

Introducción al Procesamiento de Lenguaje Natural

Grupo PLN – InCo

Sintaxis

Sintaxis

Del gr. σύνταξις – “syntaxis” – “organización” compuesto por el prefijo συν – “syn” : “con” y el verbo τάσσω – “tásso” : “poner en orden” u “ordenar”

- La sintaxis refiere al modo en el que las palabras se disponen juntas en la oración
 - Tiene que ver con la **estructura** que tienen las oraciones **NO** con su significado
 - Se refiere en general a las **categorías gramaticales**, especie de “clases de equivalencia” para las palabras y cómo estas se relacionan, no a las palabras en sí
-

Sintaxis – Categorías gramaticales

- nombre/sustantivo (*casa, casas, amigo, lluvia, felicidad*)
 - determinante (incluye artículo) (*el, una, unos, esos, nuestro, algún*)
 - adjetivo (*alto, alta, lindos, inteligente, soberbio*)
 - verbo (*cantar, cantando, cantaremos, sabe, sé, tomar*)
 - adverbio (*medio, simplemente, solo, ayer*)
 - pronombre (*yo, él, mí, se, aquello, donde, que*)
 - preposición (*a, de, por, contra, ante, bajo, en*)
 - conjunción (*y, o, si, que, pero*)
-

Sintaxis

- Dos aspectos fundamentales: **constituyentes**
orden
 - Constituyentes: palabra o grupos de palabras que funcionan como una *unidad* dentro de la estructura **oración**
 - El conjunto de todos los constituyentes de una oración es un *conjunto parcialmente ordenado*
-

Sintaxis – Constituyentes

¿Qué tienen en común :

un estudiante de letras de Montevideo

la mayoría de los partidos del Mundial

ella

Barack Obama

el vecino que trabaja en la UTE

?

Sintaxis – Constituyentes

¿Qué tienen en común :

un estudiante de letras de Montevideo
la mayoría de los partidos del Mundial
ella
Barack Obama
el vecino que trabaja en la UTE

?

- **Son Grupos Nominales**
-

Sintaxis – Constituyentes

¿Qué tienen en común :

un estudiante de letras de Montevideo
la mayoría de los partidos del Mundial
ella
Barack Obama
el vecino que trabaja en la UTE

?

- **Son Grupos Nominales**
 - **Funcionan en conjunto como una unidad**
-

Sintaxis – Constituyentes

- preposición – GRUPO PREPOSICIONAL
 - nombre/sustantivo – GRUPO NOMINAL
 - verbo – GRUPO VERBAL
 - adjetivo – GRUPO ADJETIVAL
 - adverbio – GRUPO ADVERBIAL
-

Gramática

La gramática comprende dos partes:

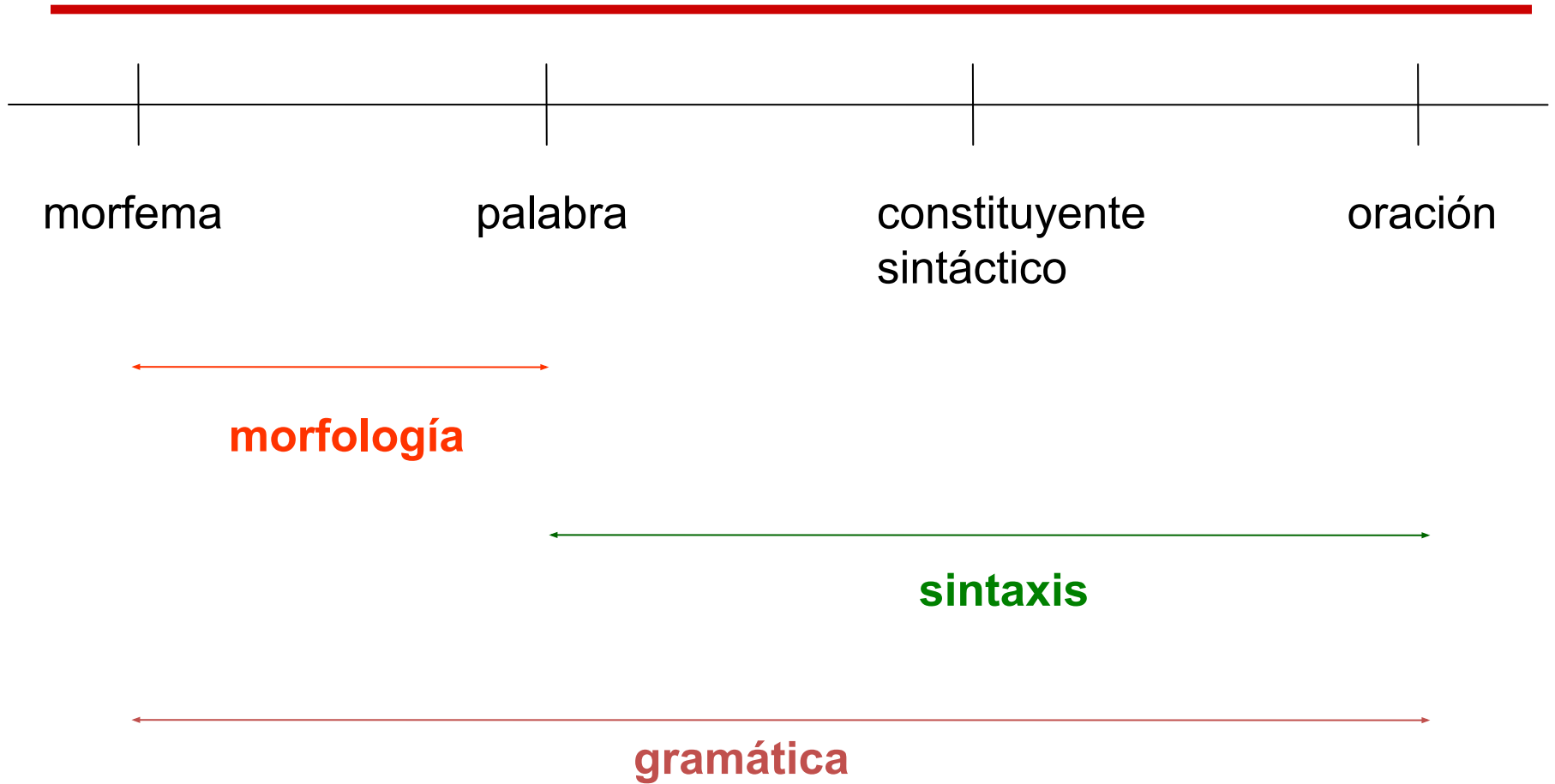
- MORFOLOGÍA

Estudia la estructura interna de las palabras.

- SINTAXIS

Estudia la estructura de la oración, es decir, la combinatoria de las palabras.

Gramática



Gramática Libre de Contexto

- Capturan la noción de constituyente sintáctico y la noción de orden.
 - Herramienta formal que puede ser vista tanto desde un punto de vista generador como estructurador.
 - Propiedades computacionales interesantes: se puede reconocer en tiempo polinómico.
-

Gramática Libre de Contexto

Una Gramática Libre de Contexto es una tupla con 4 elementos:

$$G = (V, T, P, S)$$

- $V \rightarrow$ conjunto de símbolos variables
 - $T \rightarrow$ conjunto de símbolos terminales
 - $S \in V$, símbolo inicial
 - $P \rightarrow$ conjunto de reglas de producción :
 $A \rightarrow \alpha$, con α secuencia de símbolos de $V \cup T$,
eventualmente vacía ($\alpha = \epsilon$)
-

Gramática Libre de Contexto

Una Gramática Libre de Contexto es una tupla con 4 elementos:

$$G = (V, T, P, S)$$

- $V \rightarrow$ conjunto de símbolos variables
- $T \rightarrow$ conjunto de símbolos terminales
- $S \in V$, símbolo inicial
- $P \rightarrow$ conjunto de reglas de producción :
 $A \rightarrow \alpha$, con α secuencia de símbolos de $V \cup T$,
eventualmente vacía ($\alpha = \epsilon$)

Ejemplo: $S \rightarrow aSb \mid \epsilon$;

$$G = (\{S\}, \{a, b\}, S, \{S \rightarrow aSb \mid \epsilon\})$$

Gramática Libre de Contexto

- Una GLC es un “dispositivo” **generador**
- El concepto de **derivación** permite definir *lenguajes*
- Una tira de variables y terminales **deriva** en otra si puede ser reescrita en la segunda como resultado de aplicaciones de reglas de la gramática
- Definimos el lenguaje $L = L(G)$ generado por una gramática G del siguiente modo:
$$L = \{ w / S \Rightarrow^* w \}, \text{ siendo } \Rightarrow^* \text{ una “especie” de clausura transitiva de } \Rightarrow$$

y w una tira de terminales
- Para el ejemplo se podría obtener:

$$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aabb$$

GLC en Computación

Teoría de Lenguajes Formales

- $L_1 = \{ w / w = a^n b^n c^n, n > 0 \}$
 - $L_2 = \{ w / w = a^n b^n, n > 0 \}$
 - $L_3 = \{ w / w = (ab)^n, n > 0 \}$
-

GLC en Computación

Jerarquía de Chomsky

Teoría de Lenguajes Formales

- $L_1 = \{ w / w = a^n b^n c^n, n > 0 \}$ **Sensibles al Contexto (tipo 1)**
 - $L_2 = \{ w / w = a^n b^n, n > 0 \}$ **Libres de Contexto (tipo 2)**
 - $L_3 = \{ w / w = (ab)^n, n > 0 \}$ **Regulares (tipo 3)**
 - *la jerarquía se completa con los lenguajes tipo 0 (Recursivamente Enumerables)*
-

GLC en Computación

Lenguajes de programación

Backus–Naur Form

- $\langle \text{sentence} \rangle ::= \langle \text{if} \rangle \mid \langle \text{while} \rangle \mid \langle \text{cond} \rangle \mid \langle \text{sequence} \rangle$
 - $\langle \text{while} \rangle ::= \text{while } \langle \text{cond} \rangle \text{ do } \langle \text{sequence} \rangle$
-

GLC en Computación

Lenguajes de programación

Backus–Naur Form

- $\langle \text{sentence} \rangle ::= \langle \text{if} \rangle \mid \langle \text{while} \rangle \mid \langle \text{cond} \rangle \mid \langle \text{sequence} \rangle$
SENT $\rightarrow$ IF
SENT $\rightarrow$ WHILE
....
- $\langle \text{while} \rangle ::= \text{while } \langle \text{cond} \rangle \text{ do } \langle \text{sequence} \rangle$
WHILE $\rightarrow$ while COND do SEQ

es una especificación (metalenguaje) con el formalismo de las GLC para especificar sintaxis

GLC para GN sencillos

el avión

un perro andaluz

París

GLC para GN sencillos

GN $\rightarrow$ Det Nom

GN $\rightarrow$ Det Nom Adj

GN $\rightarrow$ NomProp

GRAMÁTICA

Det $\rightarrow$ el | un | la | ...

Nom $\rightarrow$ avión | perro | ...

Adj $\rightarrow$ andaluz | viejo | ...

NomProp $\rightarrow$ París | Jorge | ...

LÉXICO

GLC para GN sencillos

GN $\rightarrow$ Det Nom

GN $\rightarrow$ Det Nom Adj

GN $\rightarrow$ NomProp

GRAMÁTICA

**símbolos terminales y no
terminales**

Det $\rightarrow$ el | un | la | ...

Nom $\rightarrow$ avión | perro | ...

Adj $\rightarrow$ andaluz | viejo | ...

NomProp $\rightarrow$ París | Jorge | ...

LÉXICO

es parte de la gramática

Ejercicio

Cuando salí de Monte con un currículum abusivamente sobresaliente y bajo el brazo un recién nacido título de ingeniero, el profesor Paley estaba a mi lado y no me abandonó hasta que pisamos Santamaría.

Identificar los grupos nominales (GN) contenidos en el texto anterior y escribir una GLC que los genere.

Ejercicio

Monte

*un currículum abusivamente sobresaliente
el brazo*

*un recién nacido título de ingeniero
el profesor Paley*

mi lado

Santamaría

Ejercicio: una (posible) solución

GN → Nom
GN → NomProp
GN → Det Nom
GN → Det GAdj Nom
GN → Det Nom GAdj
GN → GN GP
GN → GN NomProp

GAdj → Adj
GAdj → Adv Adj

GP → Prep GN

GRAMÁTICA

Det → el | un | mi

Nom → currículum | brazo | título |
ingeniero | profesor | lado

Adj → sobresaliente | nacido

NomProp → Monte | Paley | Santamaría

Adv → abusivamente | recién

Prep → de

LÉXICO

La oración

Diferentes enfoques:

- *criterio lexicográfico*: comienza en mayúscula y termina en punto
 - *criterio gramatical*: verbo conjugado + independencia sintáctica
 - *criterio semántico*: expresa algo que puede ser verdadero o falso y que comunica una idea completa
-

La oración

Es una unidad de predicación en donde se establece una *relación* entre dos constituyentes:

- el sujeto (suele ser un grupo nominal)
- el predicado (suele ser un grupo verbal)

El predicado está *nucleado* por un verbo.

La oración

Varias formas/tipos posibles:

- Aseverativa: *Julio vende la casa.*
 - Imperativa: *Traeme ya el informe.*
 - Interrogativa: *¿Podés pasarme la sal?*
 - Exclamativa: *¡Cuánto ha crecido!*
 - Desiderativa: *Ojalá salga el sol.*
 - Dubitativa: *Quizá venga María.*
-

Oración aseverativa

Los niños juegan.

Julio vende la casa.

Compré un libro de sintaxis.

Compré un libro en la librería.

El resultado de la interpelación permanece incierto.

El hombre llegó muy tranquilamente a su casa.

Oración aseverativa

[Los niños]_{GN} [juegan]_{GV}.

[Julio]_{GN} [vende la casa]_{GV}.

[Compré un libro de sintaxis]_{GV}.

[Compré un libro en la librería]_{GV}.

[El resultado de la interpelación]_{GN} [permanece incierto]_{GV}.

[El hombre]_{GN} [llegó muy tranquilamente a su casa]_{GV}.

Oración aseverativa

[Los niños]_{GN} [juegan]_{GV}.

[Julio]_{GN} [vende [la casa]_{GN}]_{GV}.

[Compré [un libro de sintaxis]_{GN}]_{GV}.

[Compré [un libro]_{GN} [en la librería]_{GP}]_{GV}.

[El resultado de la interpelación]_{GN} [permanece [incierto]_{GAdj}]_{GV}.

[El hombre]_{GN} [llegó [muy [tranquilamente]_{GAdv}]_{GAdv}]_{GV} [a su casa]_{GP}]_{GV}.

Reglas para oraciones aseverativas

(esbozo)

$O \rightarrow GN\ GV$

$O \rightarrow GV$

$GV \rightarrow V$

$GV \rightarrow V\ GN$

$GV \rightarrow V\ GAdj$

$GV \rightarrow V\ GAdv$

$GV \rightarrow GV\ GP$

...

Notaciones para la estructura

Para representar el análisis que resulta de la aplicación de las reglas utilizamos:

- Notación con paréntesis rectos (bracketed notation)
 - Árboles de análisis sintáctico
-

Notaciones para la estructura

Ejemplo:

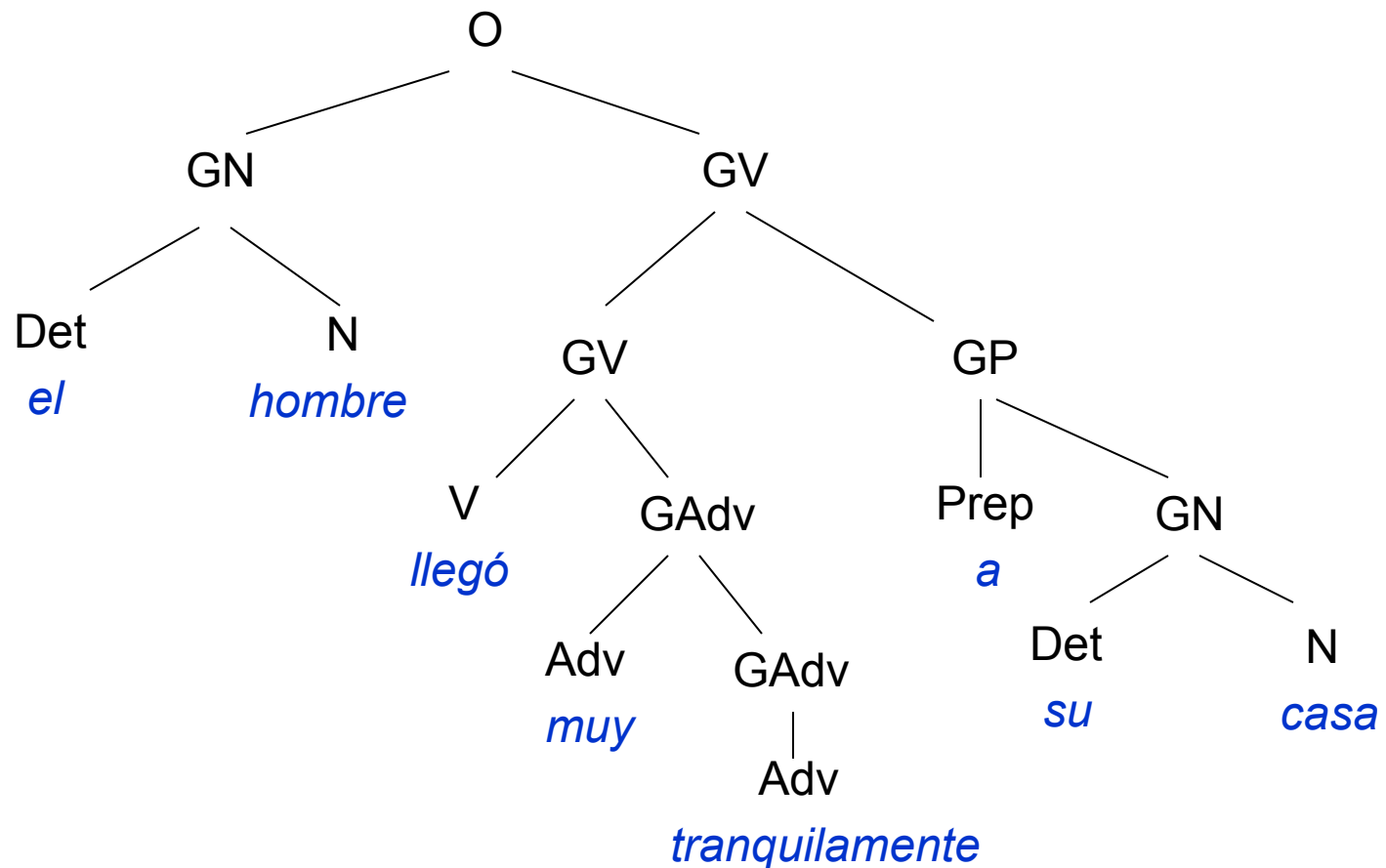
El hombre llegó muy tranquilamente a su casa.

Notaciones para la estructura

[El hombre]_{GN} [[llegó [muy [tranquilamente]_{GAdv}]_{GAdv}]_{GV} [a [su casa]_{GN}]_{GP}]_{GV}.

Notaciones para la estructura

El hombre llegó muy tranquilamente a su casa.



Coordinación

Cleo y Manu leen a Herrera y Reissig.

Mandame un capuccino y una medialuna.

Valentina come y llora al mismo tiempo.

El hombre encendió el auto y arrancó sin mirar si venía alguien.

Coordinación

Cleo y Manu leen a Herrera y Reissig.

GN GN

GN GN

Mandame un capuccino y una medialuna.

Valentina come y llora al mismo tiempo.

El hombre encendió el auto y arrancó sin mirar si venía alguien.

Coordinación

Cleo y Manu leen a Herrera y Reissig.

GN GN

GN GN

Mandame un capuccino y una medialuna.

Valentina come y llora al mismo tiempo.

El hombre encendió el auto y arrancó sin mirar si venía alguien.

Coordinación

Cleo y Manu leen a Herrera y Reissig.

GN GN

GN GN

Mandame un capuccino y una medialuna.

GN

GN

Valentina come y llora al mismo tiempo.

V

V

El hombre encendió el auto y arrancó sin mirar si venía alguien.

O

O

Reglas para coordinación

$O \rightarrow O \text{ ConjCoord } O$

$GN \rightarrow GN \text{ ConjCoord } GN$

$GV \rightarrow GV \text{ ConjCoord } GV$

$\text{ConjCoord} \rightarrow y \mid o \mid \text{pero} \mid \dots$

Características del enfoque GLC

- Reconocimiento eficiente
 - Formalismo general, lingüísticamente neutro pero con cierta dificultad para expresar adecuadamente algunas propiedades.
-

Algunos problemas

GN → Det Nom

Det → el | la | los | las

Nom → libro | libros | casa | casas

¿Qué GNs se pueden generar?

Algunos problemas

Sobregeneración

Una solución posible: “partir” los símbolos

Det --- DetFS, DetMS, DetFP, DetMP

Nom --- NomFS,

GN $\rightarrow$ DetFD NomFS | ...

Algunos problemas

Juan quiere salir.

Juan comió pizza.

Jun quiere pizza.

** Juan comió salir.*

Subcategorización:

un verbo determina la sintaxis posible de sus complementos

Algunos problemas

- Concordancia – multiplicidad de reglas
 - GNMascSing → DetMascSing NomMascSing
(x 4 + cosas como “el agua”)
 - Subcategorización
 - sintáctica

*Juan quiere salir. // * Juan comió + Vinf*

No todas las combinaciones de categorías gramaticales son posibles, las reglas deben preverlo.
 - semántica

*Juan dijo la verdad. // * Juan dijo la mesa.*
*Presenció el incendio. // * Presenció la silla.*
-

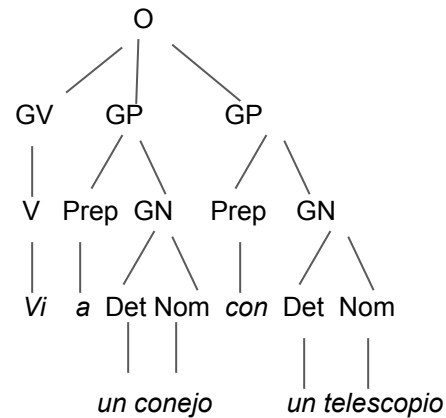
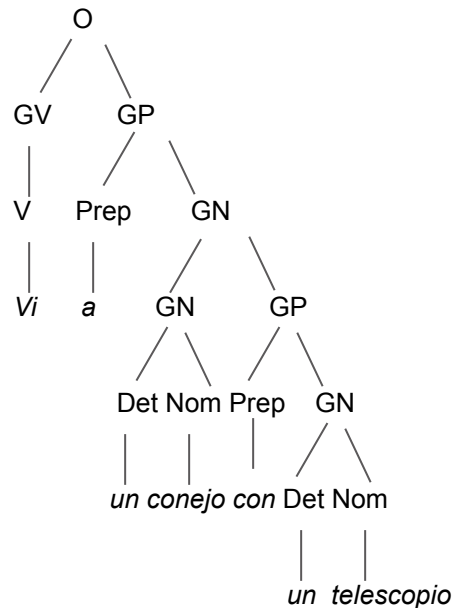
Algunos problemas

- Dificultad en tener una gramática de cobertura completa
 - Múltiples análisis – ambigüedad
 - *Vi a un conejo con un telescopio* (PP attachment)
-

Algunos problemas

Ejemplo:

Vi a un conejo con un telescopio



Algunos problemas

Para sistemas de parsing (fines de los 80's)

- Los parsers asociados a GLC tienen una cobertura pobre
 - Un porcentaje importante de oraciones quedan sin análisis
 - Incluso oraciones simples tienen muchos análisis
 - Los parsers no tienen método de selección entre los distintos análisis
 - Una propuesta de solución a este problema:
 - GLC extendidas con un mecanismo de puntaje
 - Parsing que selecciona el análisis mejor “puntuado”
 - Ha dado resultado (con diversas modificaciones y extensiones)
-

Ranqueo de árboles

Se propone un enfoque tal que:

1– No se escribe una *gramática*, sino que se escriben *árboles sintácticos* para un conjunto amplio de oraciones (un corpus anotado).

- Con suficientes casos, podremos aspirar a una buena cobertura.
- Los ejemplos anotados definen de modo implícito la gramática.

2– Para el análisis de una oración nueva, se utiliza la similitud con los árboles del corpus como medida de “cuan bueno es” (*gold standard*)

- Utilizaremos la probabilidad del árbol como medida.
 - El paradigma es GLCP: Gramáticas Libres de Contexto Probabilistas.
-

Gramáticas Libres de Contexto Probabilistas (GLCP)

$G = (V, T, S, R, P)$

- V conjunto de no terminales (variables)
- T conjunto de terminales
- S símbolo inicial
- R conjunto de reglas de la forma $X \rightarrow \gamma$,
 - >> X es un no terminal
 - >> γ es una secuencia de terminales y/o variables
- P conjunto de probabilidades de cada regla r

Se cumple:

$$\forall X \in N, \sum_{X \rightarrow \gamma \in R} P(X \rightarrow \gamma) = 1$$

Gramáticas Libres de Contexto Probabilistas (GLCP)

La gramática G genera un *modelo de lenguaje* L

(conjunto de tiras con una distribución de probabilidades sobre ellas)

Probabilidad de un árbol

$P(t)$ – La probabilidad de un árbol es el producto de las probabilidades de las reglas usadas para generarlo.

- Se supone independencia en la aplicación de las distintas reglas

Probabilidad de una tira

$P(w_{1n})$ – La probabilidad de una tira es la suma de las probabilidades de los árboles que tienen esa tira como alcance.

- La probabilidad de una tira interesa para modelos de lenguaje (transcripción de habla, resumen automático, traducción automática, etc.)

Ejemplo (adapt. Manning)

S	→	NP VP	1.0	NP	→	NP PP	0.4
VP	→	V NP	0.7	NP	→	<i>Juan</i>	0.1
VP	→	VP PP	0.3	NP	→	<i>María</i>	0.18
PP	→	P NP	1.0	NP	→	<i>arroz</i>	0.04
P	→	<i>con</i>	1.0	NP	→	<i>cuchara</i>	0.18
V	→	<i>comió</i>	1.0	NP	→	<i>queso</i>	0.1

Juan comió arroz con cuchara

Probabilidades de árboles y tiras

w_{15} = *Juan comió arroz con cuchara*

- $P(t_1) = 1.0 * 0.1 * 0.7 * 1.0 * 0.4 * 0.04 * 1.0 * 1.0 * 0.18$
 $= 0.0002016$
 - $P(t_2) = 1.0 * 0.1 * 0.3 * 0.7 * 1.0 * 0.04 * 1.0 * 1.0 * 0.18$
 $= 0.0001512$
 - $P(w_{15}) = P(t_1) + P(t_2)$
 $= 0.0002016 + 0.0001512 = \mathbf{0.0003528}$
-

Estimación de parámetros (probabilidades de las reglas)

Usar un corpus de árboles:

La probabilidad de una regla se aproxima por la frecuencia relativa de su utilización en el corpus

$$P(A \rightarrow \alpha) = C(A \rightarrow \alpha) / \sum_{\beta} C(A \rightarrow \beta) = C(A \rightarrow \alpha) / C(A)$$

$C(A \rightarrow \alpha)$: cantidad de veces que se usa la regla $A \rightarrow \alpha$ en el corpus de árboles

$C(A)$: cantidad de veces que aparece el símbolo A en el corpus

Estimación de parámetros (probabilidades de las reglas)

Usar un corpus de árboles (cont.):

- La probabilidad de una regla se aproxima por la frecuencia relativa de su utilización en el corpus
 - La idea es entrenar gramáticas *treebanks*
 - Los *treebanks* son corpus en los que cada oración está asociada a un árbol sintáctico
 - Se pueden crear:
 - Directamente “a mano”
 - Con parsing automático y posterior corrección manual
-

Penn Treebank

- El *Penn Treebank* es un corpus anotado ampliamente usado (inglés), mantenido por el LDC (*Linguistic Data Consortium*)
 - *Es una base de datos de árboles lingüísticos*
 - Cada oración ha sido anotada con su estructura sintáctica (árbol) y algo de información semántica
 - Se trabajó fundamentalmente con el Wall Street Journal: 1 millón de palabras (unas 40.000 oraciones de ediciones del Wall Street Journal del período 1987–1989)
-

PENN TREEBANK

(NP-SBJ (DT The) (NN move))
(VP (VBD followed)
(NP
(NP (DT a) (NN round))
(PP (IN of)
(NP
(NP (JJ similar) (NNS increases))
(PP (IN by)
(NP (JJ other) (NNS lenders)))
(PP (IN against)
(NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
(, ,)
(S-ADV
(NP-SBJ (-NONE- *))
(VP (VBG reflecting)
(NP
(NP (DT a) (VBG continuing) (NN decline))
(PP-LOC (IN in)
(NP (DT that) (NN market))))))
(. .)))

The move followed a round of similar increases by other lenders against Arizona real estate loans, reflecting a continuing decline in that market.

Ejemplo

Juan ama a María

(S (NP (NNP *Juan*))
 (VP (VBZ *ama*)
 (PP (TO *a*)
 (NP (NNP *María*))))

Categorías gramaticales (*tagset*)

- Más de 40 categorías distintas (corpus Brown)
- Incluyen variantes en número para sustantivos
 - NN – sustantivo singular *cat*
 - NNS – sustantivo plural *cats*

En inglés no hay variación en género (solo pronombres) y ni los adjetivos ni los determinantes varían en número

- Incluyen variantes en forma, persona y tiempo para verbos
 - VB forma base *eat*
 - VBD pasado *ate*
 - VG gerundio *eating*
 - VBN participio *eaten*
 - VBP presente, no 3era persona *eat*
 - VBZ presente, 3era persona *eats*

con 6 tags alcanzan

Penn Treebank – Tagset

Tag	Description	Example	Tag	Description	Example
CC	coordin. conjunction	<i>and, but, or</i>	SYM	symbol	<i>+, %, &</i>
CD	cardinal number	<i>one, two, three</i>	TO	“to”	<i>to</i>
DT	determiner	<i>a, the</i>	UH	interjection	<i>ah, oops</i>
EX	existential ‘there’	<i>there</i>	VB	verb, base form	<i>eat</i>
FW	foreign word	<i>mea culpa</i>	VBD	verb, past tense	<i>ate</i>
IN	preposition/sub-conj	<i>of, in, by</i>	VBG	verb, gerund	<i>eating</i>
JJ	adjective	<i>yellow</i>	VCN	verb, past participle	<i>eaten</i>
JJR	adj., comparative	<i>bigger</i>	VBP	verb, non-3sg pres	<i>eat</i>
JJS	adj., superlative	<i>wildest</i>	VBZ	verb, 3sg pres	<i>eats</i>
LS	list item marker	<i>1, 2, One</i>	WDT	wh-determiner	<i>which, that</i>
MD	modal	<i>can, should</i>	WP	wh-pronoun	<i>what, who</i>
NN	noun, sing. or mass	<i>llama</i>	WP\$	possessive wh-	<i>whose</i>
NNS	noun, plural	<i>llamas</i>	WRB	wh-adverb	<i>how, where</i>
NNP	proper noun, singular	<i>IBM</i>	\$	dollar sign	<i>\$</i>
NNPS	proper noun, plural	<i>Carolinas</i>	#	pound sign	<i>#</i>
PDT	predeterminer	<i>all, both</i>	“	left quote	<i>‘ or “</i>
POS	possessive ending	<i>’s</i>	”	right quote	<i>’ or ”</i>
PRP	personal pronoun	<i>I, you, he</i>	(	left parenthesis	<i>[, (, {, <</i>
PRP\$	possessive pronoun	<i>your, one’s</i>	)	right parenthesis	<i>],), }, ></i>
RB	adverb	<i>quickly, never</i>	,	comma	<i>,</i>
RBR	adverb, comparative	<i>faster</i>	.	sentence-final punc	<i>. ! ?</i>
RBS	adverb, superlative	<i>fastest</i>	:	mid-sentence punc	<i>: ; ... --</i>
RP	particle	<i>up, off</i>			

Gramática a partir de corpus

- Los corpus anotados con árboles definen implícitamente una gramática
 - Se trata de tomar todos los sub-árboles locales en el conjunto de árboles del corpus
 - Si el corpus es suficientemente grande, la cobertura de la gramática será bastante buena
 - Pero.... se depende de los criterios utilizados para anotar con árboles las oraciones del corpus
 - En el caso del *Penn*, se suelen tener estructuras muy chatas y gran cantidad de reglas
-

Gramática a partir de corpus

...

(NP

(NP (DT a) (NN round))

(PP (IN of)

(NP

(NP (JJ similar) (NNS increases))

(PP (IN by)

(NP (JJ other) (NNS lenders)))

(PP (IN against)

(NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))

...

Reglas “inferidas” de los árboles

NP → NP PP // NP → DT NN // NP → JJ NNS

NP → NP PP PP // NP → NNP JJ NN NNS *(estructuras “chatas”)*

PP → IN NP

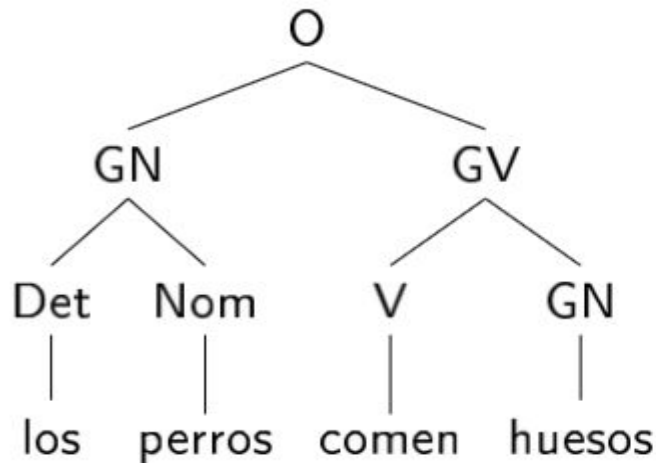
Otros formalismos

- Gramáticas de dependencias
 - HPSG
 - Lexical heads
 - Gramáticas categoriales combinatorias
 - ...
-

Gramáticas de Dependencias

- Enfoque completamente distinto al de las GLC
- No hay constituyentes

Ejemplo: *Los perros comen huesos.*



Gramáticas de Dependencias

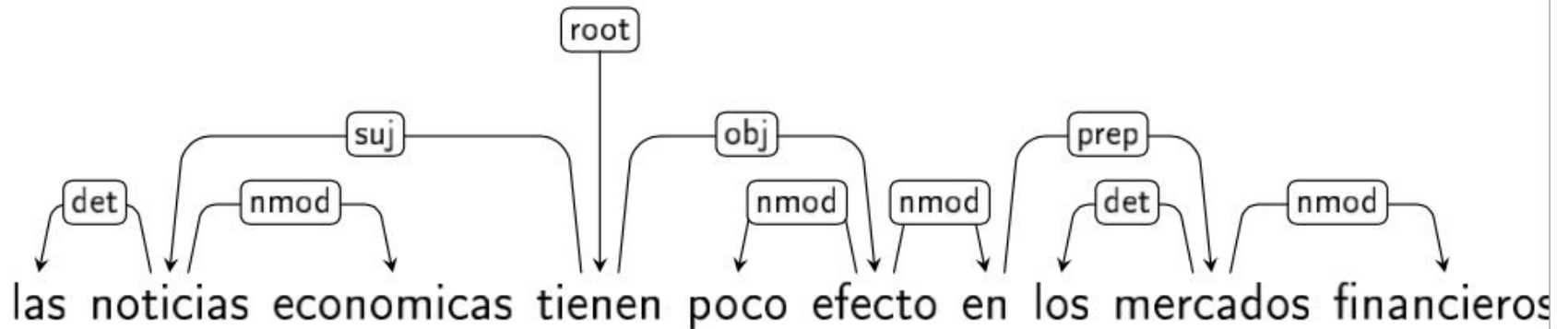
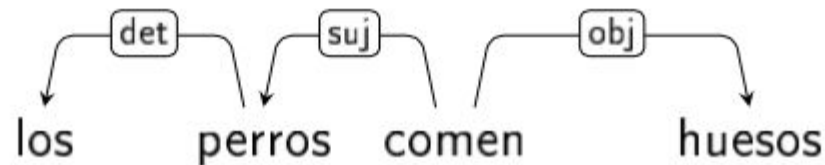
- Relaciones *biléxicas* (palabra a palabra)
- Asimétricas: una de las palabras gobierna a la otra
- Etiquetadas (en general) con la función

En el ejemplo: *Los perros comen huesos*

"comen" gobierna a "perros" y "huesos"
"perros" gobierna a "los"

Gramáticas de Dependencias

Los perros comen huesos.



Gramáticas de Dependencias

- Las etiquetas de función pueden estar orientadas a la sintaxis (*subj, obj, nmod*), o...
- Genéricamente orientadas a la semántica (*agente, tema*), o...
- Una mezcla de ambas
- Se dibuja un:
 1. Grafo conexo, dirigido
 2. Acíclico
 3. Single head: un solo padre por cada nodo y un nodo raíz (sin padre)

Estas restricciones implican que el grafo es un árbol.
De hecho, 1 y 3 implican 2

Gramáticas de Dependencias

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	Description
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	Description
CONJ	Conjunct
CC	Coordinating conjunction

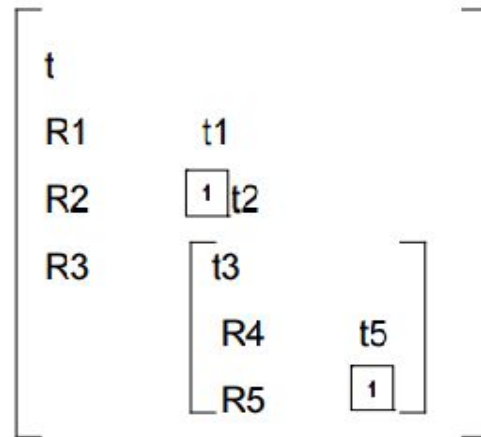
subconjunto de relaciones del Proyecto *Universal Dependencies* (Nivre et al. 2016)

Gramáticas HPSG

- Propuesta original de Pollard y Sag (1987).
 - Gramática fuertemente lexicalista: las posibilidades combinatorias de las palabras están definidas en las entradas léxicas.
 - Se definen pocas reglas.
 - Se toman ideas de la teoría X Barra.
 - Todos los elementos de la gramática se especifican por medio de *tfs*, entre ellas se aplica la operación de Unificación.
 - *tfs* (*typed featured structures* – estructuras de rasgos tipificadas) que representan sus restricciones
-

Gramáticas HPSG

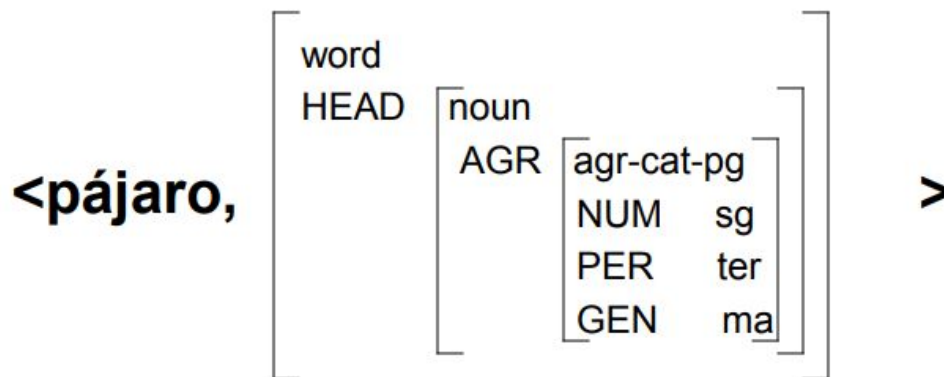
Una *tfs* es una matriz con rasgos y valores asociados a ellos:



Gramáticas HPSG

- Una gramática con *tfs* consiste en un conjunto de principios y reglas *R*, un conjunto de entradas léxicas *L*
- Todos estos elementos son *tfs* bien formadas respecto a una jerarquía de tipos con sus restricciones.

Ejemplo de entrada léxica:



Gramáticas HPSG

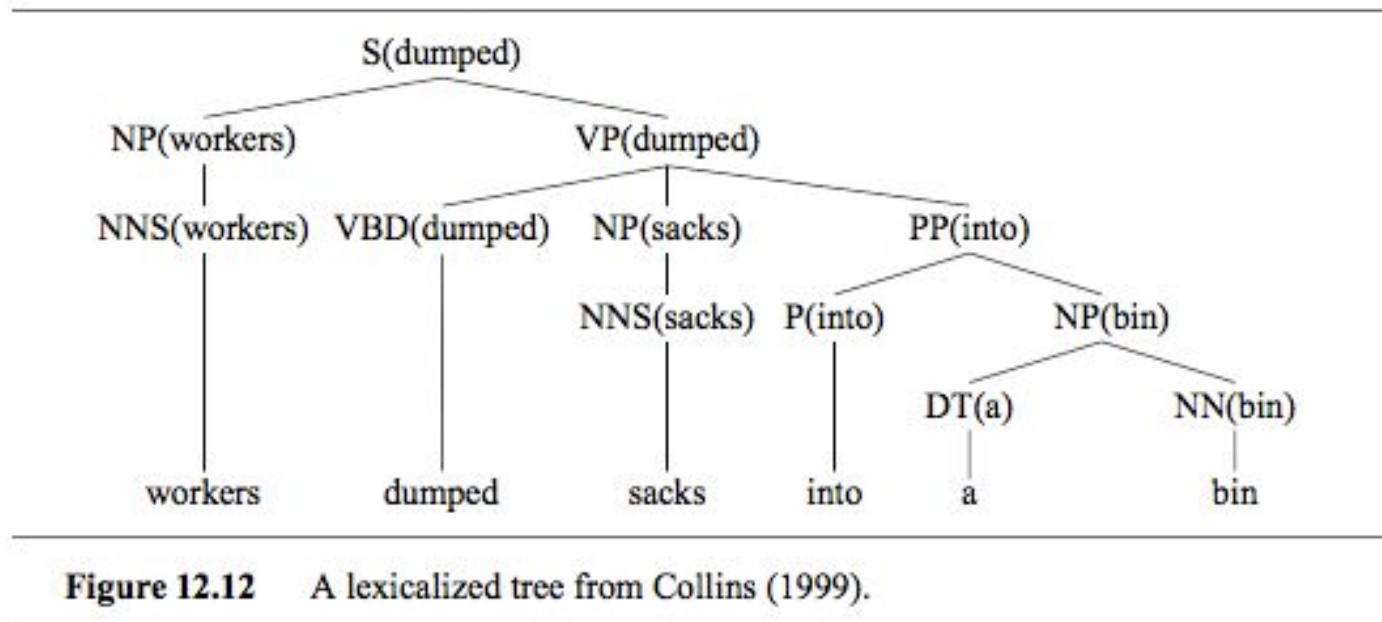
Ejemplo de una regla:

```
np-rule := phrase &  
[ CATEG np,  
  AGR #agr & agr,  
  ARGS < [CATEG det,  
           AGR #agr],  
  [CATEG n,  
   AGR #agr] > ]
```

Gramáticas con *lexical heads*

- A cada constituyente sintáctico se le puede asociar un cabezal (*head*)
 - GN (head=N), GV (head=V)
 - Para la generación del árbol, cada regla de la GLC es “aumentada” para identificar un componente del lado derecho que sea “el principal”; es decir, se le asocia a la derecha el *head* correspondiente
 - Simple para frases sencillas
-

Gramáticas con *lexical heads*



las *heads* se “van pasando” a través de los nodos en el árbol sintáctico

Gramáticas Categoriales

- Formalismo de descripción lingüística cuyos orígenes son de la 1era mitad del siglo XX (Adjuckiewicz, Bar-Hillel)
 - Fundamentos lógicos: la estructura de una oración según la gramática, se obtiene a partir de una derivación en un sistema formal
 - Las CCG (Steedman, 1986) son una variante con operadores adicionales que permiten flexibilidad en las derivaciones
-

Gramáticas Catoriales

Componentes básicos:

- *Categorías*: especifican tipos de palabras o constituyentes
 - *Reglas combinatorias*: especifican cómo se pueden combinar las palabras o constituyentes
 - *Léxico*: especifica las categorías que puede tener una palabra
 - *Derivaciones*: proceso sistemático de combinación de palabras o constituyentes
-

Gramáticas Catoriales

Categorías simples:

Conjunto limitado de categorías atómicas, por ej. O, GN, etc.

Categorías complejas (del tipo Functor Argumento):

Son funciones, retornan un resultado cuando obtienen un argumento

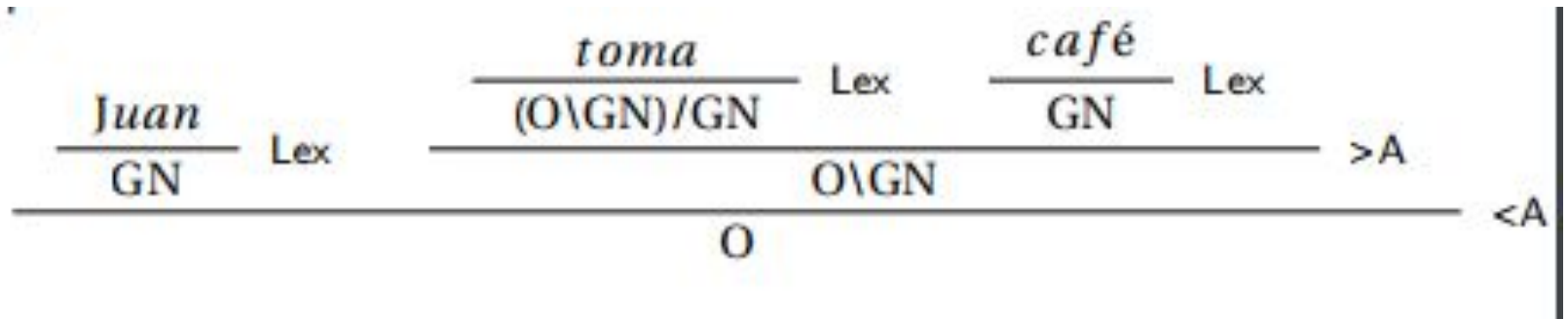
Juan toma café

Juan, café $\rightarrow$ GN

toma $\rightarrow$ (O\GN)/GN

Gramáticas Categoriales

Ejemplo de una derivación:



Parsers para el español

Parsers para el español:

- Freeling: <http://nlp.lsi.upc.edu/freeling/>
- Stanford: <http://nlp.stanford.edu:8080/parser/index.jsp>
- MaltParser:

http://ws04.iula.upf.edu/soaplab2-axis/#syntactic_tagging.malt_parser_row

Referencias

- Jurafsky, D. y J. Martin. 2019. *Speech and Language Processing*. Third edition Capítulo 12
 - Hopcroft & Ullman. 2013. *Introduction to Automata Theory, Languages and Computation*. Third edition Capítulo 5
-