Recursos lingüísticos

Miguel A. Alonso Jorge Graña Jesús Vilares

Departamento de Computación, Facultad de Informática, Universidade da Coruña

Índice

Tipos de principales recursos lingüísticos

- Corpus etiquetados
- Diccionarios (lexicones)
- Bancos de árboles (Treebanks)

Corpus etiquetado

- Un **corpus** es una colección de documentos
- Un **corpus etiquetado** es aquél en el que todas y cada una de las palabras aparecen acompañadas al menos de su **etiqueta correcta**
- También suele aparecer el **lema**
- Sirven como referencia de un estilo de uso de una lengua durante el proceso de ajuste o puesta a punto de las herramientas de etiquetación

Diccionario

- En su forma más simple, una lista de palabras
- Generalmente acompañada de sus posibles etiquetas o papeles candidatos que puede desempeñar dentro de la frase
- También puede contener acepciones semánticas
- También puede contener relaciones entre palabras
  - derivativas (poco común)
  - relaciones semánticas (WordNet, EuroWordNet, . . .)
  - . . .
- También puede contener información de un domicion específico (SentiWordNet, . . .)

Treebank

- Cada frase aparece completamente analizada sintácticamente
- Dos grandes paradigmas (no excluyentes):
  - Treebanks de constituyentes
  - Treebanks de dependencias
- Permiten extraer:
  - las reglas de la gramática a la que obedece un determinado idioma, o al menos un subconjunto relevante de él, incluso con una probabilidad de uso asignada a cada una de esas reglas.
  - Las relaciones de dependencias entre palabras, así como sus proabalidades relativas

Ejemplo: El corpus ITU o CRATER

- Corpus etiquetado **paralelo** inglés-francés-español (contiene los mismos documentos en varios idiomas)
- Dominio de los textos: telecomunicaciones (ITU es la International Telecommunications Union)
- Tamaño:
  - Texto original: 5 millones de palabras
  - Texto anotado automáticamente: 5 millones de palabras
  - Textos anotado **revisado**: medio millón de palabras

```
En/PREP/en esta/DMPXFS/este colaboraci&oacute;n/NCFS/colaboraci&oacute;n debe/VMPI3S/deber
reconocerse/VCLI/reconocer el/ARTDMS/el car&aacute;cter/NCMS/car&aacute;cter consultivo/ADJGMS/consultivo
de/PREP/de las/ARTDFP/el organizaciones/NCFP/organizaci&oacute;n que/CQUE/que participan/VLPI3P/participar
en/PREP/en los/ARTDMP/el trabajos/NCMP/trabajo del/PDEL/del CCITT/NPTOS/CCITT ,/CM/,
en_particular/ADVN/en_particular a/PREP/a la/ARTDFS/el ISO/ACRNM/ISO ,/CM/, desde/PREP/desde el/ARTDMS/el
punto/NCMS/punto de/PREP/de vista/NCFS/vista de/PREP/de su/PPOSPS/suyo labor/NCFS/labor
con_respecto_a/PREP/con_respecto_a los/ARTDMP/el sistemas/NCMP/sistema de/PREP/de datos/NCMP/dato y/CC/y
a/PREP/a las/ARTDFP/el comunicaciones/NCFP/comunicaci&oacute;n ./FS/.
```

Estadísticas

- 486.073 palabras (7.564.781 bytes)
- 14.919 frases (media de 33 palabras por frase)
- cada línea es una frase:
  `palabra/etiqueta/lema  palabra/etiqueta/lema  ...`
- 12.462 lemas, 17.138 formas con 18.917 etiquetas posibles:
  - 15.745 formas con 1 etiqueta
  - 1.097 formas con 2 etiquetas
  - 223 formas con 3 etiquetas
  - 61 formas con 4 etiquetas
  - 8 formas con 5 etiquetas
  - 3 formas con 6 etiquetas
  - 1 forma con 7 etiquetas
- $8,13\%$ de formas ambiguas
- $1,10$ etiquetas por forma

Estadísticas

- Estadísticas sobre palabras:
  - 263.592 palabras con 1 etiqueta
  - 107.746 palabras con 2 etiquetas
  - 79.751 palabras con 3 etiquetas
  - 7.013 palabras con 4 etiquetas
  - 18.949 palabras con 5 etiquetas
  - 8.528 palabras con 6 etiquetas
  - 494 palabras con 7 etiquetas
- 486.073 palabras con 895.760 etiquetas posibles.
- $45,77\%$ de palabras ambiguas
- $1,84$ de etiquetas por palabra

Tagset

ACRNM	Sigla (`ADN`)
ADJCP	Adjetivo comparativo general plural (`mayores`, `menores`)
ADJCS	Adjetivo comparativo general singular (`mayor`, `menor`)
ADJGFP	Adjetivo positivo general femenino plural
ADJGFS	Adjetivo positivo general femenino singular
ADJGMP	Adjetivo positivo general masculino plural
ADJGMS	Adjetivo positivo general masculino singular
ADJSFP	Adjetivo superlativo general femenino plural (`máximas`, `mínimas`)
ADJSFS	Adjetivo superlativo general femenino singular (`máxima`, `mínima`)
ADJSMP	Adjetivo superlativo general masculino plural (`máximos`, `mínimos`)
ADJSMS	Adjetivo superlativo general masculino singular (`máximo`, `mínimo`, `grandísimo`)
ADVGR	Adverbio con grado positivo (`muy`, `demasiado`, `mucho`)
ADVGRC	Adverbio con grado comparativo (`más`, `menos`)
ADVGRS	Adverbio con grado superlativo (`abundantísimamente`)
ADVINT	Adverbio interrogativo (`cómo`)
ADVL	Adverbio locativo direccional (`abajo`)
ADVLD	Adverbio locativo dinámico (`adelante`)
ADVLE	Adverbio locativo estático (`dentro`)
ADVLID	Adverbio locativo interrogativo dinámico (`adónde`)
ADVLIN	Adverbio locativo interrogativo (`dónde`)
ADVLP	Adverbio locativo con deixis próxima (`aquí`)
ADVLR	Adverbio locativo con deixis remota (`allí`)
ADVLRD	Adverbio locativo relativo dinámico (`adonde`)
ADVLRE	Adverbio locativo relativo direccional (`donde`)
ADVMRE	Adverbio relativo modal (`como`)
ADVN	Adverbio general (`salvajemente`, `bien`, `probablemente`)
ADVNEG	Adverbio general negativo (`tampoco`)
ADVT	Adverbio temporal (`ahora`, `ayer`)
ADVTIN	Adverbio temporal interrogativo (`cuándo`)
ADVTNE	Adverbio temporal negativo (`nunca`)
ADVTRE	Adverbio temporal relativo (`cuando`)
ALFP	Letra del alfabeto en plural (`As/Aes`, `bes`)

Ejemplo: El CORGA

- COrpus de Referencia do Galego Actual
- http://corpus.cirp.es/corga/
- Documentos en XML
- Dominio general: 457 periódicos, 103 revistas, 415 libros (novela, ensayo, relato corto y teatro) desde 1975 a la actualidad.
- Tamaño
  - 25 millones de palabras
  - 59.193 palabras etiquetadas y revisadas manualmente(767 noticias)

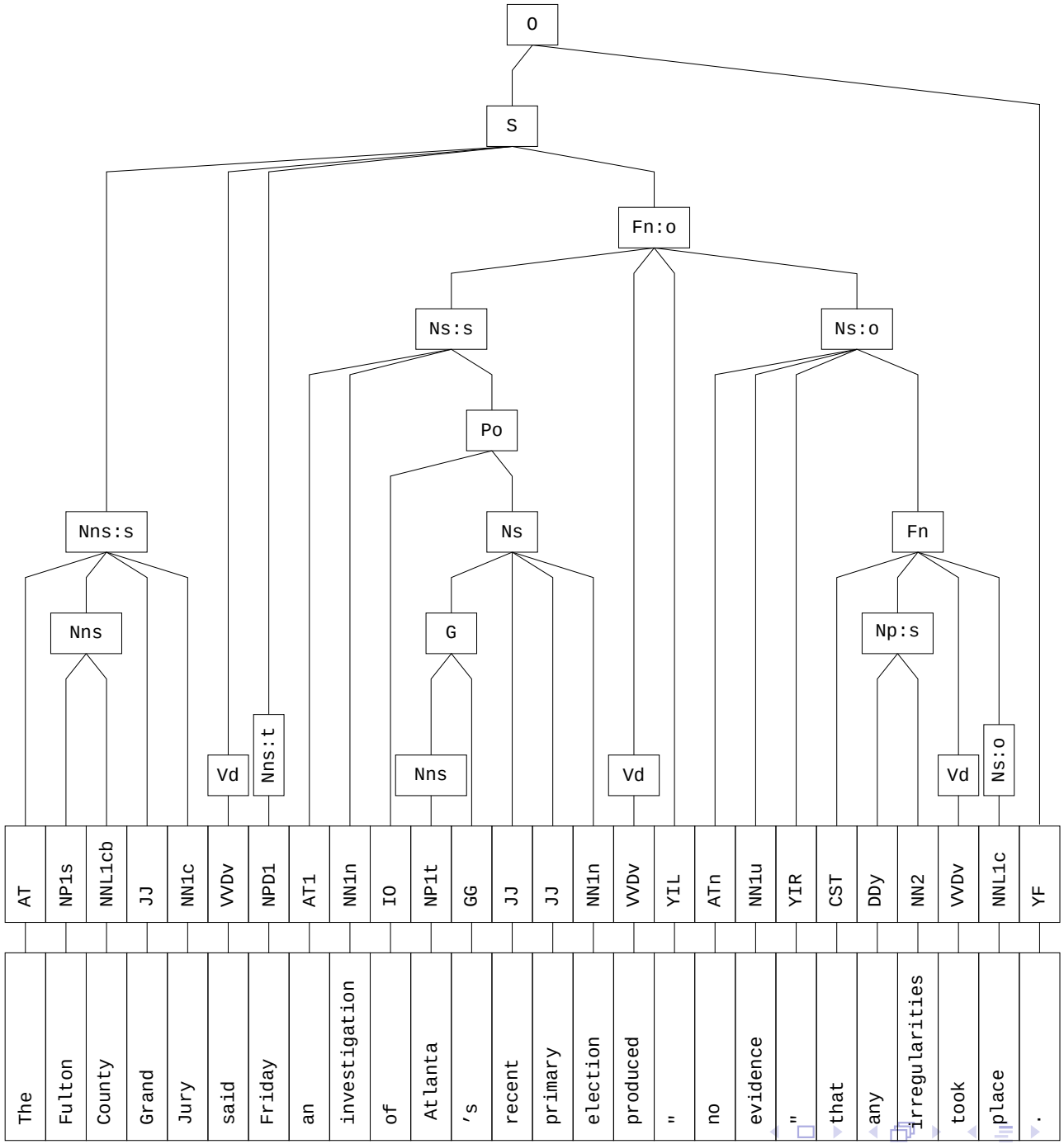
Tagset

			m- masculino **f-** feminino **a-** masculino ou feminino **n-** neutro **0-** non aplica	**s-** singular **p-** plural **a-** singular ou plural **0-** non aplica	**c-** comparativo **s-** superlativo **0-** non aplica	**1-** primeira **2-** segunda **3-** terceira **a-** primeira ou terceira **0-** non aplica	**n-** nominativo **a-** acusativo **d-** dativo **p-** preposicional **f-** formante léxico **o-** outros casos	**p-** presente **i-** copretérito **l-** antepretérito **f-** futuro **e-** pretérito **c-** pospretérito **0-** non aplica	**i-** indicativo **s-** subxuntivo **m-** imperativo **f-** infinitivo **x-** xerundio **p-** participio	**s-** singular **p-** plural **a-** singular ou plural	**d-** determinante **n-** non determinante **a-** determinante ou non determinante
Categoría	**Tipo**	**Subtipo**	**Xénero**	**Número**	**Grao**	**Persoa**	**Caso**	**Tempo Verbal**	**Modo**	**Posuidor**	**Valor**
Substantivo **S** 1	**c-** común **p-** propio 2		m/f/a/0 3	s/p/a/0 4							
Adxectivo **A** 1			m/f/a 3	s/p/a 4	c/s/0 2						
Verbo **V** 1			m/f/0 5	s/p/0 6		1/2/3/a/0 4		p/i/f/e/c/0 2	i/s/m/f/x/p 3		
Preposición **P** 1											
Conxunción **C** 1	**c-** coordinante **s-** subordinante 2										
Adverbio **W** 1	**n-** nuclear **m-** modificador **a-** nuclear ou modificador **r-** relativo **g-** interrogativo/ exclamativo 2										
Artigo **D** 1	**d-** determinado **i-** indeterminado 2		m/f 3	s/p 4							
Demostrativo **E** 1			m/f/n 3	s/p 4							d/n/a 2
Relativo **T** 1			m/f/a 3	s/p/a 4							d/n/a 2
Posesivo **M** 1			m/f/a 5	s/p/a 6		1/2/3 3				s/p/a 4	d/n/a 2
Indefinido **I** 1			m/f/a/0 3	s/p/a/0 4							d/n/a 2
Numeral **N** 1	**c-** cardinal **s-** ordinal 2		m/f/a 4	s/p 5							d/n/a 3
Pronome **R** 1	**t-** tónico **a-** átono 2		m/f/n/a 5	s/p/a 6		1/2/3 4	n/a/d/p/f/o 3				
interrogativo/ exclamativo **G** 1			m/f/a 3	s/p/a 4							d/n/a 2
Locución **L** 1	**a-** adverbial **p-** preposicional **c-** conxuntiva 2	**c-** coordinante **s-** subordinante **0-** non aplica 3									
Interxección **Y**											

Ejemplo: El corpus Susanne

- Treebank de constituyentes del inglés
- Tamaño: 150.000 palabras (10.500lemas diferentes) del corpus BROWN para el inglés americano

```
...
A01:0010b  -  AT     The            the            [O[S[Nns:s.
A01:0010c  -  NP1s   Fulton         Fulton         [Nns.
A01:0010d  -  NNL1cb County         county         .Nns]
A01:0010e  -  JJ     Grand          grand          .
A01:0010f  -  NN1c   Jury           jury           .Nns:s]
A01:0010g  -  VVDv   said           say            [Vd.Vd]
A01:0010h  -  NPD1   Friday         Friday         [Nns:t.Nns:t]
A01:0010i  -  AT1    an             an             [Fn:o[Ns:s.
A01:0010j  -  NN1n   investigation  investigation  .
A01:0020a  -  IO     of             of             [Po.
A01:0020b  -  NP1t   Atlanta        Atlanta        [Ns[G[Nns.Nns]
A01:0020c  -  GG     +<apos>s       -              .G]
A01:0020d  -  JJ     recent         recent         .
A01:0020e  -  JJ     primary        primary        .
A01:0020f  -  NN1n   election       election       .Ns]Po]Ns:s]
A01:0020g  -  VVDv   produced       produce        [Vd.Vd]
A01:0020h  -  YIL    <ldquo>        -              .
A01:0020i  -  ATn    +no            no             [Ns:o.
A01:0020j  -  NN1u   evidence       evidence       .
A01:0020k  -  YIR    +<rdquo>       -              .
A01:0020m  -  CST    that           that           [Fn.
A01:0030a  -  DDy    any            any            [Np:s.
A01:0030b  -  NN2    irregularities irregularity   .Np:s]
A01:0030c  -  VVDv   took           take           [Vd.Vd]
A01:0030d  -  NNL1c  place          place          [Ns:o.Ns:o]Fn]Ns:o]Fn:o]S]
```


O
S
Fn:o
Ns:s
Ns:o
Po
Nns:s
Ns
Fn
Nns
G
Np:s
Vd
Nns:t
Nns
Vd
Vd
Ns:o
AT NP1s NNL1cb JJ NN1c VVDv NPD1 AT1 NN1n IO NP1t GG JJ JJ NN1n VVDv YIL ATn NN1u YIR CST DDy NN2 VVDv NNL1c YF
The Fulton County Grand Jury said Friday an investigation of Atlanta 's recent primary election produced " no evidence " that any irregularities took place .

Tagset

APPGf	her como posesivo ≠ PPHO1f
APPGh1	its
APPGh2	their
APPGi1	my como posesivo
APPGi2	our
APPGm	his excepto como pronombre ≠ PPGm
APPGy	your
AT	the como determinante
AT1	Artículo indefinido a, an
AT1e	every
ATn	no como determinante o calificador ≠ UH
BTO21	in en la secuencia in order + infinitivo
BTO22	order en la secuencia in order + infinitivo
CC	Conjunción coordinada
CC31	Conjunción coordinada de 3 palabras (1a. palabra), e.g. as well as
CC32	Conjunción coordinada de 3 palabras (2a. palabra), e.g. as well as
CC33	Conjunción coordinada de 3 palabras (3a. palabra), e.g. as well as
CCB	but como conjunción coordinada ≠ ICSx RR
CCn	nor
CCr	or
CS	Conjunción subordinada
CS21	Conjunción subordinada de 2 palabras (1a. palabra), e.g. even though
CS22	Conjunción subordinada de 2 palabras (2a. palabra), e.g. even though
CS31	Conjunción subordinada de 3 palabras (1a. palabra), e.g. as long as
CS32	Conjunción subordinada de 3 palabras (2a. palabra), e.g. as long as
CS33	Conjunción subordinada de 3 palabras (3a. palabra), e.g. as long as
CSA	as como conjunción subordinada o como preposición en sentido comparativo ≠ IIa RGa
CSN	than en todos los usos
CST	that como conjunción subordinada ≠ DD1a
CST21	as how
CST22	as how
CSW	whether en todos los usos

Extracción de una gramática a partir del treebank

```
[ Fa [ CSf For CSf ] [ Ds:s [ DD1q each DD1q ] [ Po [ IO of IO ] [ Np [ DD2i these DD2i ] [ NN2 lines NN2 ] Np ] Po ] Ds:s ] [ Vz [ VVZv meets VVZv ] Vz ] [ YTL <bital> YTL ] [ N:o [ FOx Q FOx ] N:o ] [ YTR <eital> YTR ] [ P:p [ II in II ] [ Np [ MC three MC ] [ NN2 points NN2 ] [ YC , YC ] [ REX namely REX ] [ N@ [ MC two MC ] [ NN2 points NN2 ] [ P [ II on II ] [ FOx g FOx ] P ] [ Ns+ [ CC and CC ] [ MC1 one MC1 ] [ NNL1n point NNL1n ] [ P [ II on II ] [ Ms [ MC1 one MC1 ] [ Po [ IO of IO ] [ Np [ AT the AT ] [ JJ multiple JJ ] [ NN2 secants NN2 ] Np ] Po ] Ms ] P ] Ns+ ] N@ ] Np ] P:p ] Fa ]
```

- Añadir un axioma `X`
- eliminar los no terminales que aparecen repetidos antes de cada corchete de cierre

```
[ X [ Fa [ CSf For ] [ Ds_s [ DD1q each ] [ Po [ IO of ] [ Np [ DD2i these ] [ NN2 lines ] ] ] ] [ Vz [ VVZv meets ] ] [ YTL <bital> ] [ N_o [ FOx Q ] ] [ YTR <eital> ] [ P_p [II in ] [ Np [ MC three ] [ NN2 points ] [ YC , ] [ REX namely ] [ N@ [ MC two ] [ NN2 points ] [ P [ II on ] [ FOx g ] ] [ Ns+ [ CC and ] [ MC1 one ] [ NNL1n point ] [ P [ II on ] [ Ms [ MC1 one ] [ Po [ IO of ] [ Np [ AT the ] [ JJ multiple ] [ NN2 secants ] ] ] ] ] ] ] ] ] ]
```

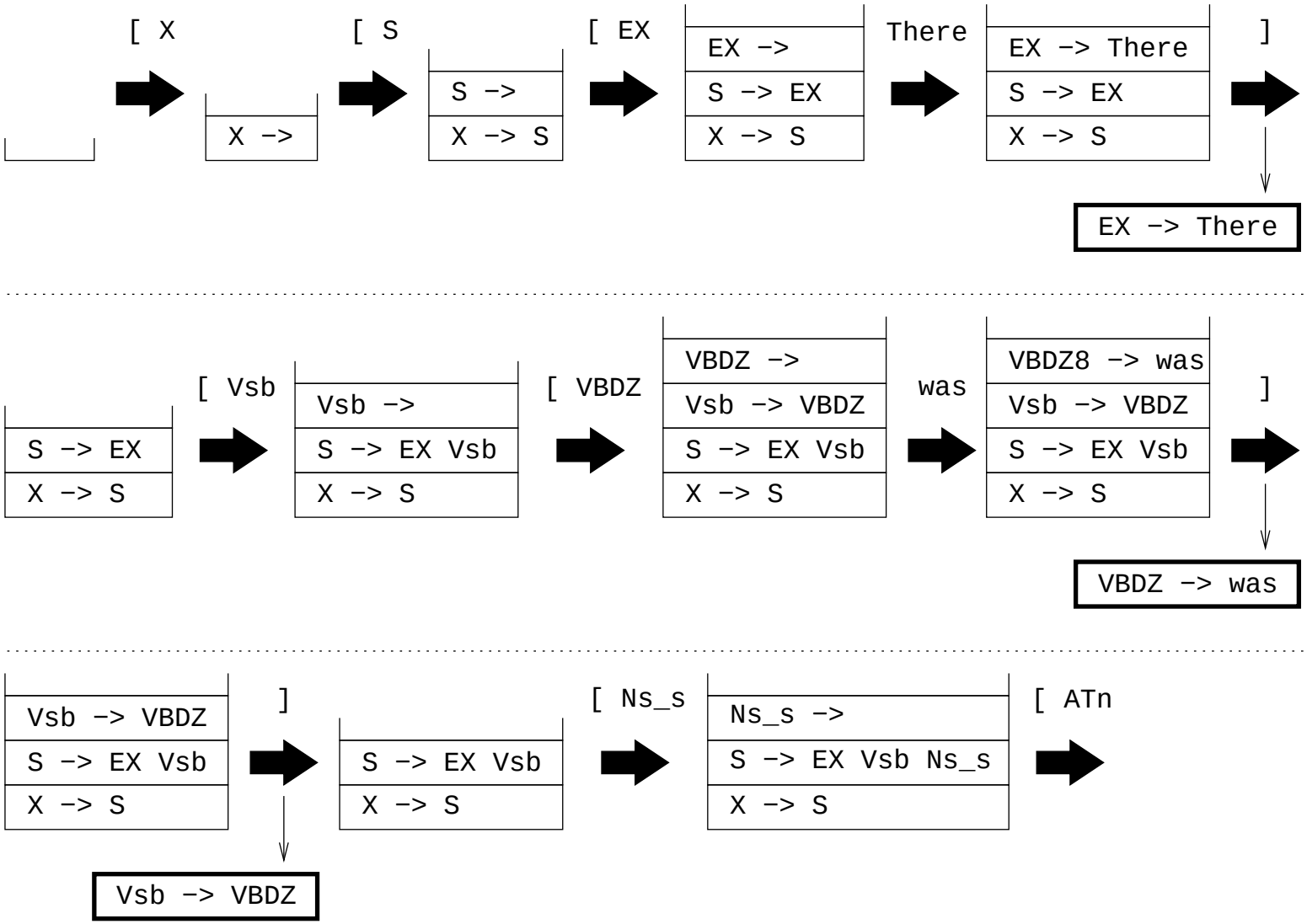
Extracción de reglas gramaticales

- El símbolo [indica el comienzo de una regla: se toma el siguiente símbolo del árbol parentizado, se añade a la parte derecha de la regla que está en la cima de la pila, y se crea un nuevo elemento en la pila para una nueva regla cuya parte izquierda es dicho símbolo
- El símbolo] indica el final de una regla: la regla que está en la cima de la pila en está ya completa, por lo que se extrae y se elimina de la pila
- Cualquier otro símbolo se añade a la parte derecha de la regla que está en la cima de la pila

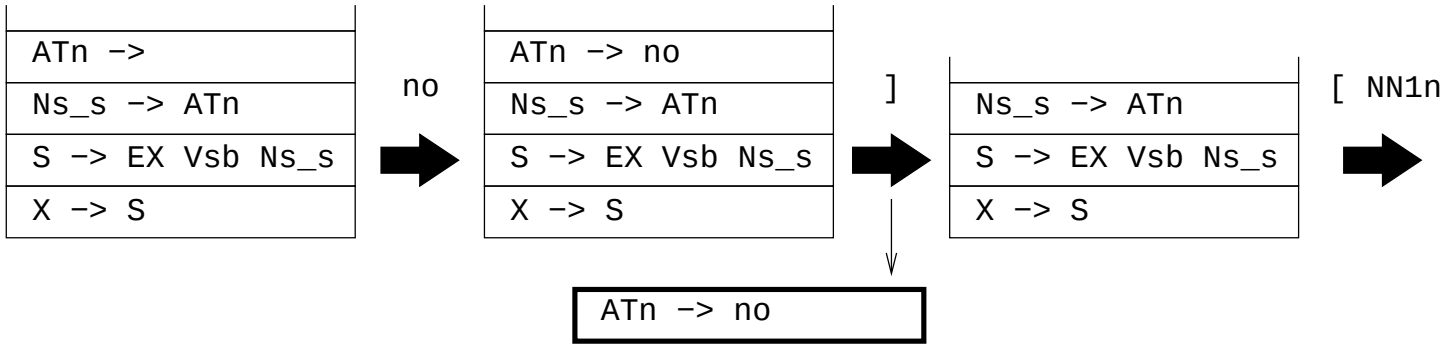
Algoritmo de extracción de reglas gramaticales

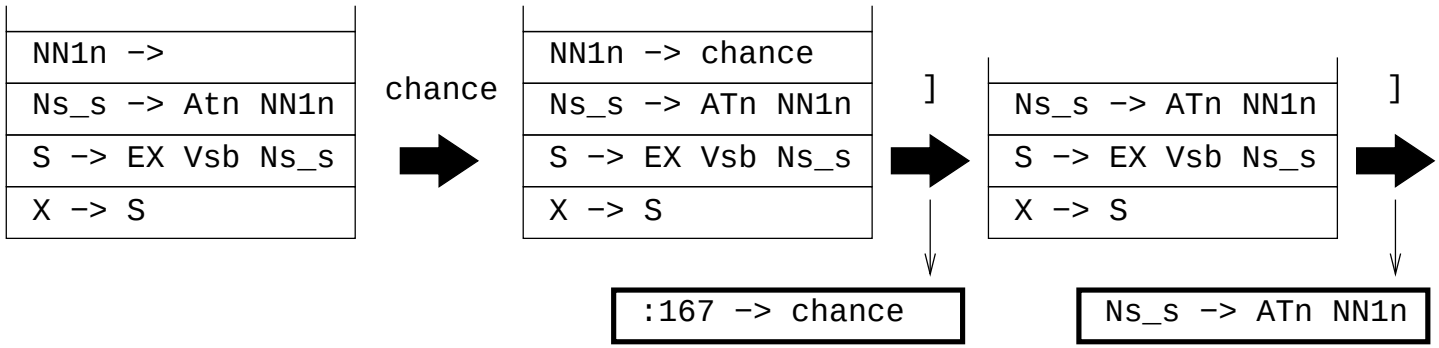
$P \leftarrow$ *Pila_Vacía*;
$R \leftarrow$ *Conjunto_Vacío*;
while (*queden símbolos del Árbol_Parentizado por procesar*) **do begin**
    $A \leftarrow$ *siguiente símbolo del Árbol_Parentizado*;
    case A **of**
        [: $A \leftarrow$ *siguiente símbolo del Árbol_Parentizado*;
            Regla $\leftarrow$ **pop** (P);
            Añadir A en la parte derecha de Regla;
            push $(P, Regla)$;
            push $(P, A \rightarrow)$;
        ] : *Regla* $\leftarrow$ **pop** (P);
            Añadir Regla a R;
        cualquier otro símbolo :
            Regla $\leftarrow$ **pop** (P);
            Añadir A en la parte derecha de Regla;
            push $(P, Regla)$
    endcase
endwhile;
return R

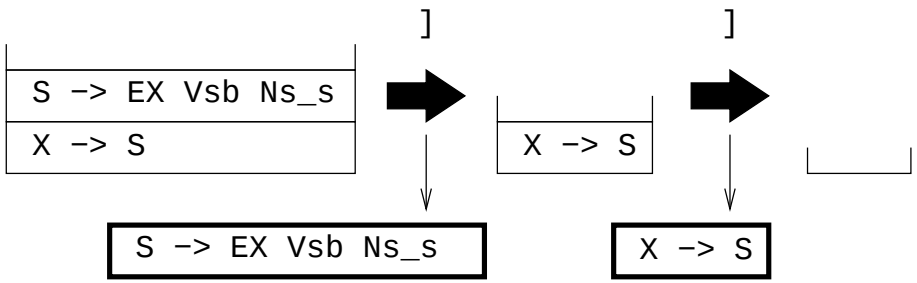
Ejemplo de extracción de reglas gramaticales



Ejemplo de extracción de reglas gramaticales

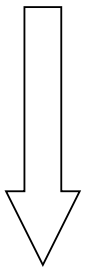






Treebank SUSANNE

Corpus sin trazas	Corpus con trazas
4.292 frases	2.188 frases
77.275 palabras	60.759 palabras
18 palabras por frase	28 palabras por frase
44,93% de las palabras son ambiguas	43,50% de las palabras son ambiguas
2,30 etiquetas por palabra	2,39 etiquetas por palabra



Gramática SUSANNE

17.669 reglas

1.525 no terminales

28.117 nodos