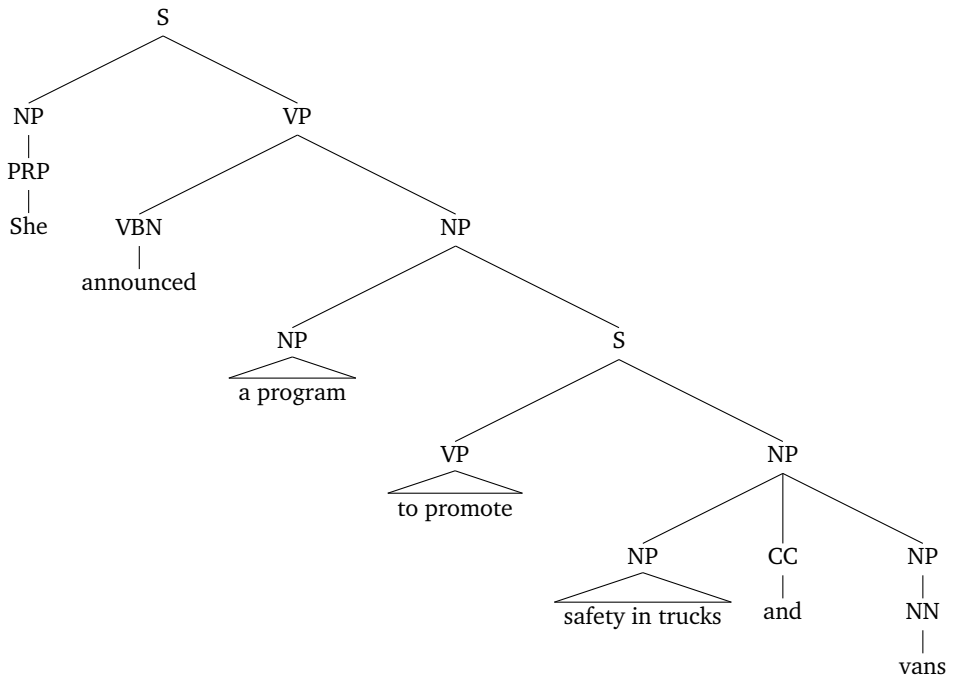
Processament del Llenguatge Humà (GIA-PLH)

Exercicis sobre anàlisi de constituents

CFGs

Ejercicio 1.

Considera la oración "*She announced a program to promote safety in trucks and vans*" y el siguiente análisis sintáctico correspondiente a una de las posibles interpretaciones (el programa promueve la seguridad en camiones y también promueve furgonetas):

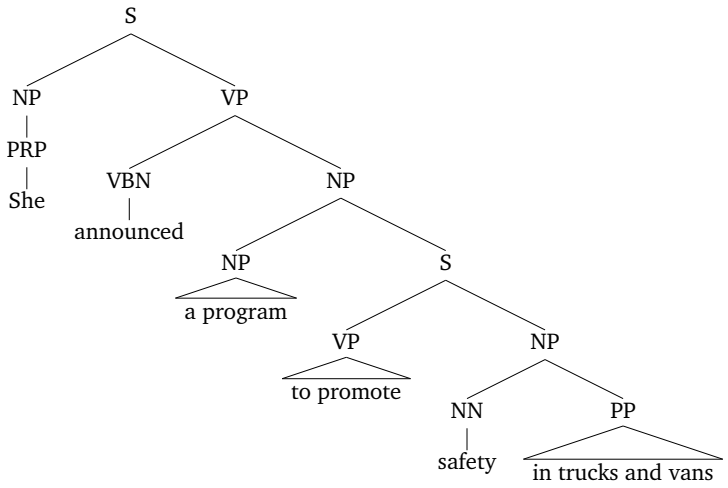


1. Dibuja los árboles de constituyentes de, al menos, otras tres interpretaciones de esta oración
2. Dibuja los árboles de, al menos, dos interpretaciones para cada una de las siguientes oraciones:
    - *The post office will hold out discounts and service concessions as incentives*
    - *They are hunting lions and tigers*
    - *Monty flies like mosquitoes*

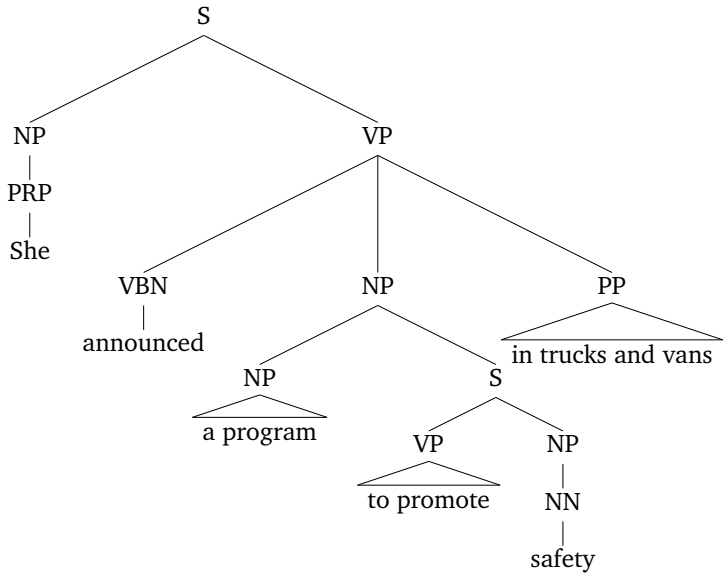
SOLUCION

1. Tres interpretaciones:

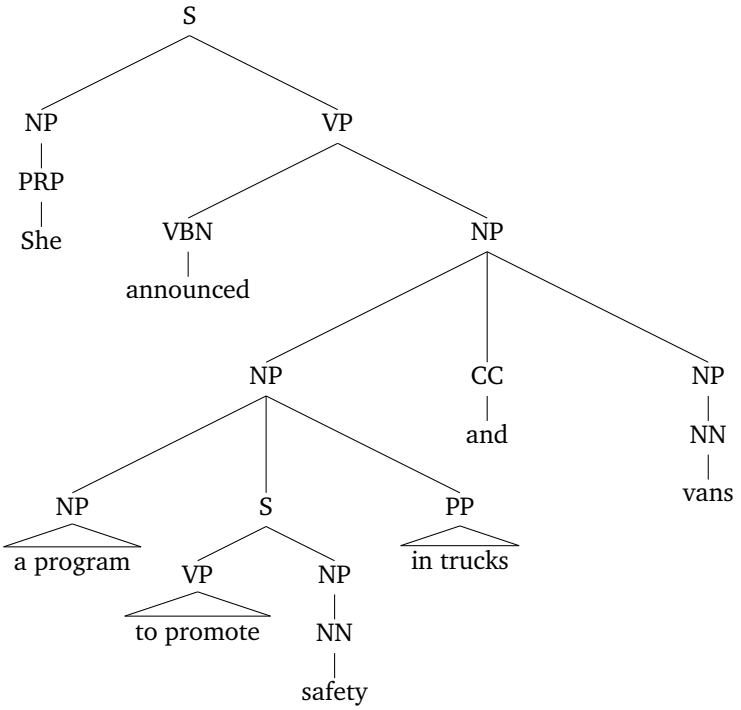
   1: El programa anunciado promueve la seguridad en ambos transportes.



   2: El programa se anuncia en camiones y furgonetas.

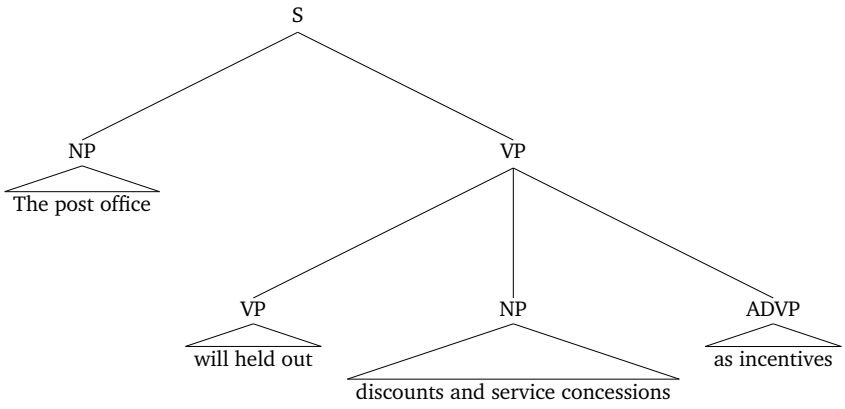


3: Ella anuncia un programa para promover la seguridad en camiones. Ella también anuncia furgonetas.

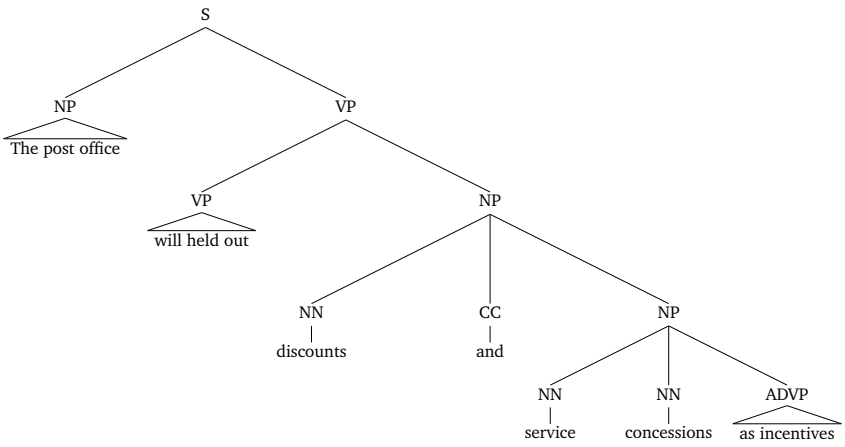


2. Encontrar dos interpretaciones para cada oración

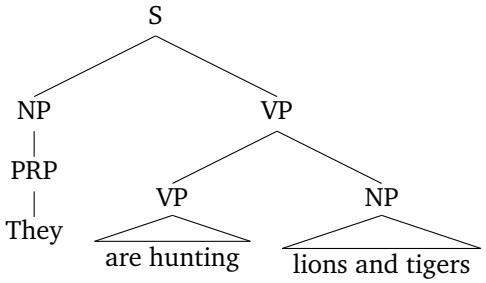
Oración 1, 1: Se mantienen los descuentos y las concesiones



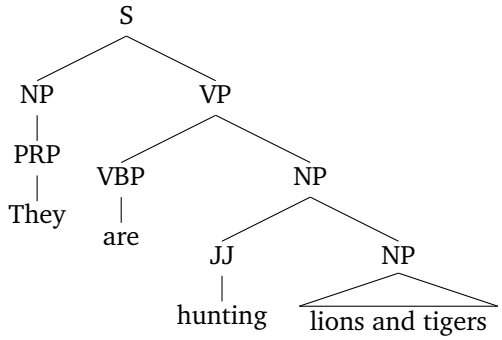
Oración 1, 2: Se mantienen los descuentos y las concesiones, estas últimas como incentivos



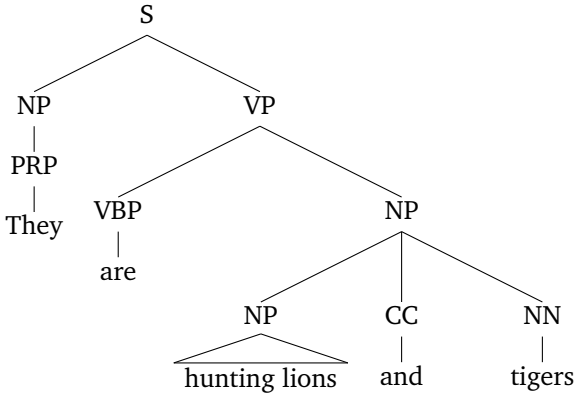
Oración 2, 1: Están cazando leones y tigres.



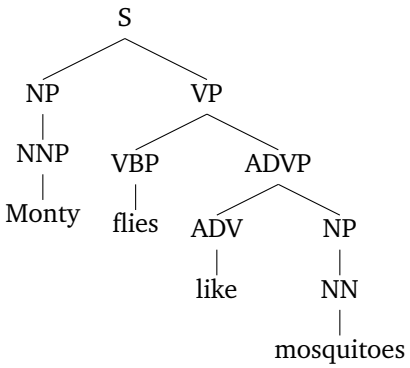
Oración 2, 2: Ambos felinos son cazadores



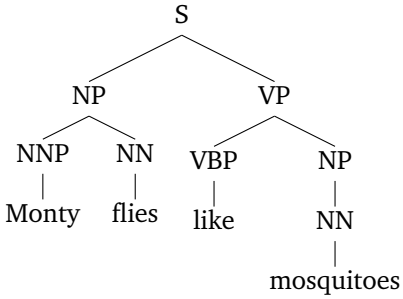
Oracion 2, 3: Ellos son leones cazadores y tigres



Oración 3, 1: Monty vuela como los mosquitos

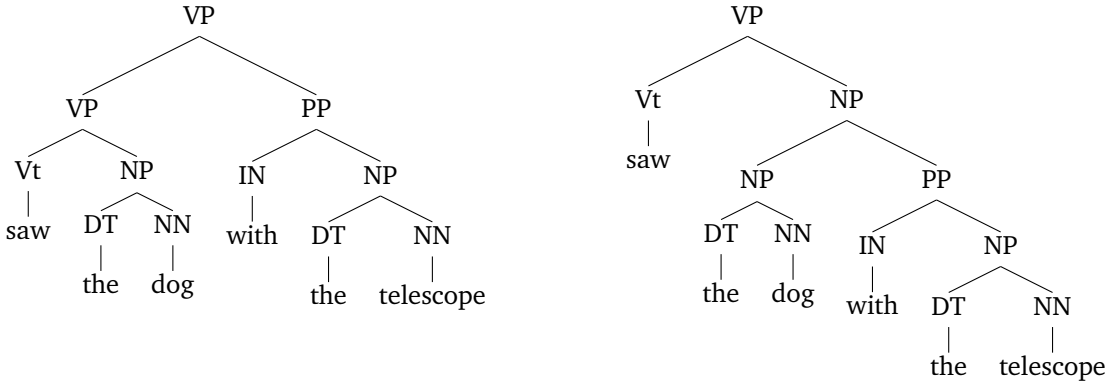


Oración 3, 2: A las moscas que están en Monty les gustan los mosquitos



Ejercicio 2.

Supón la oración "*saw the dog with the telescope*", su árbol de constituyentes correcto (izquierda) y uno producido por un analizador sintáctico (derecha): and we are given the gold parse tree (left) and the predicted parse tree (right):



¿Cuales son la precisión y la cobertura del árbol producido por el analizador?

SOLUCION

El árbol correcto tiene 11 nodos no terminales, cada uno producido por una posible regla.

El árbol del analizador tiene 11 nodos no terminales, cada uno producido por una regla que puede ser la correcta o no. En particular, las reglas $\mathrm{VP} \rightarrow \mathrm{Vt\ NP}$ y $\mathrm{NP} \rightarrow \mathrm{NP\ PP}$ no se usan para obtener el árbol correcto. Todas las otras reglas sí que se usan en él, por lo que esas otras son correctas. Por lo tanto tenemos 9 nodos correctos.

$$Prec = \frac{\#nodos\_correctos}{\#nodos\_producidos} = \frac{9}{11} = 81.8\%$$

$$Cob = \frac{\#nodos\_correctos}{\#nodos\_esperados} = \frac{9}{11} = 81.8\%$$

Los nodos producidos no tienen porque ser los nodos esperados, pero en esta ocasión coinciden, con lo que se obtiene el mismo valor para precisión y cobertura.

Ejercicio 3.

Considera la siguiente CFG:

$\mathrm{S} \rightarrow \mathrm{NP\ VP}$	$\mathrm{DT} \rightarrow \mathrm{the}$	$\mathrm{NN} \rightarrow \mathrm{park}$
$\mathrm{NP} \rightarrow \mathrm{DT\ NN}$	$\mathrm{NN} \rightarrow \mathrm{man}$	$\mathrm{VB} \rightarrow \mathrm{saw}$
$\mathrm{NP} \rightarrow \mathrm{NP\ PP}$	$\mathrm{NN} \rightarrow \mathrm{dog}$	$\mathrm{IN} \rightarrow \mathrm{with}$
$\mathrm{PP} \rightarrow \mathrm{IN\ NP}$	$\mathrm{NN} \rightarrow \mathrm{cat}$	$\mathrm{IN} \rightarrow \mathrm{under}$
$\mathrm{VP} \rightarrow \mathrm{VB\ NP}$		

1. ¿Cuántos árboles genera la gramática para la oración "*the man saw the dog in the park*"?
2. ¿Cuántos para la oración "textitthe man saw the dog in the park with the cat"?

SOLUCION

1. La oración "*the man saw the dog in the park*" tiene un único análisis in esa gramática porque solo existe la posibilidad de asociar el PP "*in the park*" al NP "*the dog*" mediante la regla $\mathrm{NP} \rightarrow \mathrm{NP\ PP}$. No hay ninguna regla para que ese PP se pueda asociar al verbo
2. La oración "*the man saw the dog in the park with the cat*" tiene dos análisis: uno donde el gato está con el perro y otro donde el gato están con el parque

Ejercicio 4.

Considera la siguiente CFG:

S → NP VP	DT → the	NNS → cats
NP → DT NN	NN → man	NNS → parks
NP → DT NNS	NN → dog	VB → see
NP → NP PP	NN → cat	VB → sees
PP → IN NP	NN → park	IN → in
VP → VB NP	NNS → dogs	IN → with
VP → VP PP		

Esta gramática sobregenera oracines incorrectas para el inglés, como por ejemplo:

the dog see the cat

the dog in the park see the cat

the dog in the park see the cat in the park

the dogs sees the cat

the dogs in the park sees the cat

the dogs in the park sees the cat in the park

1. Modifica la gramática para que todas las oraciones generadas respeten la concordancia de 3a persona entre el sujeto y el verbo

SOLUCION

La regla que se requiere cambiar es S → NP VP. Necesitamos modificarla para que combine solo NP singular con VP en 3a persona y NP plural con VP de otro número de persona. Para ello, necesitamos diferentes reglas para obtener NP singulares/plurales y VP 3a persona/otra persona.
La regla S → NP VP debe ser sustituida por:
S → NPs VPs
S → NPp VPp

Todas las reglas NP deben duplicarse para distinguir singulares de plurales:
NPs → DT NN
NPp → DT NNS
NPs → NPs PP
NPp → NPp PP

Finalmente, las reglas VP deben duplicarse para distinguir 3a persona de otra:
VBs → sees
VBp → see
VPs → VBs NP
VPp → VBp NP
VPs → VPs PP
VPp → VPp PP

Para evitar una explosión de reglas podemos conservar un NP genérico que sea usado para obtener frases nominales después de un verbo o dentro de un PP:
NP → NPs
NP → NPp

PCFGs

Ejercicio 5.

Usando la siguiente PCFG en CNF:

$S \rightarrow NP\ VP$	1.0	$P \rightarrow with$	1.0
$NP \rightarrow NP\ PP$	0.4	$V \rightarrow saw$	1.0
$PP \rightarrow P\ NP$	1.0	$NP \rightarrow astronomers$	0.1
$VP \rightarrow V\ NP$	0.7	$NP \rightarrow ears$	0.18
$VP \rightarrow VP\ PP$	0.3	$NP \rightarrow saw$	0.04
		$NP \rightarrow stars$	0.18
		$NP \rightarrow telescopes$	0.1

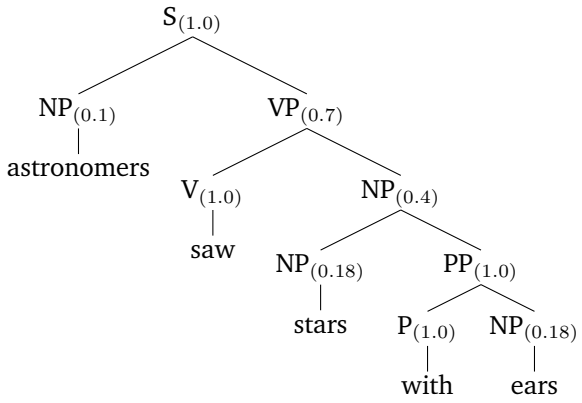
sobre la oración *astronomers saw stars with ears*

- ¿Cuántos análisis correctos se obtienen?
- Escríbelos junto con sus probabilidades.

SOLUCION

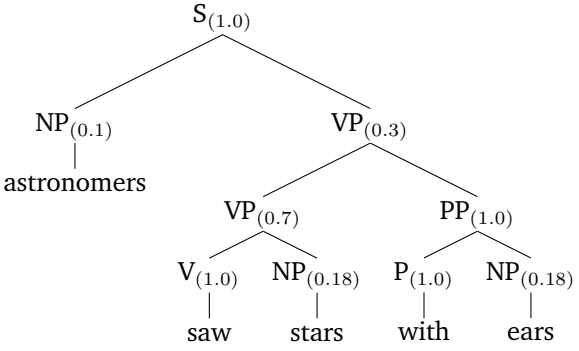
Hay 2 análisis posibles:

1: (las estrellas tienen orejas)



Probabilidad: $1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.00091$

2: (Los astronomos usan sus orejas para observar las estrellas)



Probabilidad: $1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.00068$

Ejercicio 6.

Dada la siguiente PCFG:

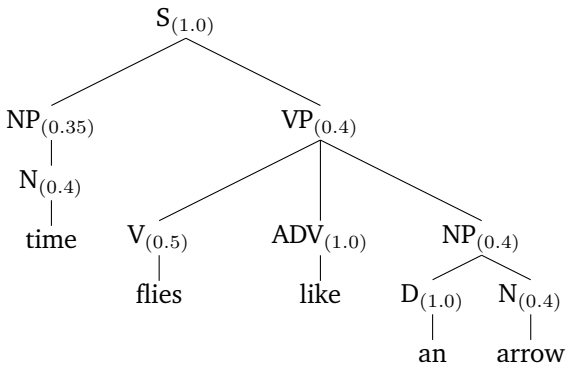
$S \rightarrow NP\ VP$	1.0	$N \rightarrow time$	0.4
$NP \rightarrow N\ N$	0.25	$N \rightarrow flies$	0.2
$NP \rightarrow D\ N$	0.4	$N \rightarrow arrow$	0.4
$NP \rightarrow N$	0.35	$D \rightarrow an$	1.0
$VP \rightarrow V\ NP$	0.6	$ADV \rightarrow like$	1.0
$VP \rightarrow V\ ADV\ NP$	0.4	$V \rightarrow flies$	0.5
		$V \rightarrow like$	0.5

y la oración *time flies like an arrow*

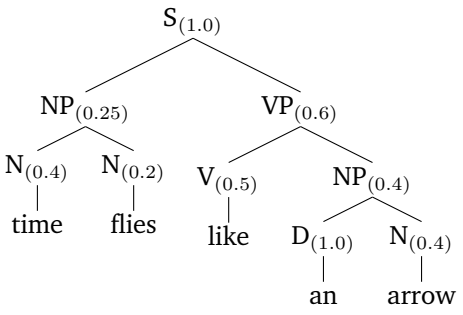
1. Escribe 2 árboles de constituyentes que esa PCFG genere para la oración dada.
2. Computa la probabilidad de cada uno de esos árboles.
3. Convierte la gramática a CNF y simula el comportamiento del algoritmo PCKY sobre la oración. Proporciona el chart final con toda la información involucrada en el proceso.

SOLUCION

1. Opción 1: (el tiempo pasa tan rápido que parece una flecha)



Option 2: (A las moscas 4-dimensionales les gusta una flecha)



2. Probabilidad opción 1: $1.0 \times 0.35 \times 0.4 \times 0.4 \times 0.5 \times 1.0 \times 0.4 \times 1.0 \times 0.4 = 0.00448$
   Probabilidad opción 2: $1.0 \times 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.5 \times 0.4 \times 1.0 \times 0.4 = 0.00096$

3. (a) Conversión de la gramática a FNC:

   La FNC requiere que todas las reglas tengan un no terminal a la izquierda y dos no terminales o un terminal a la derecha. Las reglas $NP \rightarrow N$ y $VP \rightarrow V\ ADV\ NP$ violan esta condición, por lo que necesitan ser transformadas.

   Por una parte, comprimimos la regla $NP \rightarrow N$ con aquellas que producen un N, es decir, con $N \rightarrow time$, $N \rightarrow flies$ y $N \rightarrow arrow$ y producimos las siguientes:

   $NP \rightarrow time\ 0.35 \times 0.4 = 0.14$

   $NP \rightarrow flies\ 0.35 \times 0.2 = 0.07$

   $NP \rightarrow arrow$,i $0.35 \times 0.4 = 0.14$

   Eliminamos la que no está en FNC, pero mantenemos las otras 3 reglas porque existen reglas en la gramática que las necesita (p.e., $NP \rightarrow N\ N$).

   Por otra parte, partimos la regla $VP \rightarrow V\ ADV\ NP$ para obtener dos reglas en FNC:

   $VP \rightarrow V\ ADVP\ 0.4$

   $ADVP \rightarrow ADV\ NP\ 1.0$

   La gramática en FNC resultante es la siguiente:

S → NP VP	1.0	N → time	0.4
NP → N N	0.25	N → flies	0.2
NP → D N	0.4	N → arrow	0.4
NP → time	0.14	D → an	1.0
NP → flies	0.07	ADV → like	1.0
NP → arrow	0.14	V → flies	0.5
VP → V NP	0.6	V → like	0.5
VP → V ADVP	0.4		
ADVP → ADV NP	1.0		

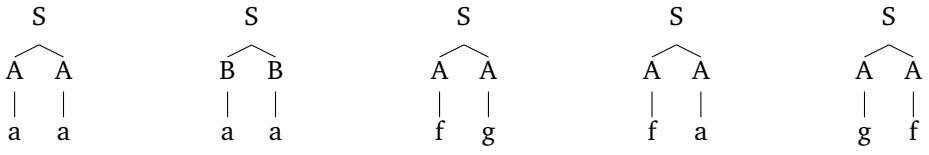
(b) chart CKY :

				15 0.00448 S → $N_{11}VP_{25}$ $(1.0 \times 0.14 \times 0.032)$ 0.00096 S → $NP_{12}VP_{35}$ $(1.0 \times 0.02 \times 0.048)$
			14	25 0.0033 S → $N_{11}VP_{25}$ $(1.0 \times 0.07 \times 0.048)$ 0.032 VP → $V_{22}ADVP_{35}$ $(0.4 \times 0.5 \times 0.16)$
		13	24	35 0.16 ADVP → $ADV_{33}NP_{45}$ $(1.0 \times 1.0 \times 0.16)$ 0.048 VP → $V_{33}NP_{45}$ $(0.6 \times 0.5 \times 0.16)$
	12 0.02 NP → $N_{11}N_{22}$ $(0.25 \times 0.4 \times 0.2)$	23	34	45 0.16 NP → $D_{44}N_{55}$ $(0.4 \times 1.0 \times 0.4)$
11 0.14 NP → time 0.4 N → time	22 0.07 NP → flies 0.2 N → flies 0.5 V → flies	33 1.0 ADV → like 0.5 V → like	44 1.0 D → an	55 0.4 N → arrow
time	flies	like	an	arrow

(La regla azul indica el subárbol más probable seleccionado en la celda)

Ejercicio 7.

Supongamos que los siguientes árboles están dentro del treebank de entrenamiento. Cada árbol se observó dentro del treebank las veces indicadas bajo ellos.



75 10 325 8 428

1. ¿Qué PCFG se consigue a partir de ellos usando máxima verosimilitud (MLE)?
2. Dada la gramática obtenida:
   - ¿Cuál es el análisis mas probable para la secuencia "*a a*"?
   - ¿Es un resultado razonable? Justifica tu respuesta.

SOLUCION

1. Se obtienen las siguientes reglas y sus contages:

$\mathrm{S} \to \mathrm{A\ A}$	$1 \times 75 + 0 \times 10 + 1 \times 325 + 1 \times 8 + 1 \times 428 = 836$
$\mathrm{S} \to \mathrm{B\ B}$	$0 \times 75 + 1 \times 10 + 0 \times 325 + 0 \times 8 + 0 \times 428 = 10$
$\mathrm{S} \to$ anything	$1 \times 75 + 1 \times 10 + 1 \times 325 + 1 \times 8 + 1 \times 428 = 846$
$\mathrm{B} \to \mathrm{a}$	$0 \times 75 + 2 \times 10 + 0 \times 325 + 0 \times 8 + 0 \times 428 = 20$
$\mathrm{B} \to$ anything	$0 \times 75 + 2 \times 10 + 0 \times 325 + 0 \times 8 + 0 \times 428 = 20$
$\mathrm{A} \to \mathrm{a}$	$2 \times 75 + 0 \times 10 + 0 \times 325 + 1 \times 8 + 0 \times 428 = 158$
$\mathrm{A} \to \mathrm{f}$	$0 \times 75 + 0 \times 10 + 1 \times 325 + 1 \times 8 + 1 \times 428 = 761$
$\mathrm{A} \to \mathrm{g}$	$0 \times 75 + 0 \times 10 + 1 \times 325 + 0 \times 8 + 1 \times 428 = 753$
$\mathrm{A} \to$ anything	$2 \times 75 + 0 \times 10 + 2 \times 325 + 2 \times 8 + 2 \times 428 = 1672$

Así que las estimaciones por MLE de las probabilidades para cada regla son:

$P(\mathrm{S} \to \mathrm{A\ A})$	=	$P(\mathrm{AA}\vert\mathrm{S})$	=	$\#(\mathrm{S} \to \mathrm{A\ A})/\#(\mathrm{S} \to$ anything$)$	=	$836/846 = 0.988$
$P(\mathrm{S} \to \mathrm{B\ B})$	=	$P(\mathrm{BB}\vert\mathrm{S})$	=	$\#(\mathrm{S} \to \mathrm{B\ B})/\#(\mathrm{S} \to$ anything$)$	=	$10/846 = 0.012$
$P(\mathrm{B} \to \mathrm{a})$	=	$P(\mathrm{a}\vert\mathrm{B})$	=	$\#(\mathrm{B} \to \mathrm{a})/\#(\mathrm{B} \to$ anything$)$	=	$20/20 = 1.000$
$P(\mathrm{A} \to \mathrm{a})$	=	$P(\mathrm{a}\vert\mathrm{A})$	=	$\#(\mathrm{A} \to \mathrm{a})/\#(\mathrm{A} \to$ anything$)$	=	$158/1672 = 0.095$
$P(\mathrm{A} \to \mathrm{f})$	=	$P(\mathrm{f}\vert\mathrm{A})$	=	$\#(\mathrm{A} \to \mathrm{f})/\#(\mathrm{A} \to$ anything$)$	=	$761/1672 = 0.455$
$P(\mathrm{A} \to \mathrm{g})$	=	$P(\mathrm{g}\vert\mathrm{A})$	=	$\#(\mathrm{A} \to \mathrm{g})/\#(\mathrm{A} \to$ anything$)$	=	$753/1672 = 0.450$

2. La secuencia "*a a*" puede ser derivada de dos maneras, correspondientes a los 2 primeros árboles de entrenamiento dados.
   La probabilidad del 1er árbol es: $P(\mathrm{S} \to \mathrm{A\ A}) \times P(\mathrm{A} \to \mathrm{a}) \times P(\mathrm{A} \to \mathrm{a}) = 0.988 \times 0.095 \times 0.095 = 0.009$ La probabilidad del 2o árbol es: $P(\mathrm{S} \to \mathrm{B\ B}) \times P(\mathrm{B} \to \mathrm{a}) \times P(\mathrm{B} \to \mathrm{a}) = 0.012 \times 1.000 \times 1.000 = 0.012$
   Por lo que el análisis mas probable es el segundo.

   El 1er árbol ocurre 75 veces en los datos de entrenamiento mientras que el 2o solo ocurre 10 veces, así que parecería que la probabilidad del 1o es mayor. Sin embargo, la probabilidad de una árbol no se computa contando cuantas veces ocurre el árbol, sino que la aproximamos multiplicando las probabilidades de las reglas que lo derivan. Dado que B produce a con mayor probabilidad que A, la probabilidad del árbol queda sesgada por este hecho.
   Esto es debido a la escasa cantidad de datos de entrenamiento y al uso de MLE: dado que no aplicamos *smoothing* para considerar la probabilidad de que B produzca otros símbolos, estamos sobreestimando la probabilidad de la regla $\mathrm{B} \to \mathrm{a}$.