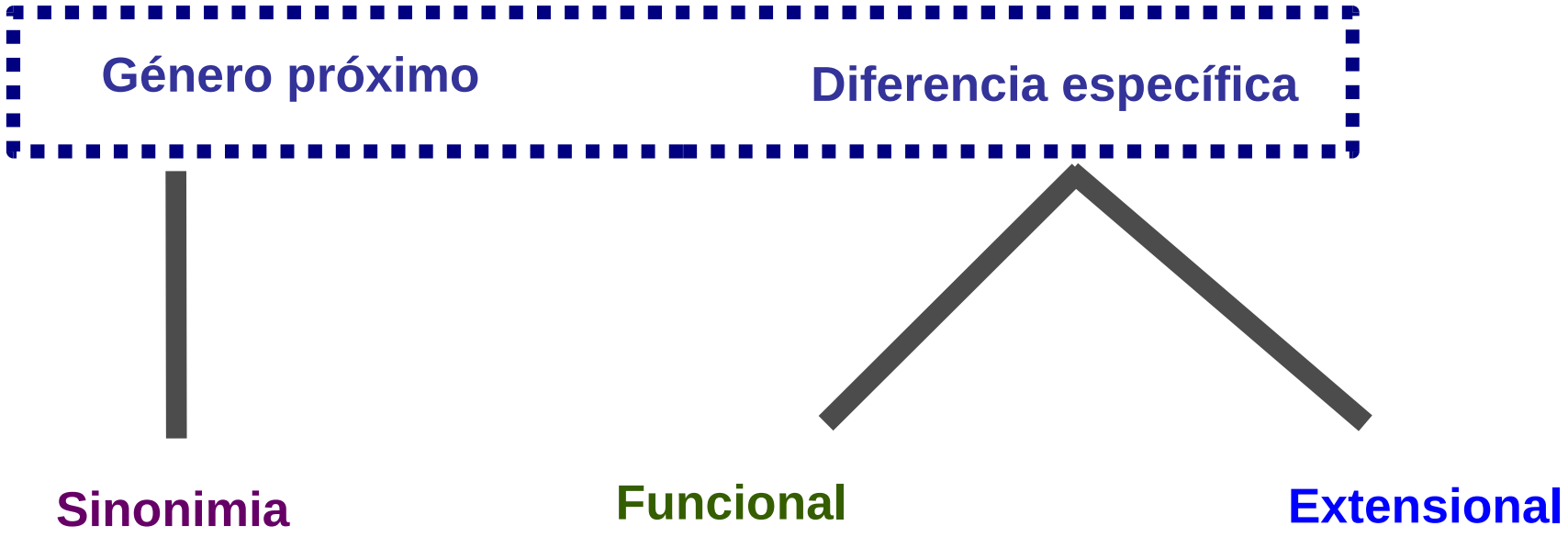
Expresión lingüística de un concepto asociado a un término. Se estructura en torno a dos unidades básicas: un género próximo y una diferencia específica.

Género próximo

Inicialmente podemos considerar un buque como un flotador que trata de permanecer en posición vertical frente a perturbaciones exteriores.

Diferencia específica

Interludio (3): contexto definitorio

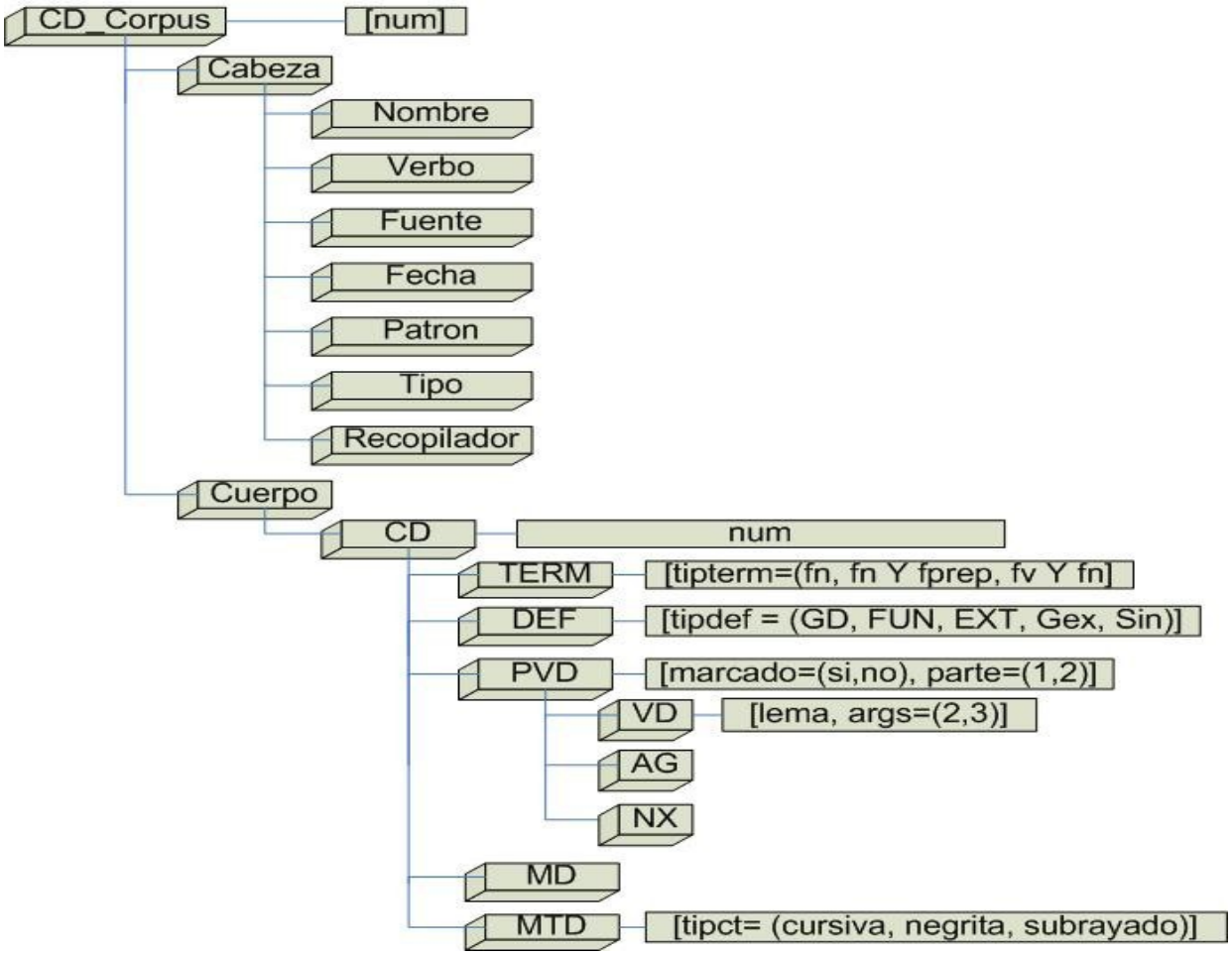


Buscando CDs (1)

Esta búsqueda de términos y definiciones ligados (que podemos llamar *contextos definitiorios* (o Cds), plantea el siguiente problema:

1. Tengo un conjunto de potenciales CDs obtenidos de diferentes textos, pero tales Cds no están organizados
2. Si quiero desplegarlos en mi pantalla, necesito estructurarlos de una forma clara, con miras a identificar qué cosa es un término y qué cosa es una definición.
3. En concreto: necesito construir un corpus de CDs, el cual me muestre por lo menos tres unidades importantes: un término, una definición y una frase verbal.

Buscando CDs (2)



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

Buscando CDs (3)

Toda hoja que etiquetamos en XML debe considerar dos partes: una **cabeza** y un **cuerpo**. Estas serían las etiquetas para la cabeza (ing. *Head*):

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.

Buscando CDs (4)

Y estas son las partes que conforman el cuerpo (ing. *Body*):

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.

Resultados (1)

¿Y qué obtenemos tras aplicar etiquetas XML a un documento? Veamos el siguiente ejemplo. Primero, un corpus 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 "

Resultados (2)

Apliquemos nuestras primeras etiquetas: ¿dónde inicia y dónde termina un CD?:

<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>

Resultados (3)

<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>

Resultados (4)

<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> como </PVD>: "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>

Resultados (5)

<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>

Aplicaciones

Veamos ahora dos aplicaciones.

1. Una mera consulta a un corpus de contextos definitorios:

http://linux.iingen.unam.mx/iling/Bpublica/V/Varios_MuestraCorpusCDs.pdf

2. Un buscador automático de CDs en colecciones de documentos técnicos:

http://brangaene.upf.edu/ecode/

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

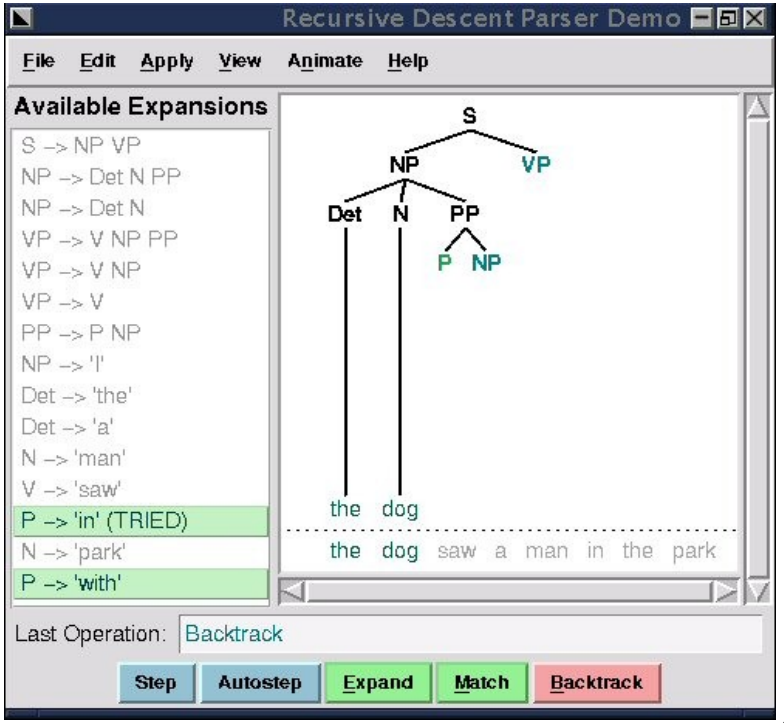
Forma	Lema	Etiqueta
chico	chico	NCMS000
chicos	chico	NCMP000
chica	chica	NCFS000
chicas	chica	NCFP000
oyente	oyente	NCCS000
oyentes	oyente	NCCP000
cortapapeles	cortapapeles	NCMN000
tesis	tesis	NCFN000
Antonio	antonio	NP00000

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 *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



Métodos de análisis

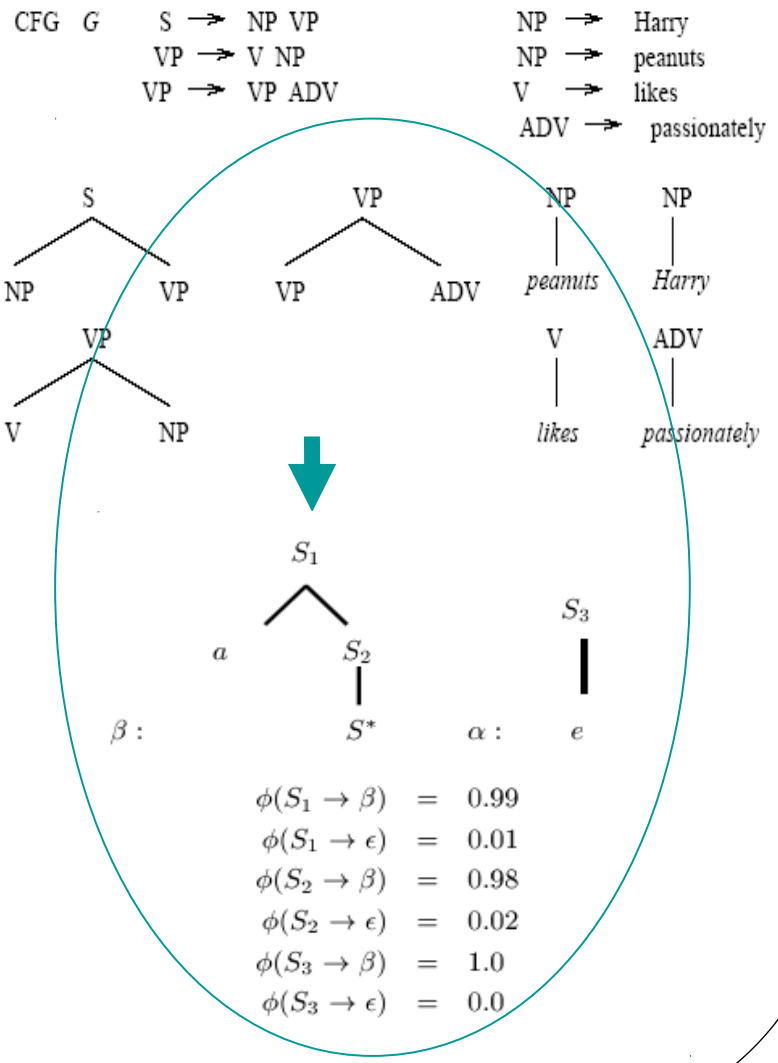
Basados en reglas formales

```
((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

$a_{i,j}::B_{i,j}$	forms	$p(b_{k,j} \mid a_{i,j})$	-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



Técnicas

Conteo de palabras

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

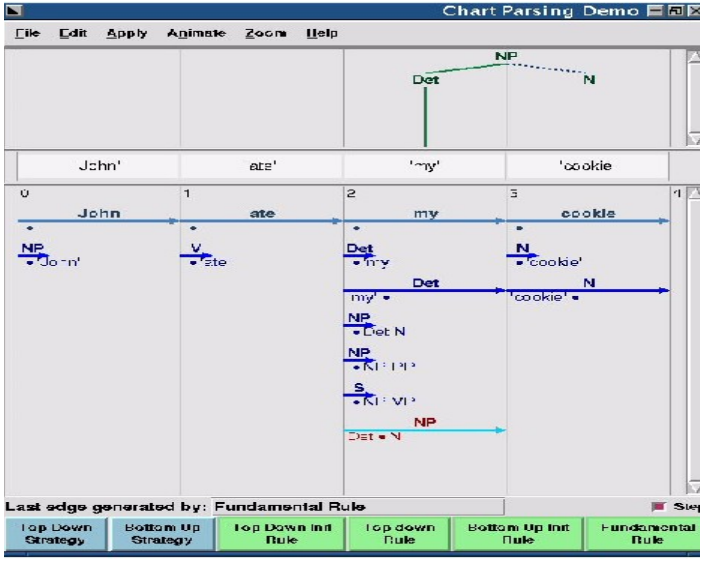
Concordancias

ocabulario terminal de la	***gramática***	o lo que es lo mis
tivos , o a través de una	***gramática***	.</s>
ocabulario terminal de la	***gramática***	).</s>
algunos restringiendo las	***gramáticas***	utilizadas y otros
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

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

Chunking y Parsing



Gracias por su atención

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