Gramáticas de Adjunción de Árboles

(TAG)

Miguel A. Alonso Pardo

Departamento de Computación, Universidade da Coruña

Análisis sintáctico de TAG

- Conceptos previos
- Gramáticas de adjunción de árboles
- Analizadores sintácticos para TAG
- Definición de los ítems
- Definición de los pasos deductivos
- Resultados experimentales

Análisis sintáctico de TAG

- **Conceptos previos**
- Gramáticas de adjunción de árboles
- Analizadores sintácticos para TAG
- Definición de los ítems
- Definición de los pasos deductivos
- Resultados experimentales

La estructura de las frases

- La **sintaxis** describe la **estructura** de las frases de un lenguaje
- Una frase bien formada se puede descomponer en constituyentes de acuerdo a unas **reglas sintácticas**
- Las reglas sintácticas se describen mediante un **formalismo gramatical**
- Un **analizador sintáctico** es un programa de ordenador que obtiene la estructura sociada a una frase dada de acuerdo con una gramática.

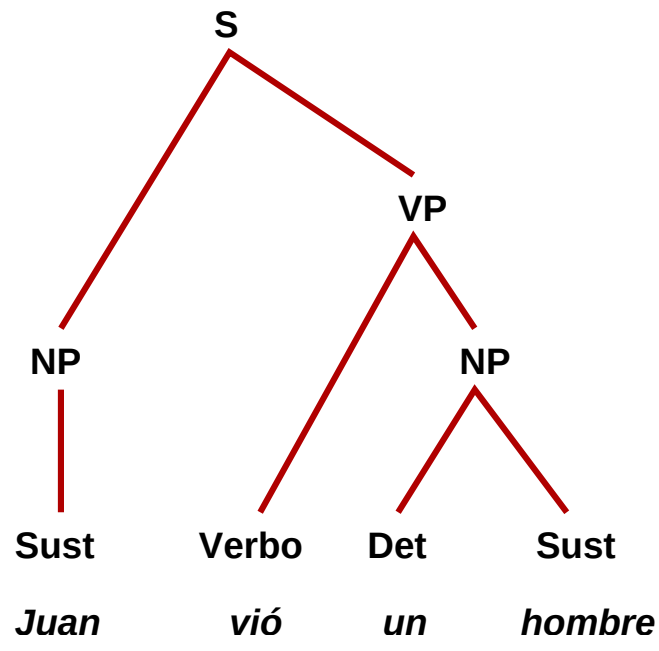
La estructura de las frases

- Frase: “Juan vio un hombre”
- Gramática: $S \rightarrow NP\ VP$

  $NP \rightarrow Sust$

  $NP \rightarrow Det\ Sust$

  $VP \rightarrow Verbo\ NP$
- Estructura:



Problema: la ambigüedad

Las gramáticas de los lenguajes naturales son ambiguas: se pueden asociar varias estructuras a una misma frase

S $\rightarrow$ NP VP

S $\rightarrow$ S PP

NP $\rightarrow$ Sust

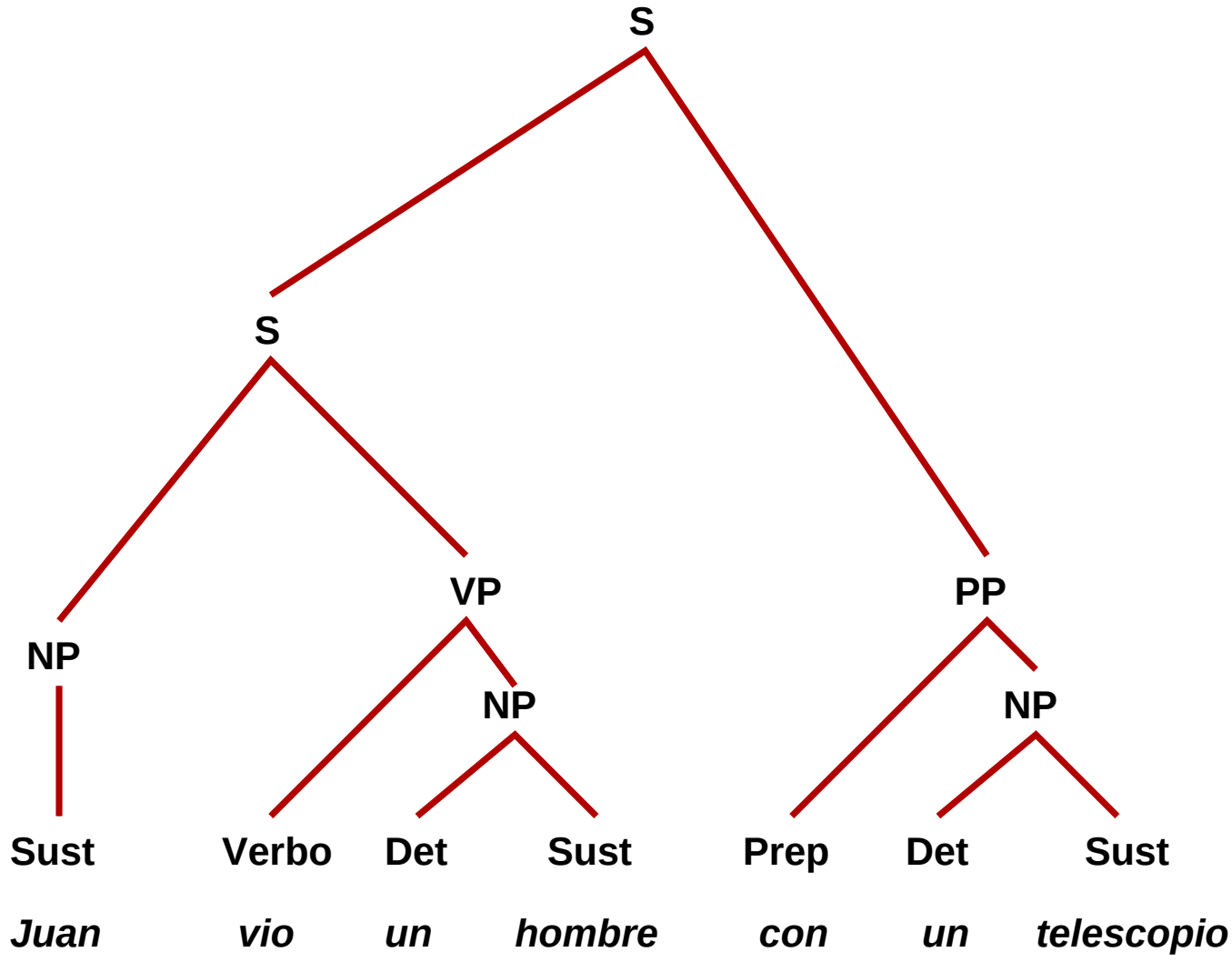
NP $\rightarrow$ Det Sust

NP $\rightarrow$ NP PP

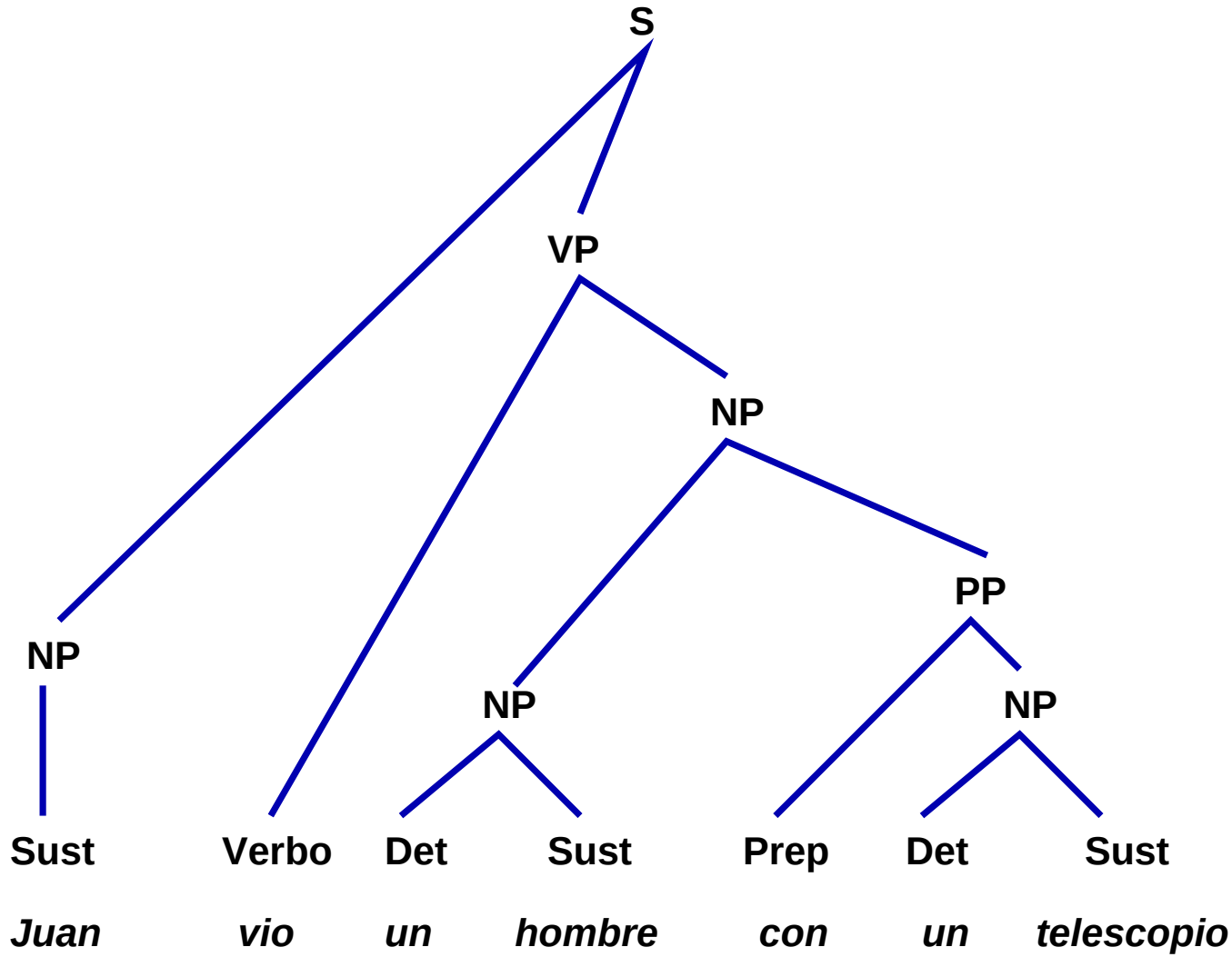
PP $\rightarrow$ Prep NP

VP $\rightarrow$ Verbo NP

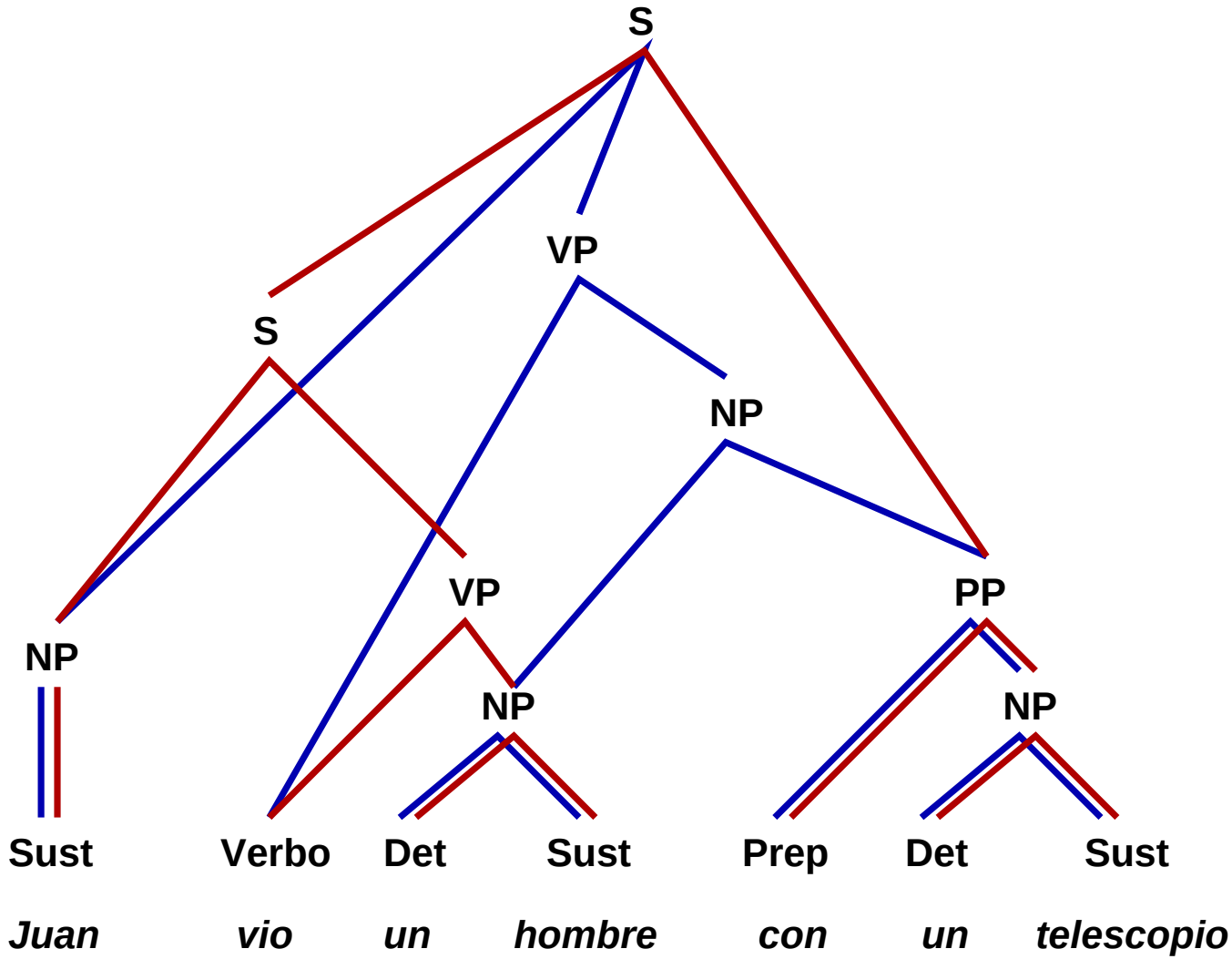
Problema: la ambigüedad



Problema: la ambigüedad



Problema: la ambigüedad



Formalismos gramaticales

Regulares $\longrightarrow$ Indep. contexto $\longrightarrow$ Sensibles contexto $\longrightarrow$ Sin restricciones

Formalismos gramaticales

Regulares $\longrightarrow$ Indep. contexto $\longrightarrow$ Sensibles contexto $\longrightarrow$ Sin restricciones

- Las **gramáticas independientes del contexto** son un formalismo muy popular, que puede ser analizado en tiempo $\mathcal{O}(n^3)$ y espacio $\mathcal{O}(n^2)$ ...

Formalismos gramaticales

Regulares $\longrightarrow$ Indep. contexto $\longrightarrow$ Sensibles contexto $\longrightarrow$ Sin restricciones

- Las **gramáticas independientes del contexto** son un formalismo muy popular, que puede ser analizado en tiempo $\mathcal{O}(n^3)$ y espacio $\mathcal{O}(n^2)$ ...
- ...pero no son lo suficientemente potentes para describir la sintaxis de los lenguajes naturales

Formalismos gramaticales

Regulares $\longrightarrow$ Indep. contexto $\longrightarrow$ Sensibles contexto $\longrightarrow$ Sin restricciones

- Las **gramáticas independientes del contexto** son un formalismo muy popular, que puede ser analizado en tiempo $\mathcal{O}(n^3)$ y espacio $\mathcal{O}(n^2)$ ...
- ...pero no son lo suficientemente potentes para describir la sintaxis de los lenguajes naturales
- Las **gramáticas sensibles al contexto** son “demasiado potentes”, con un análisis sintáctico de complejidad exponencial

Formalismos gramaticales

Regulares $\longrightarrow$ Indep. contexto $\longrightarrow$ Sensibles contexto $\longrightarrow$ Sin restricciones

- Las **gramáticas independientes del contexto** son un formalismo muy popular, que puede ser analizado en tiempo $\mathcal{O}(n^3)$ y espacio $\mathcal{O}(n^2)$ ...
- ...pero no son lo suficientemente potentes para describir la sintaxis de los lenguajes naturales
- Las **gramáticas sensibles al contexto** son "demasiado potentes", con un análisis sintáctico de complejidad exponencial

Reg. $\longrightarrow$ Indep. contexto $\longrightarrow$ **MCS** $\longrightarrow$ Sensibles contexto $\longrightarrow$ Sin restric.

Formalismos gramaticales

- Los lenguajes y formalismos **suavemente sensibles al contexto** poseen propiedades interesantes para el procesamiento del lenguaje natural:
  - Inclusión de los lenguajes independientes del contexto
  - Análisis sintáctico de complejidad polinomial
  - Dependencias anidadas y cruzadas
  - Propiedad del crecimiento constante
- Los **lenguajes de adjunción de árboles** son un subconjunto propio de los lenguajes suavemente sensibles al contexto que satisfacen estas propiedades

Análisis sintáctico de TAG

- Conceptos previos
- **Gramáticas de adjunción de árboles**
- Analizadores sintácticos para TAG
- Definición de los ítems
- Definición de los pasos deductivos

Leng. de adjunción de árboles

$$\text{CFL} \subseteq \textbf{TAL} \subseteq \textbf{MCSL} \subseteq \text{CSL}$$

Son generados por diversos formalismos, equivalentes en cuanto a su capacidad generativa:

- **Gramáticas de adjunción de árboles (TAG)**
- Gramáticas lineales de índices (LIG)
- Gramáticas categoriales combinatorias (CCG)
- Gramáticas de núcleo (HG)
- . . .

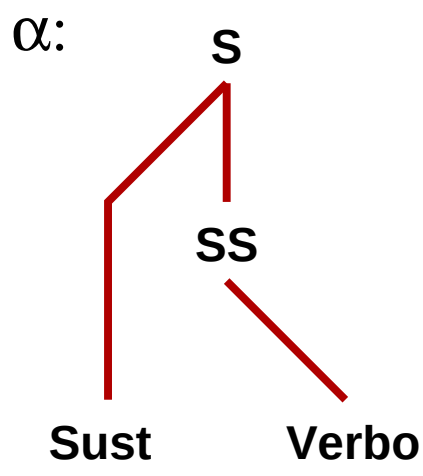
Gram. de adjunción de árboles

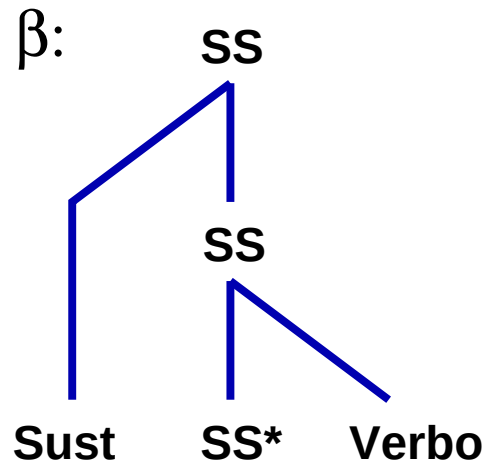
- Una extensión de los gramáticas independientes del contexto que utiliza árboles en vez de proudcciones como estructura de representación básica
- Formalmente, una tupla $(V_N, V_T, S, \boldsymbol{I}, \boldsymbol{A})$ donde:
  - V_N es un conjunto finito de símbolos no terminales
  - V_T es un conjunto finito de símbolos terminales
  - $S \in V_N$ es el símbolo inicial de la gramática
  - $\boldsymbol{I}$ es un conjunto finito de **árboles iniciales**
  - $\boldsymbol{A}$ es un conjunto finito de **árboles auxiliares**

  $\boldsymbol{I} \cup \boldsymbol{A}$ es el conjunto de **árboles elementales**

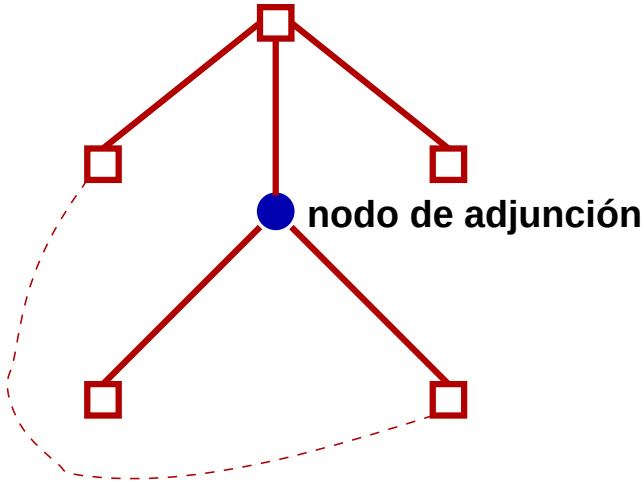
Árboles iniciales y auxiliares

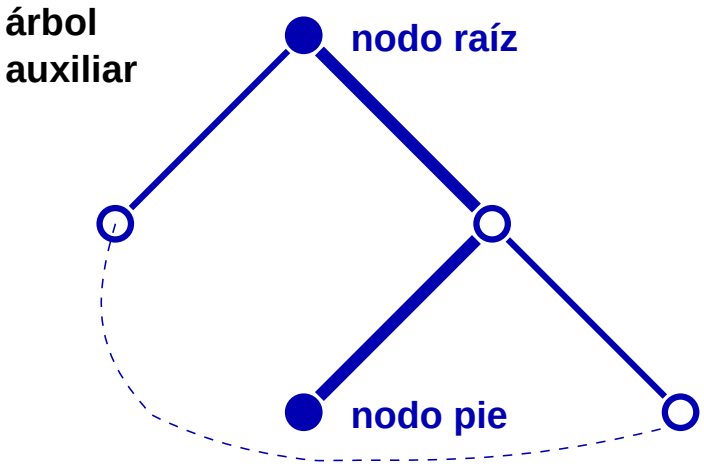
- Los árboles iniciales representan frases mínimas
- Los árboles auxiliares representan estructuras recursivas mñinimas



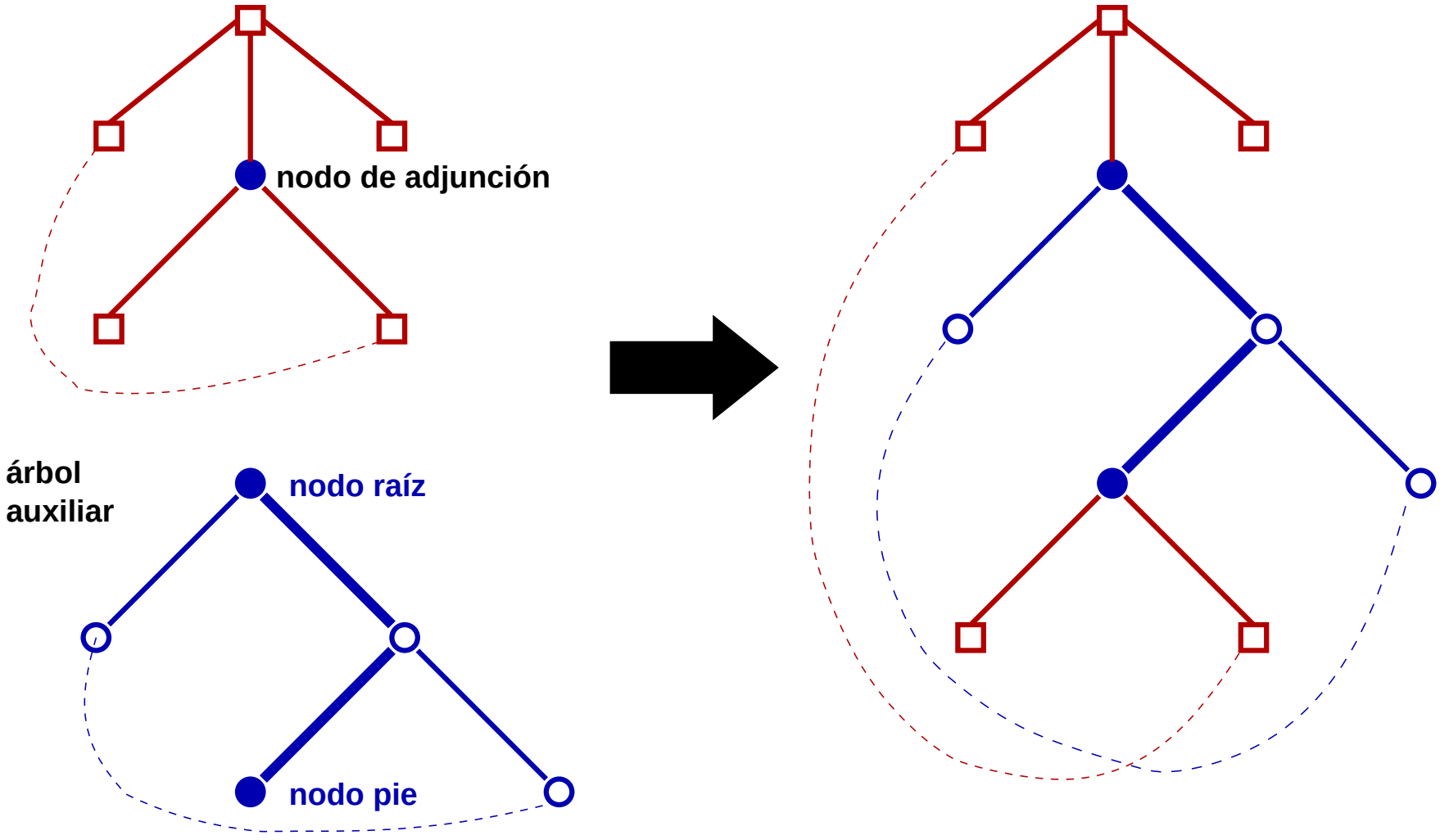


La operación de adjunción





La operación de adjunción



Motivación lingüística

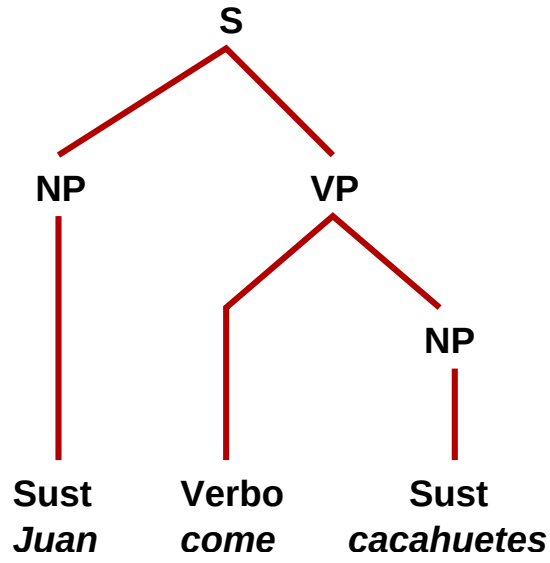
Las gramáticas de adjunción de árboles son (también) interesantes desde el punto de vista lingüístico por las siguientes características:

- La extensión del dominio de localidad
- La factorización de la recursión en el dominio de dependencias
- La posibilidad de representar dependencias cruzadas
- Su carácter lexicalizado

El dominio de localidad

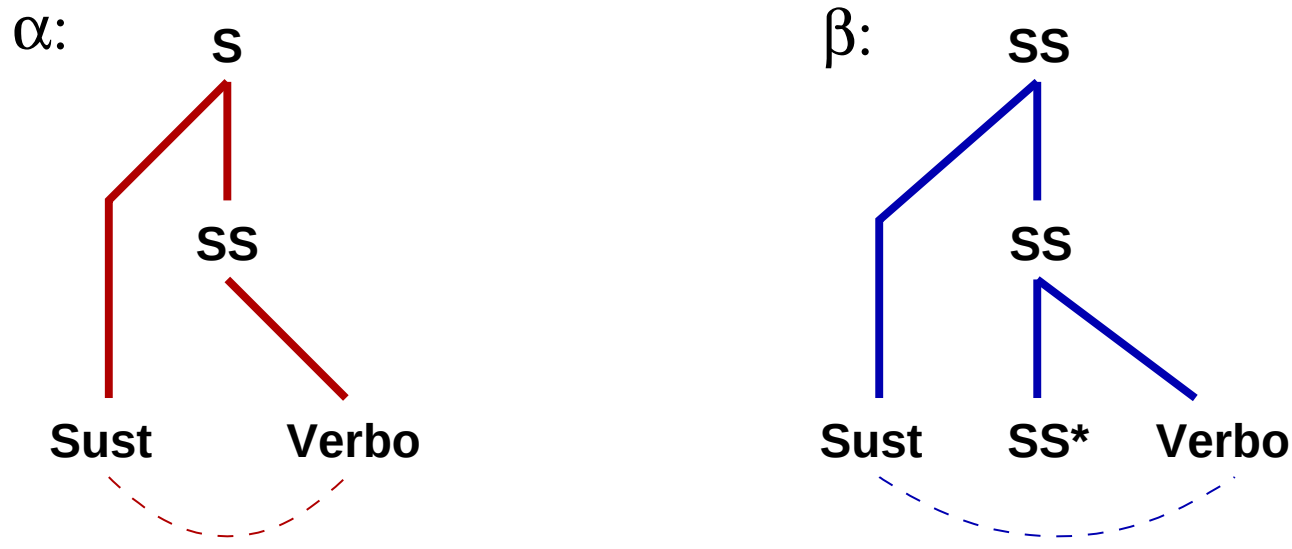
- Las TAG tienen un dominio de localidad mayor que las CFG y los formalismos basados en un esqueleto independiente del contexto
- Ejemplo: en una CFG no se pueden especificar las dependencias verbo-sujeto y verbo-objeto sin perder el nodo VP

  En cambio, estas dependencias se pueden representar fácilmente en un árbol elemental:

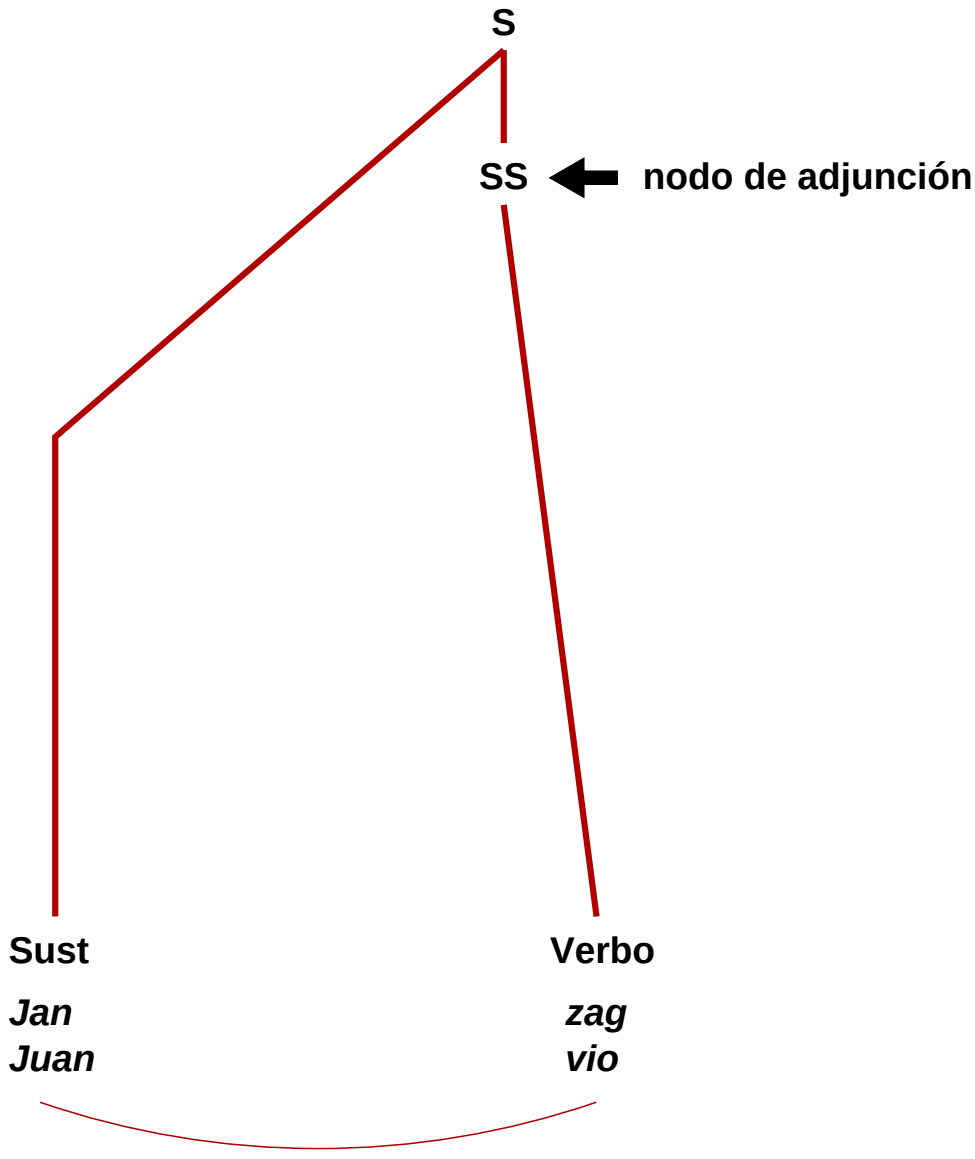


La factorización de la recursión

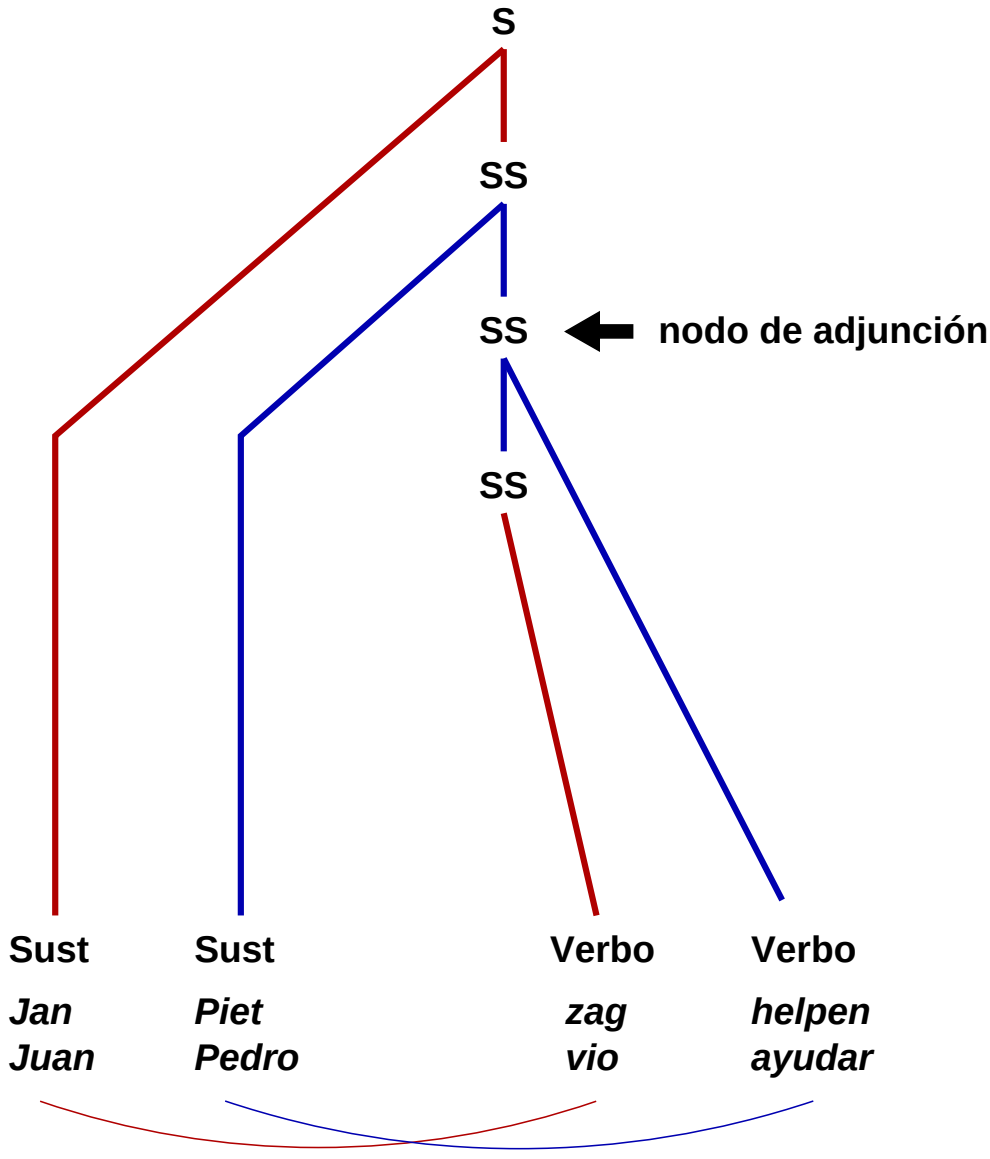
- Los árboles auxiliares son las estructuras recursivas
- El dominio de localidad extendido de los árboles permite que la unidad de recursión incluya las dependencias pertinentes
- Retomamos un ejemplo ya mostrado:



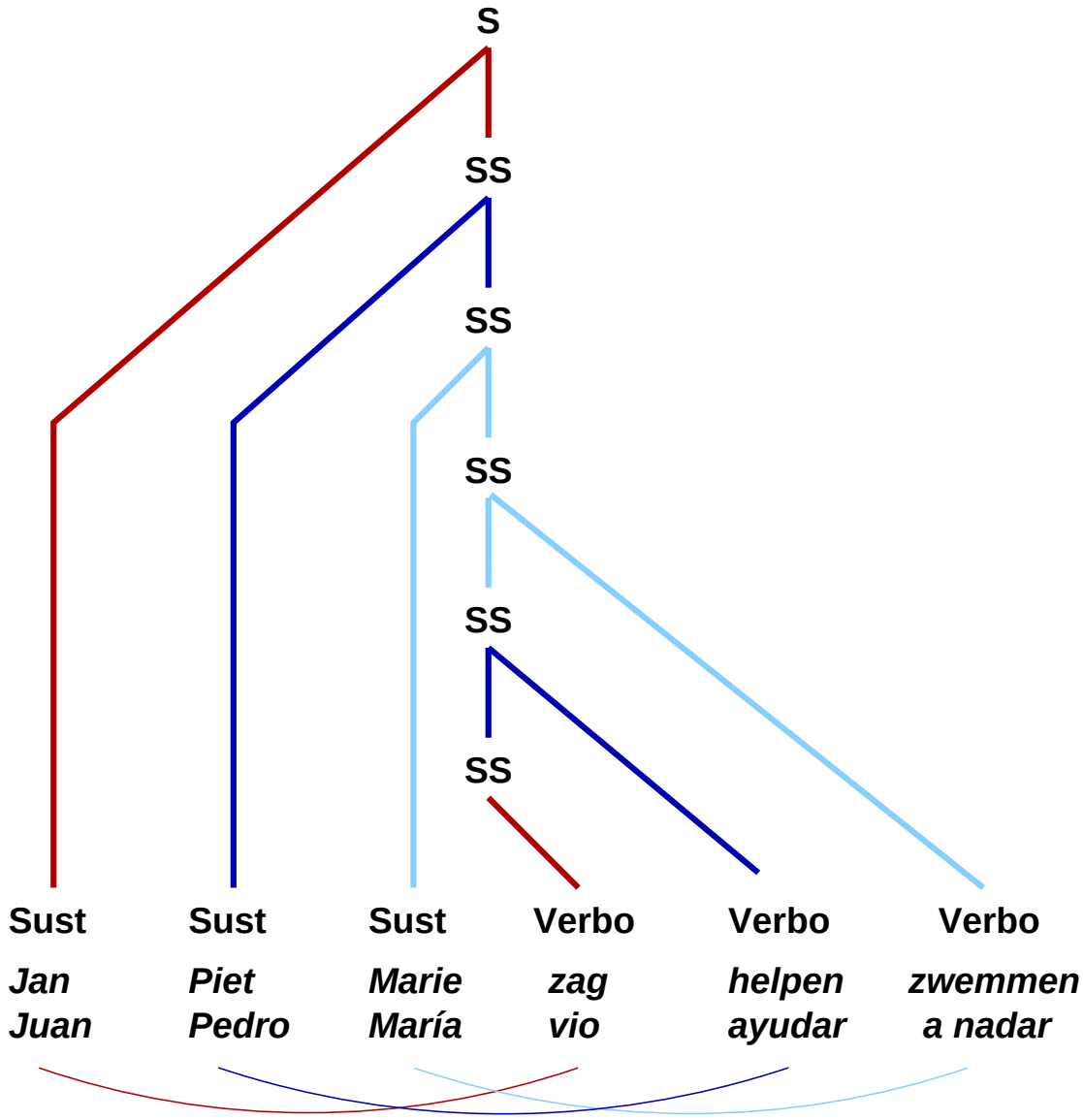
La dependencias cruzadas



La dependencias cruzadas



La dependencias cruzadas



Lexicalización

- Una gramática está lexicalizada si cada estructura elemental (producción, árbol,. . .) está asociada a un símbolo terminal, denominado **ancla**
- Las gramáticas independientes del contexto no son cerradas con respecto a la lexicalización (sin cambiar la forma de los árboles generados)
- Las gramáticas de adjunción de árboles sí son cerradas con respecto a la lexicalización

Análisis sintáctico de TAG

- Conceptos previos
- Gramáticas de adjunción de árboles
- **Analizadores sintácticos para TAG**
- Definición de los ítems
- Definición de los pasos deductivos
- Resultados experimentales

Estado del arte

1965: Algoritmo $\mathcal{O}(n^3)$ de Cocke, Younger & Kasami (CYK) para CFG

1968: Algoritmo $\mathcal{O}(n^3)$ de Earley para CFG

1975: Introducción de las TAG (Joshi, Levy & Takahashi)

1985: Algoritmo $\mathcal{O}(n^6)$ de tipo CYK (Vijay-Shanker & Joshi)

1988: Algoritmo $\mathcal{O}(n^9)$ de tipo Earley con VPP (Schabes & Joshi)

1988: Algoritmo $\mathcal{O}(n^6)$ de tipo Earley sin VPP (Lang)

1990: Algoritmo $\mathcal{O}(n^6)$ de tipo Earley sin VPP (Schabes & Joshi)

1997: Algoritmo $\mathcal{O}(n^6)$ de tipo Earley con VPP (Nederhof)

1998: Algoritmos $\mathcal{O}(n^6)$ mediante autómatas (Alonso *et al.*)

1999: Relaciones formales entre algoritmos (Alonso *et al.*)

Prop. del prefijo válido (VPP)

Los analizadores sintácticos que satisfacen la VPP garantizan que, en tanto que leen la cadena de entrada de izquierda a derecha, las subcadenas leídas son prefijos válidos del lenguaje definido por la gramática.

Formalmente:

- dada la cadena de entrada $a_1 \ldots a_k a_{k+1} \ldots a_n$
- para cada subcadena $a_1 \ldots a_k$ leída
- garantizan que $\exists\ b_1 \ldots b_m \in V_T^*$ tal que $a_1 \ldots a_k b_1 \ldots b_m \in \mathcal{L}$

Prog. dinámica o tabulación

- Los algoritmos de análisis sintáctico para gramáticas de adjunción de árboles hacen uso de la programación dinámica:
  - almacenando los resultados intermedios en **items**
  - combinando ítems mediante los **pasos deductivos** para obtener análisis de porciones cada vez mayores de la cadena de entrada
  - evitando la realización de cálculos redundantes
- Complejidad temporal $\mathcal{O}(n^6)$
- Complejidad espacial $\mathcal{O}(n^4)$ sin VPP y $\mathcal{O}(n^5)$ con VPP

Algoritmo genérico

- Entradas: una cadena de entrada
  una gramática de adjunción de árboles
- Salida: un conjunto de ítems que representa el proceso de análisis
- Estructuras de datos auxiliares: una **tabla** de ítems
  una cola de ítems (**agenda**)
- Proceso:
  1. Aplicar los pasos deductivos que no tienen ítems antecedentes, almacenando los ítems resultantes en la tabla y en la agenda
  2. Repetir hasta que la agenda esté vacía:
     a) Sacar un ítem de la agenda
     b) Aplicar todos los pasos deductivos que lo utilicen como antecedente
     c) Almacenar los ítems consecuentes en la tabla y en la agenda
- En los pasos 1) y 2.c) se comprobará que los ítems no estén repetidos

Análisis sintáctico de TAG

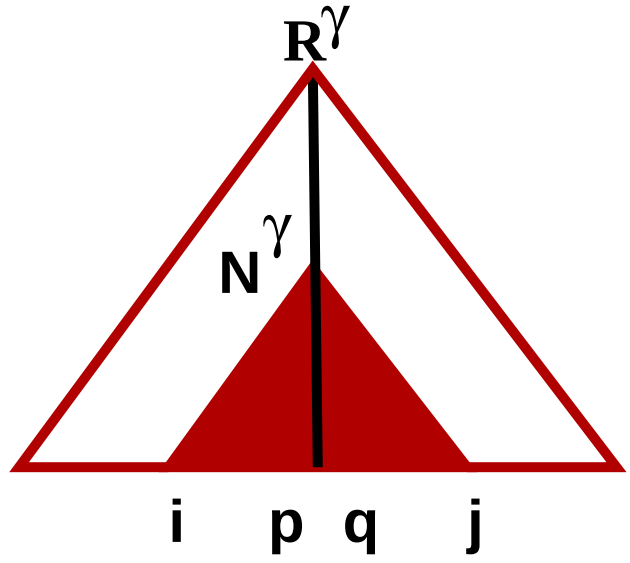
- Conceptos previos
- Gramáticas de adjunción de árboles
- Analizadores sintácticos para TAG
- **Definición de los ítems**
- Definición de los pasos deductivos
- Resultados experimentales

Ítems ascendentes

$$\mathcal{I}_{\mathrm{CYK}} = \left\{ \; [N^{\gamma}, i, j \mid p, q \mid adj] \; \right\}$$

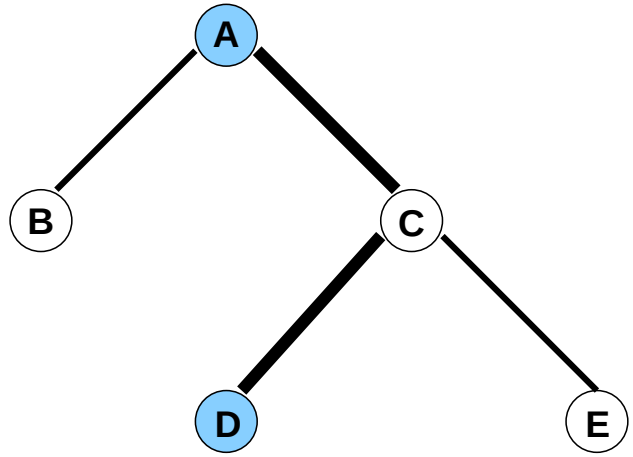
$N^{\gamma} \stackrel{*}{\Rightarrow} a_{i+1} \ldots a_p \; \mathbf{F}^{\gamma} \; a_{q+1} \ldots a_j \stackrel{*}{\Rightarrow} a_{i+1} \ldots a_j$ sii $(p, q) \neq (-, -)$

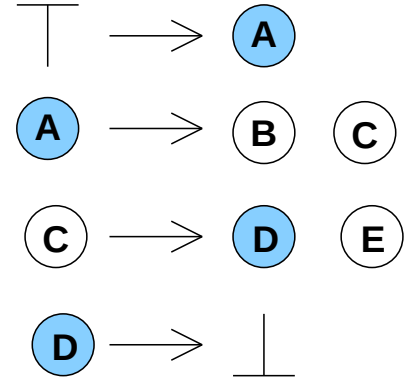
$N^{\gamma} \stackrel{*}{\Rightarrow} a_{i+1} \ldots a_j$ sii $(p, q) = (-, -)$



Ítems mixtos

- consideramos cada árbol elemental como un conjunto de producciones
- Para cada árbol inicial α añadimos la producción $\top \rightarrow \mathbf{R}^{\alpha}$
- Para cada árbol auxiliar β añadimos la producción $\top \rightarrow \mathbf{R}^{\beta}$ y $\mathbf{F}^{\beta} \rightarrow \bot$, donde $\mathbf{R}^{\beta}$ y $\mathbf{F}^{\beta}$ son los nodos ráiz y pie de β



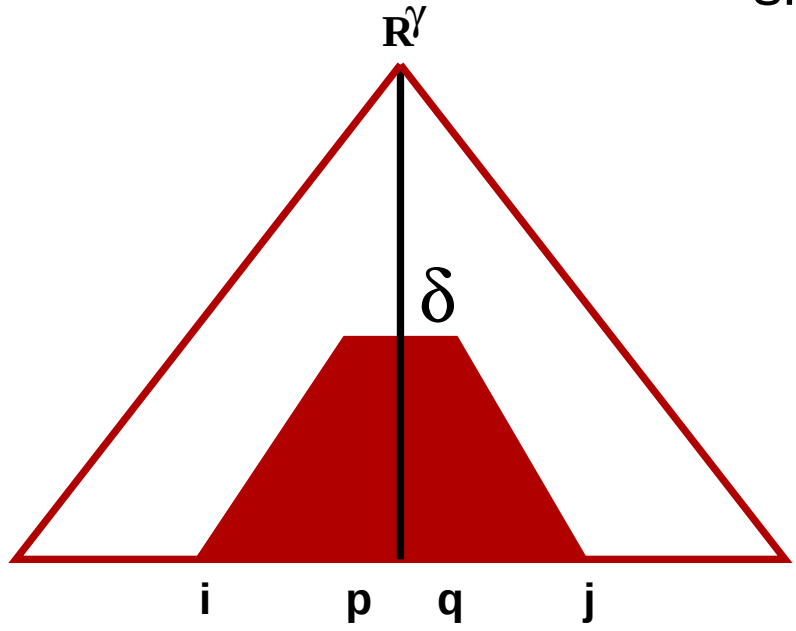


Ítems mixtos

$$\mathcal{I}_{\mathrm{E}} = \left\{ \; [N^{\gamma} \rightarrow \delta \bullet \nu, i, j \mid p, q] \; \right\}$$

$\delta \stackrel{*}{\Rightarrow} a_{i+1} \ldots a_p \; \mathbf{F}^{\gamma} \; a_{q+1} \ldots a_j \stackrel{*}{\Rightarrow} a_{i+1} \ldots a_j$ sii $(p, q) \neq (-, -)$

$\delta \stackrel{*}{\Rightarrow} a_{i+1} \ldots a_j$ sii $(p, q) = (-, -)$



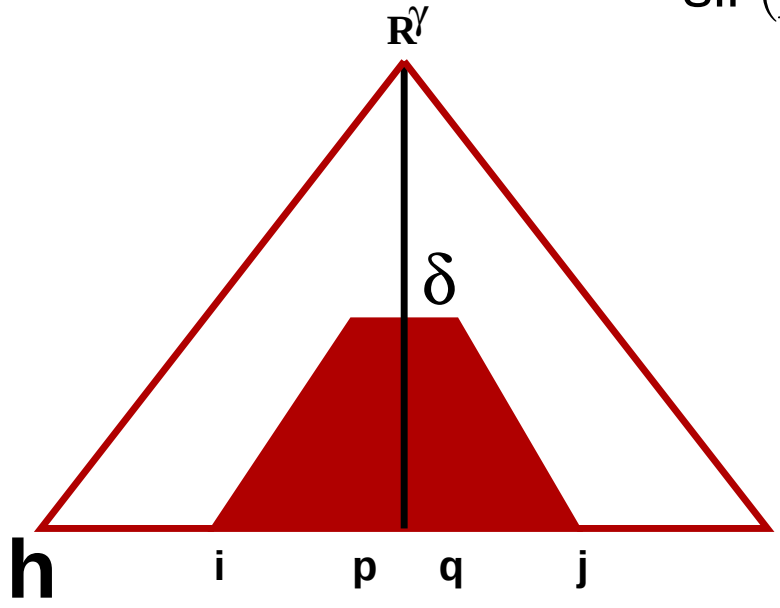
Ítems mixtos con VPP

$$\mathcal{I}_{\text{Earley}} = \left\{ \; [h, N^{\gamma} \rightarrow \delta \bullet \nu, i, j \mid p, q] \; \right\}$$

$\mathbf{R}^{\gamma} \stackrel{*}{\Rightarrow} a_{h+1} \ldots a_i \delta \nu \upsilon$ y:

$\delta \stackrel{*}{\Rightarrow} a_i \ldots a_p \; \mathbf{F}^{\gamma} \; a_{q+1} \ldots a_j \stackrel{*}{\Rightarrow} a_i \ldots a_j$ sii $(p, q) \neq (-, -)$

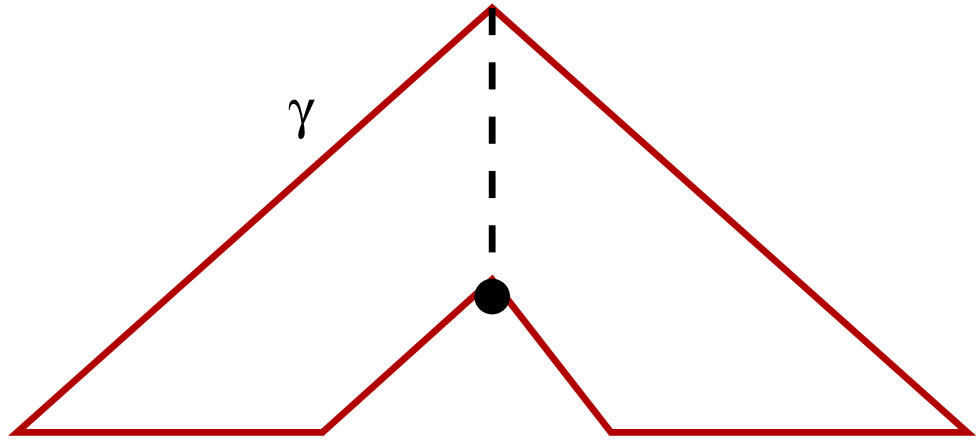
$\delta \stackrel{*}{\Rightarrow} a_i \ldots a_j$ sii $(p, q) = (-, -)$



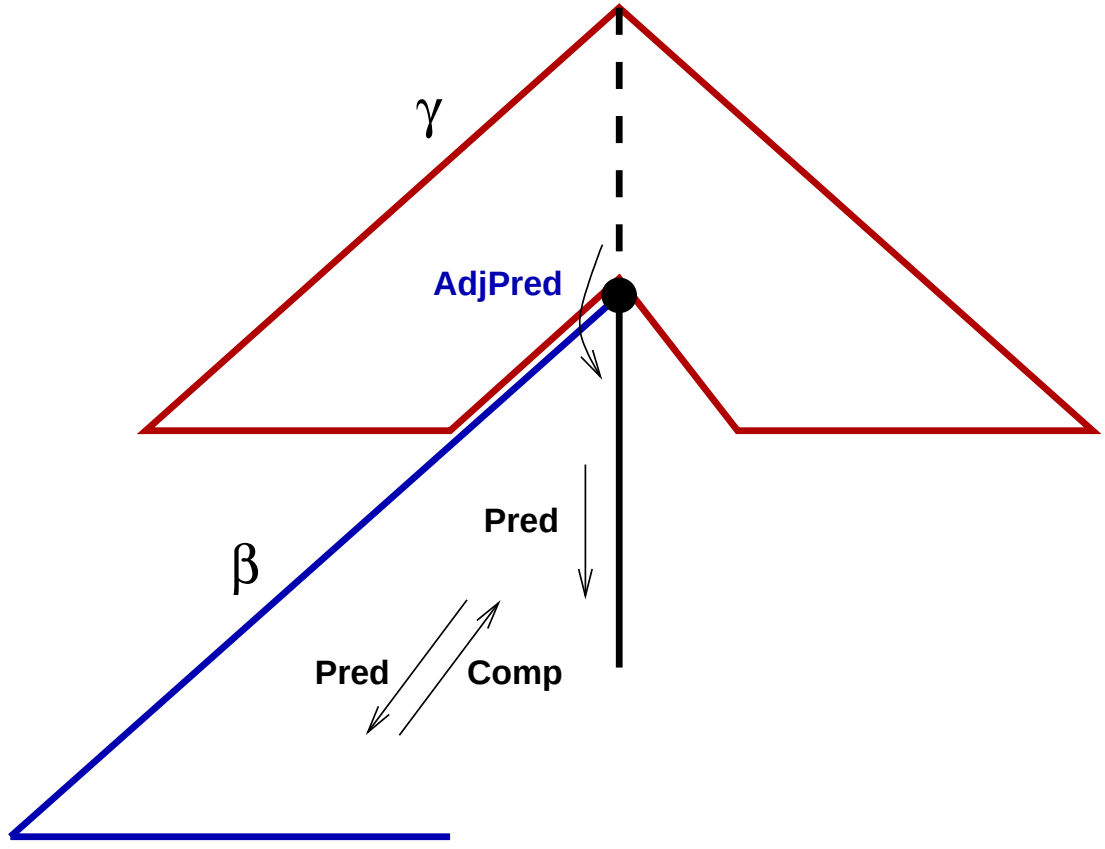
Análisis sintáctico de TAG

- Conceptos previos
- Gramáticas de adjunción de árboles
- Analizadores sintácticos para TAG
- Definición de los ítems
- **Definición de los pasos deductivos**
- Resultados experimentales

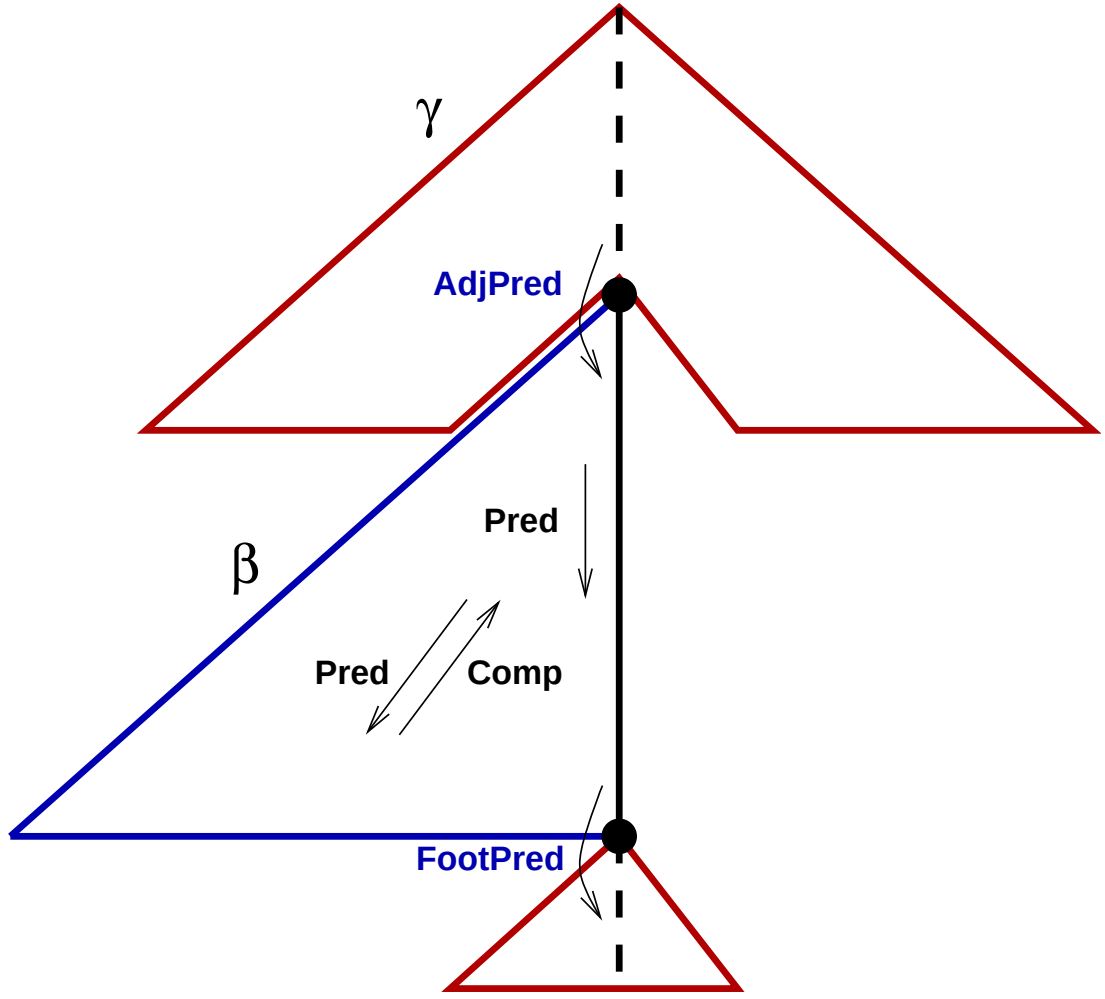
Pasos deductivos



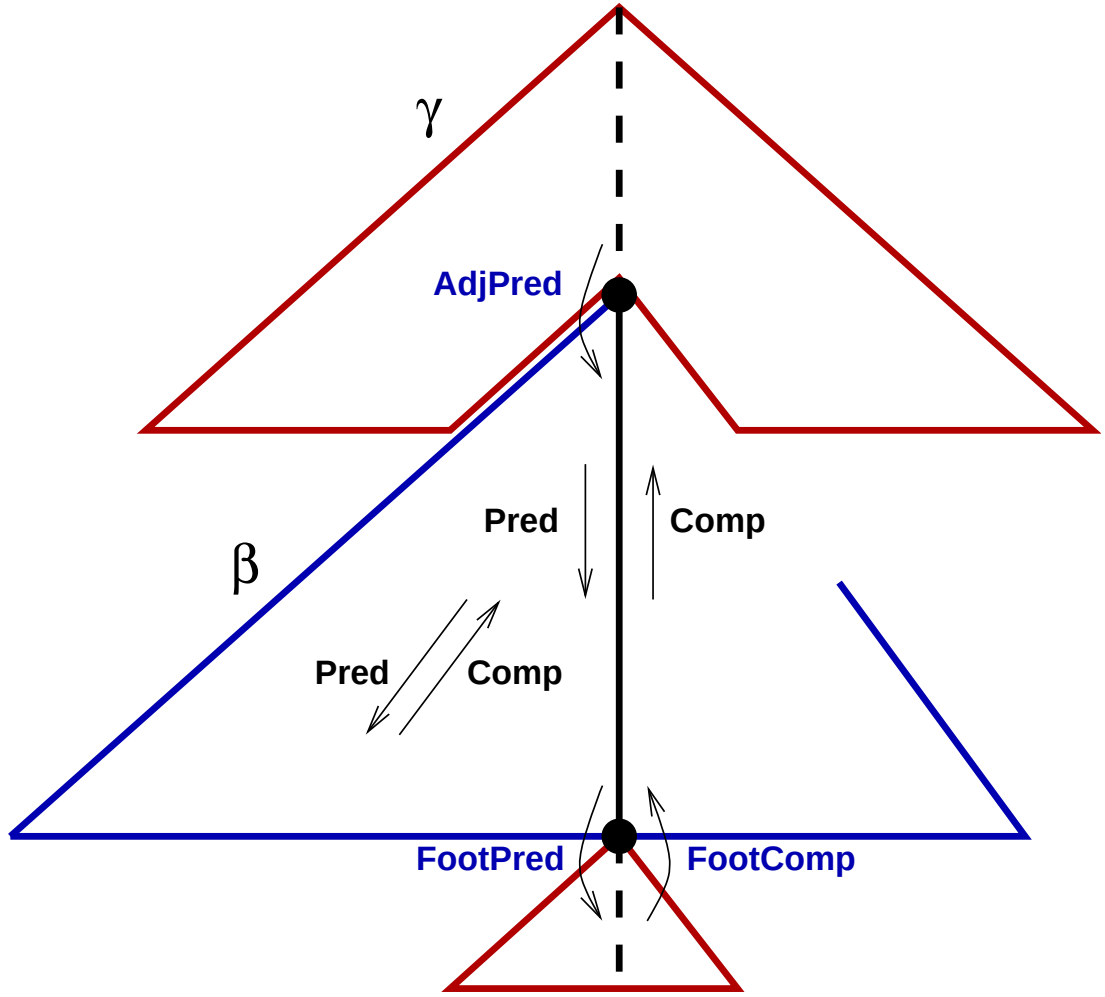
Pasos deductivos



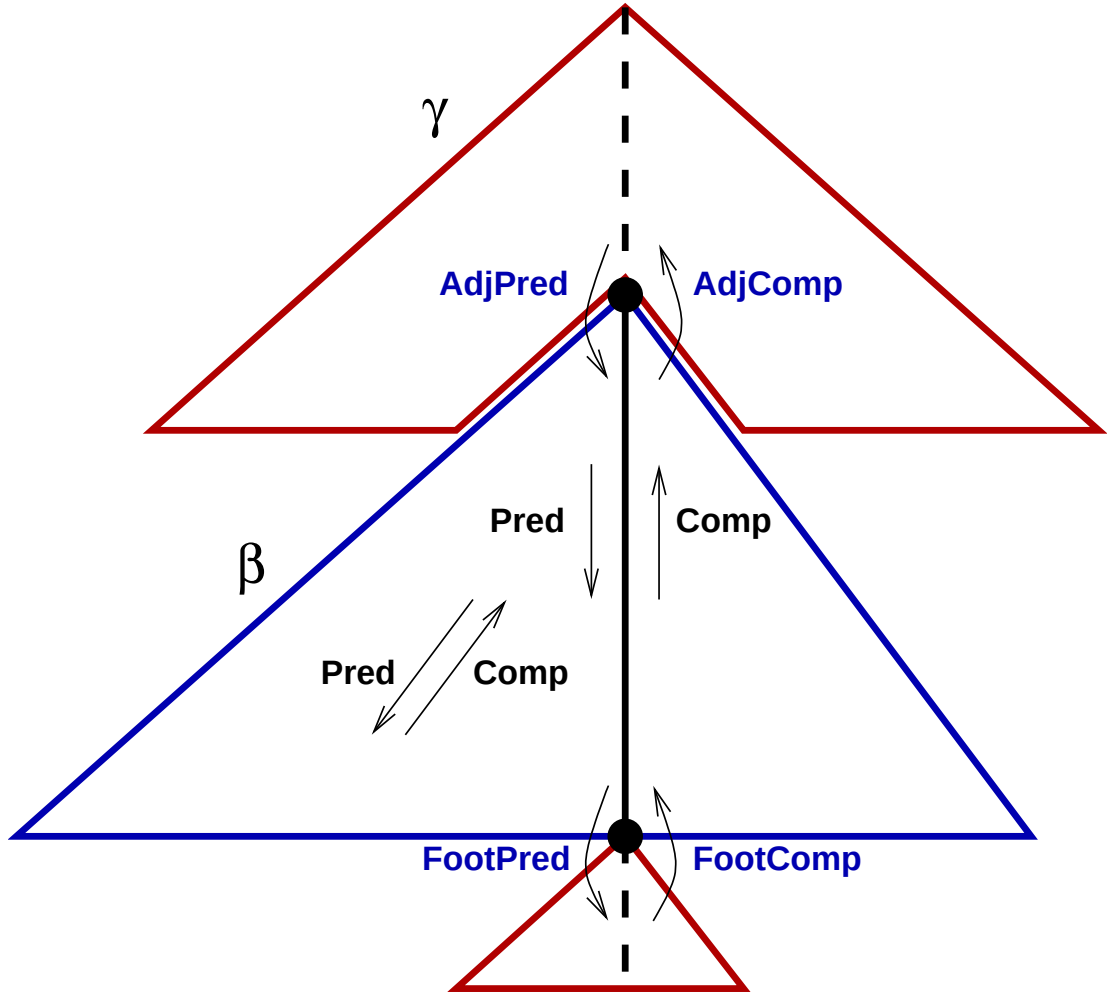
Pasos deductivos



Pasos deductivos

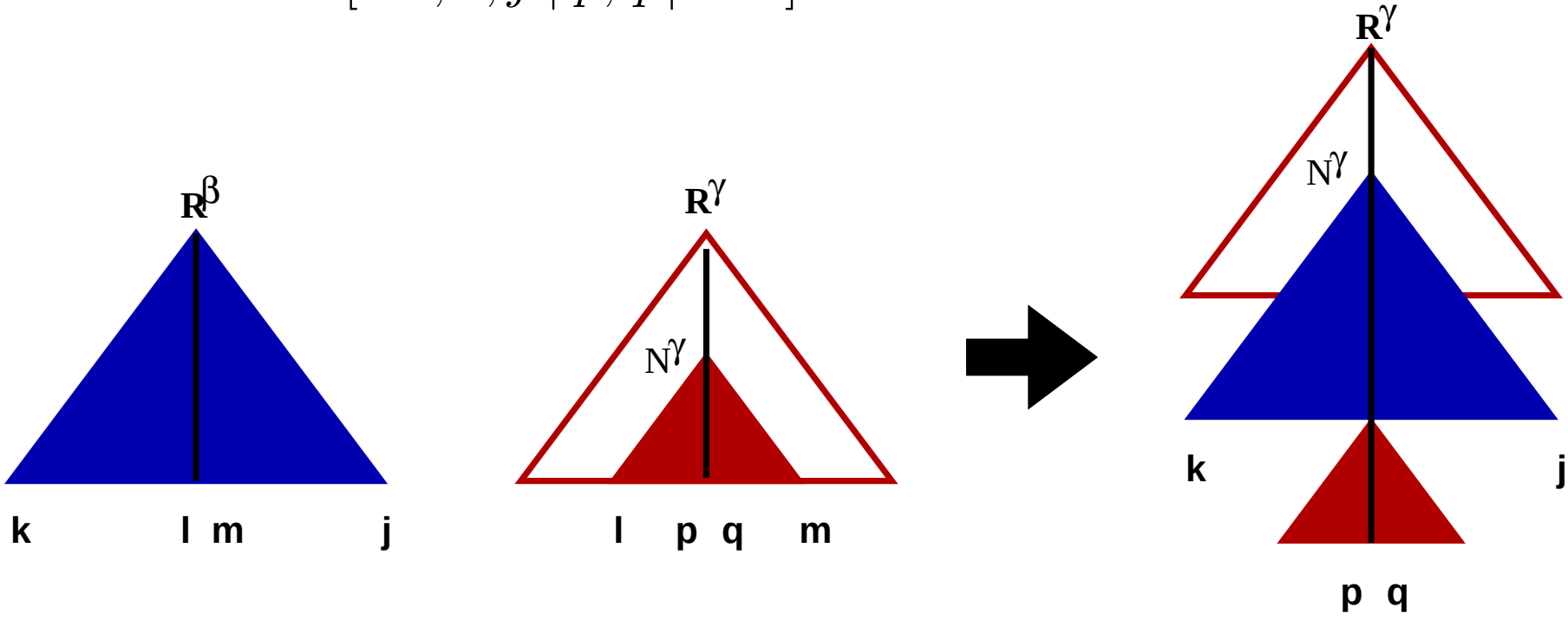


Pasos deductivos



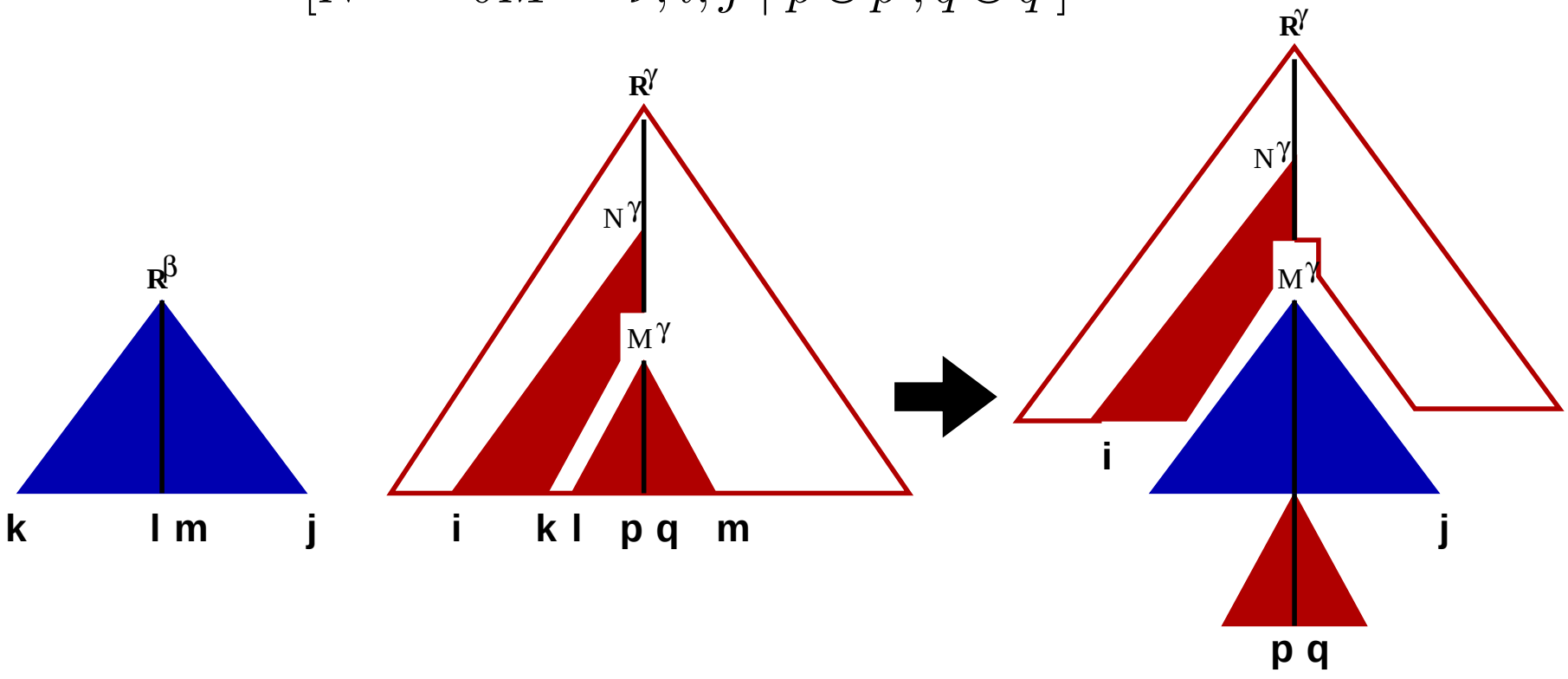
Paso AdjComp ascendente

$$\mathcal{D}^{\text{Adj}}_{\text{CYK}} = \frac{[\mathbf{R}^{\beta}, k, j \mid l, m \mid adj],\ [N^{\gamma}, l, m \mid p, q \mid \text{false}]}{[N^{\gamma}, k, j \mid p, q \mid \text{true}]} \quad \beta \in \boldsymbol{A}, \beta \in \text{adj}(N^{\gamma})$$



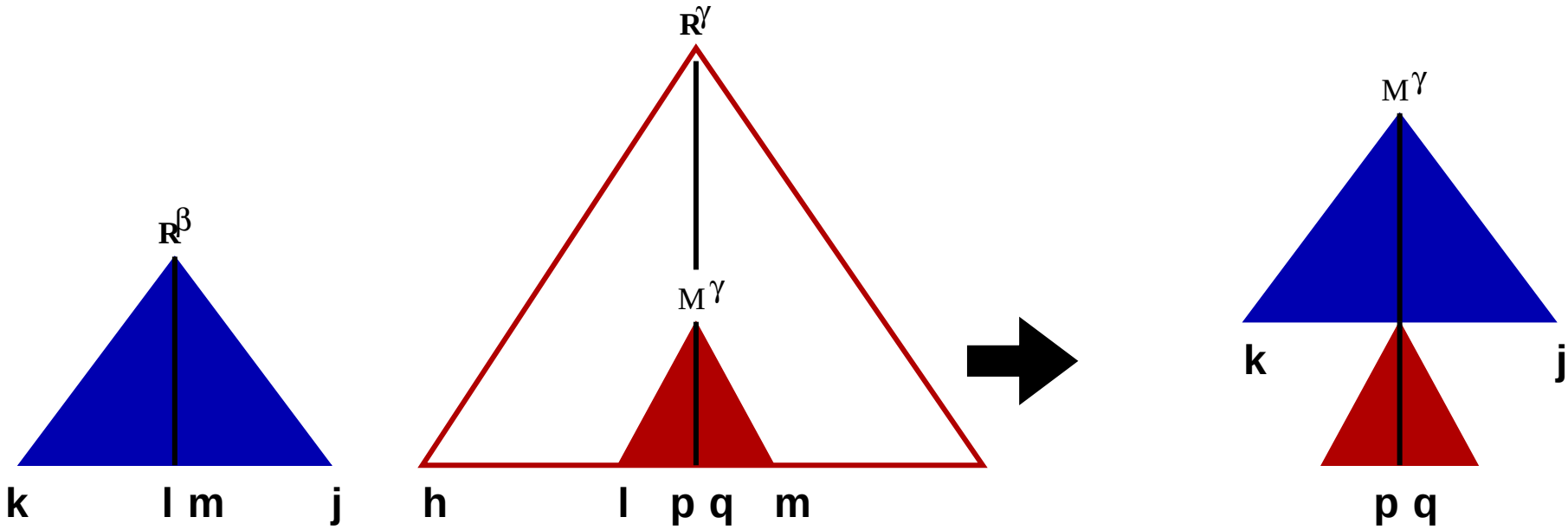
Paso AdjComp mixto

$$\mathcal{D}_{\mathrm{E}}^{\mathrm{AdjComp}} = \frac{\begin{array}{c}[\top \rightarrow \mathbf{R}^{\beta}\bullet, k, j \mid l, m],\\ [M^{\gamma} \rightarrow \upsilon\bullet, l, m \mid p, q],\\ [N^{\gamma} \rightarrow \delta \bullet M^{\gamma}\nu, i, k \mid p', q']\end{array}}{[N^{\gamma} \rightarrow \delta M^{\gamma} \bullet \nu, i, j \mid p \cup p', q \cup q']} \quad \beta \in \boldsymbol{A}, \beta \in \mathrm{adj}(M^{\gamma})$$



Paso AdjComp mixto con VPP

$$\mathcal{D}^{\text{AdjComp}^0}_{\text{Nederhof}} = \frac{[k, \top \rightarrow \mathbf{R}^{\beta}\bullet, k, j \mid l, m], \quad [h, M^{\gamma} \rightarrow \delta\bullet, l, m \mid p, q]}{[[M^{\gamma} \rightarrow \delta\bullet, k, j \mid p, q]]} \quad \beta \in \text{adj}(M^{\gamma})$$

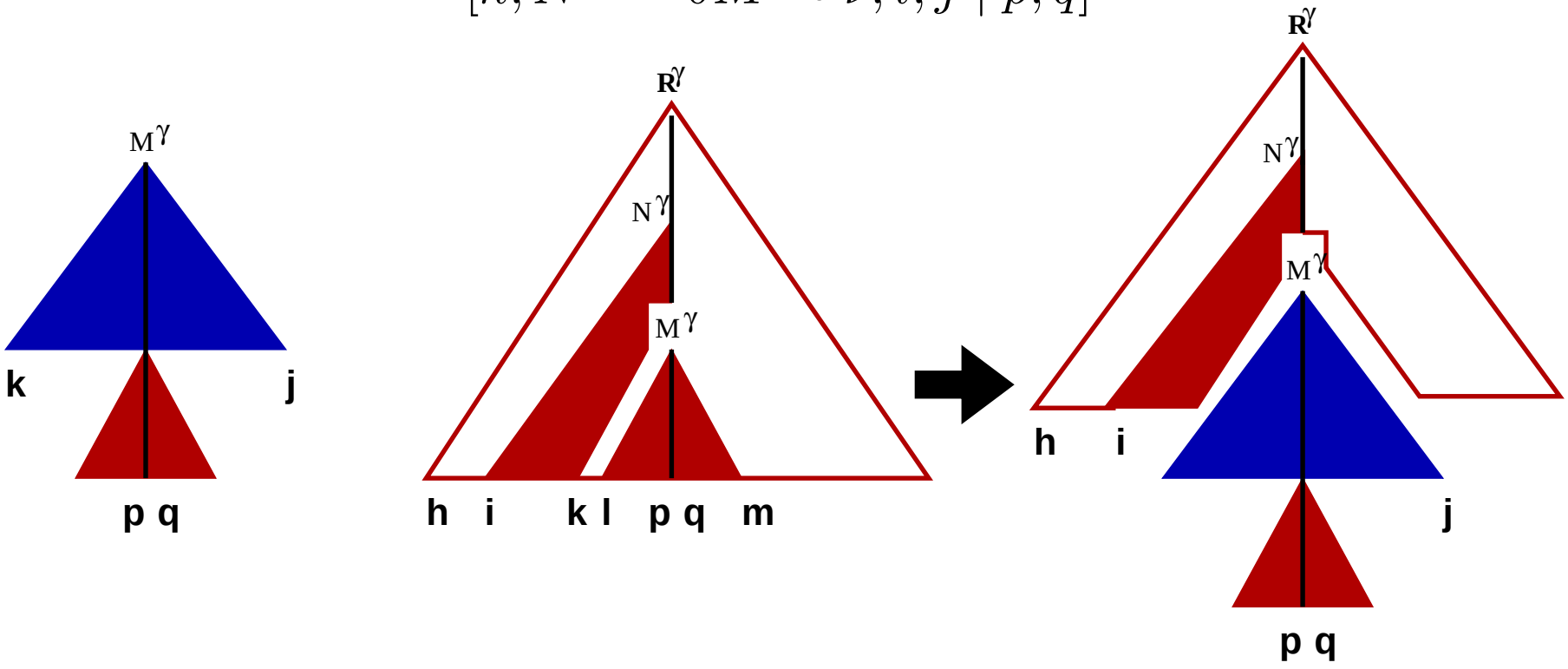


Paso AdjComp mixto con VPP

$$[[M^\gamma \rightarrow \delta\bullet, k, j \mid p, q]],$$

$$[h, M^\gamma \rightarrow \delta\bullet, l, m \mid p, q],$$

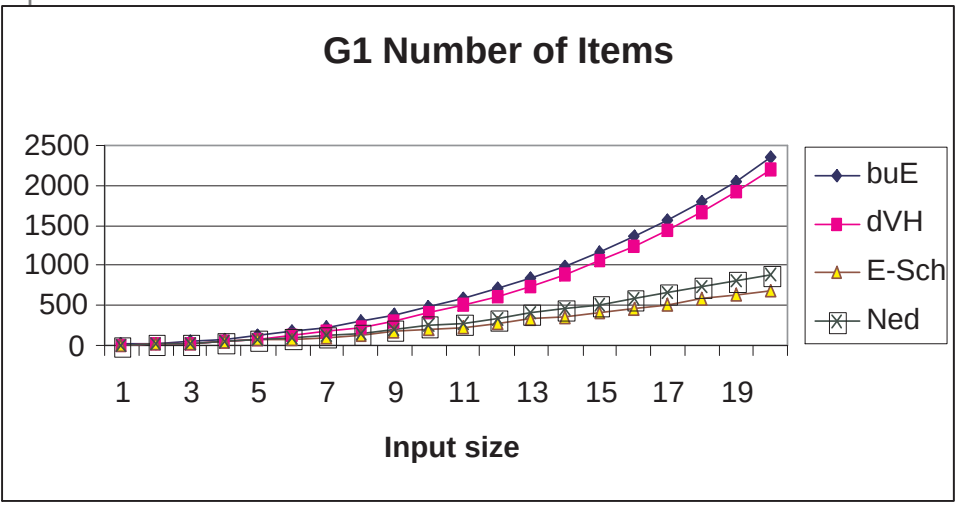
$$\mathcal{D}^{\text{AdjComp}^1}_{\text{Nederhof}} = \frac{[h, N^\gamma \rightarrow \delta \bullet M^\gamma \nu, i, k \mid -, -]}{[h, N^\gamma \rightarrow \delta M^\gamma \bullet \nu, i, j \mid p, q]} \quad \beta \in \text{adj}(M^\gamma)$$

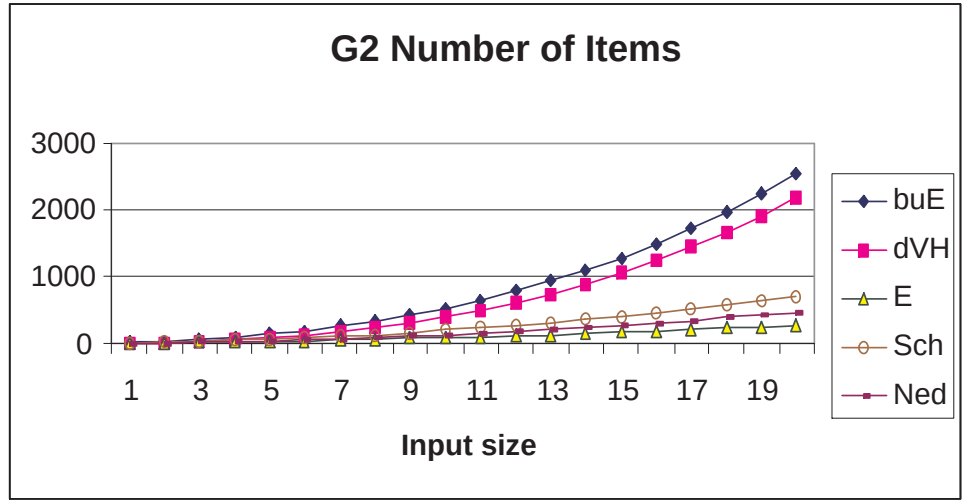


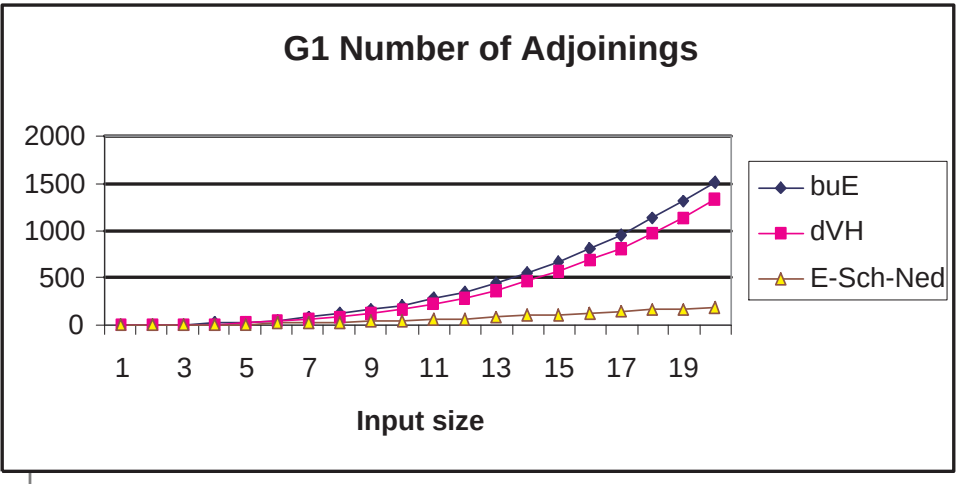
Análisis sintáctico de TAG

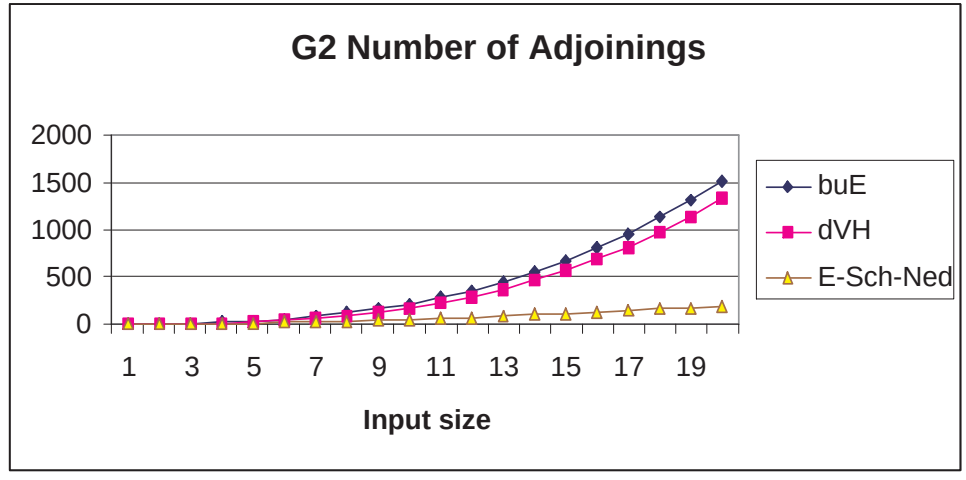
- Conceptos previos
- Gramáticas de adjunción de árboles
- Analizadores sintácticos para TAG
- Definición de los ítems
- Definición de los pasos deductivos
- **Resultados experimentales**

e^n con baja ambigüedad

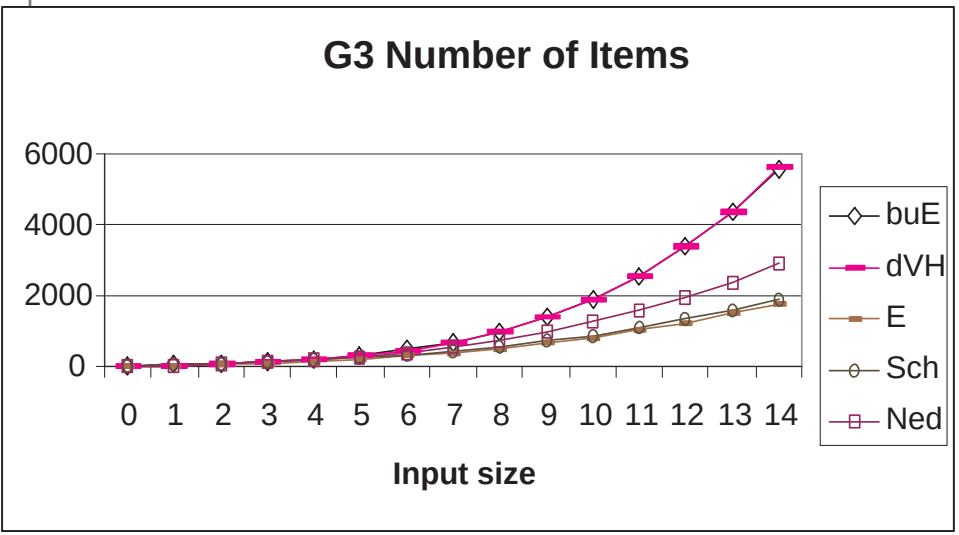


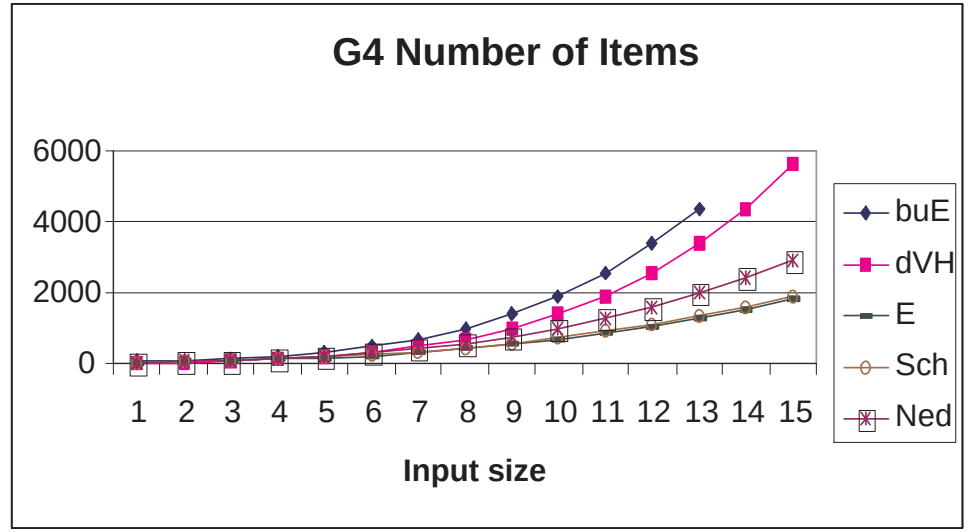


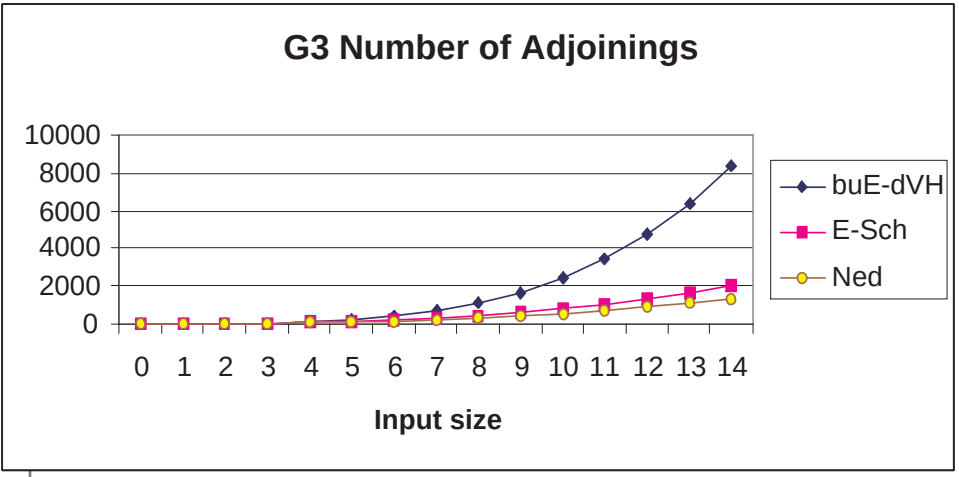


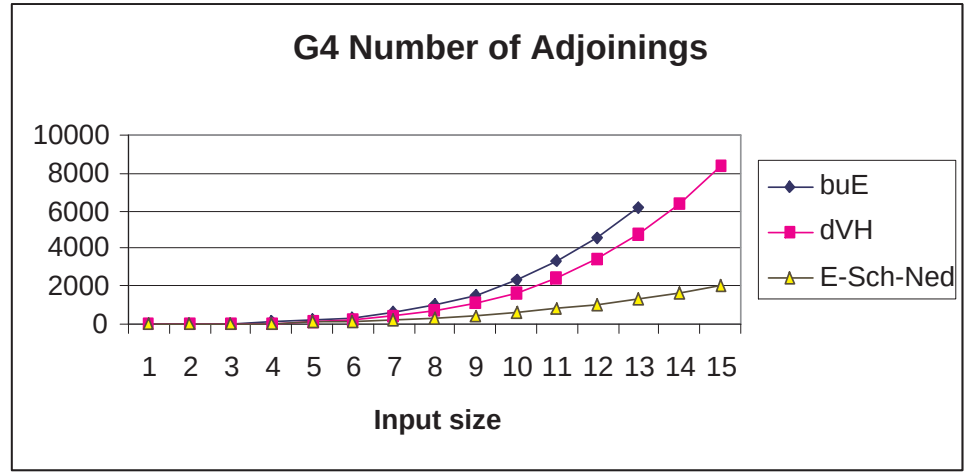


e^n con alta ambigüedad

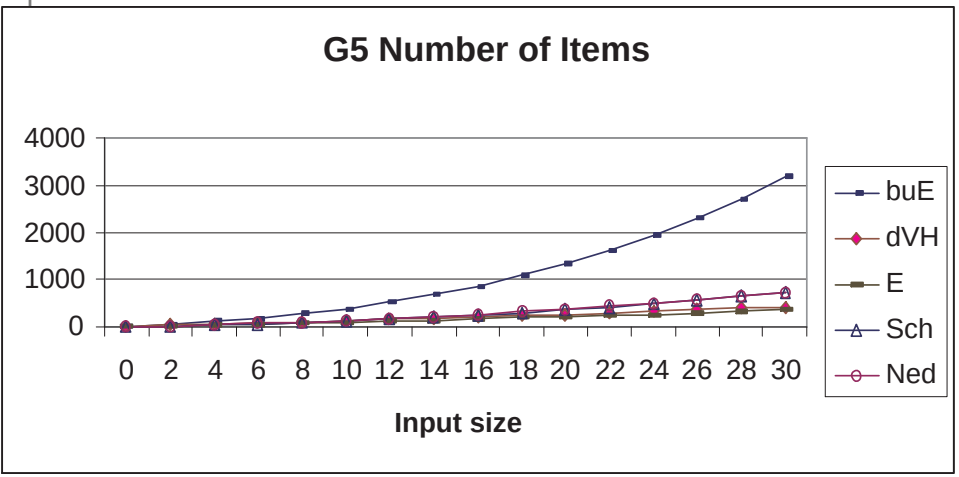


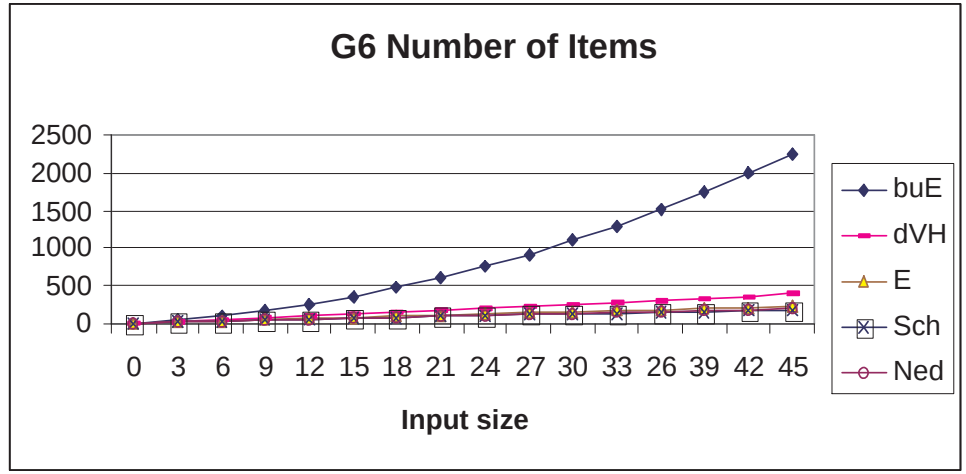


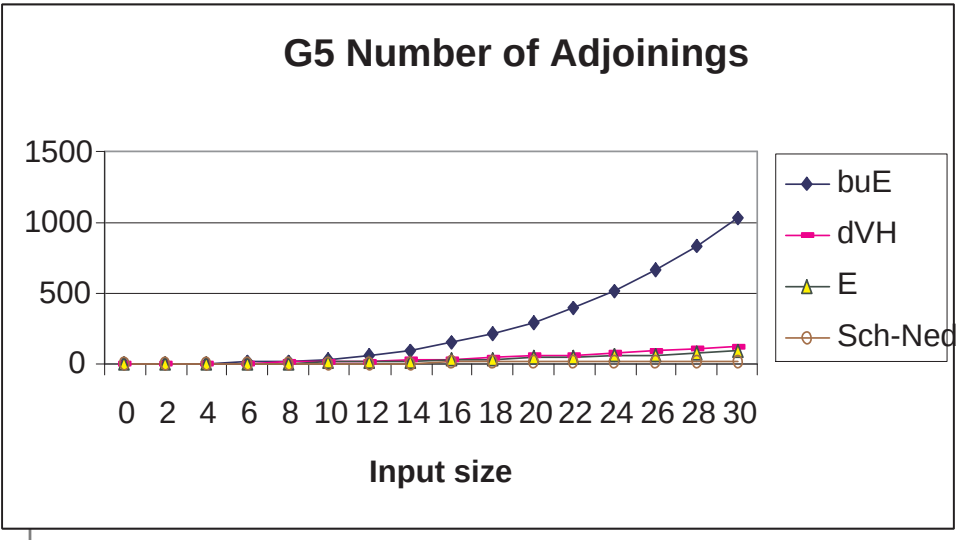


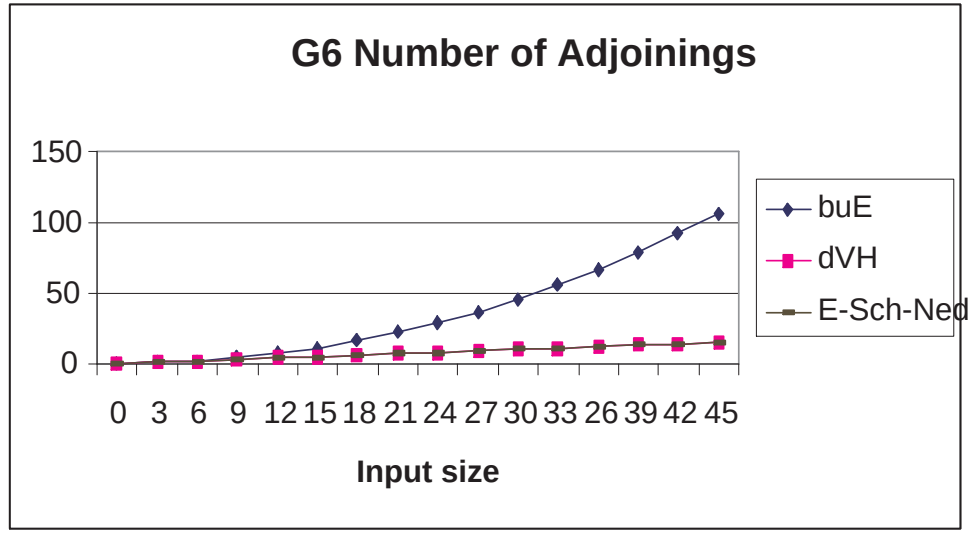


$a^n b^n$ y $a^n b^n c^n$

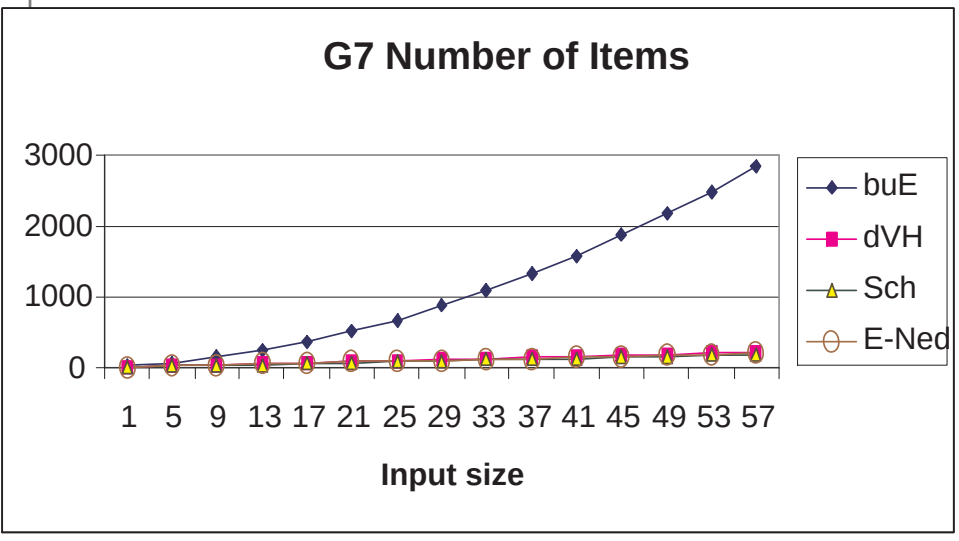


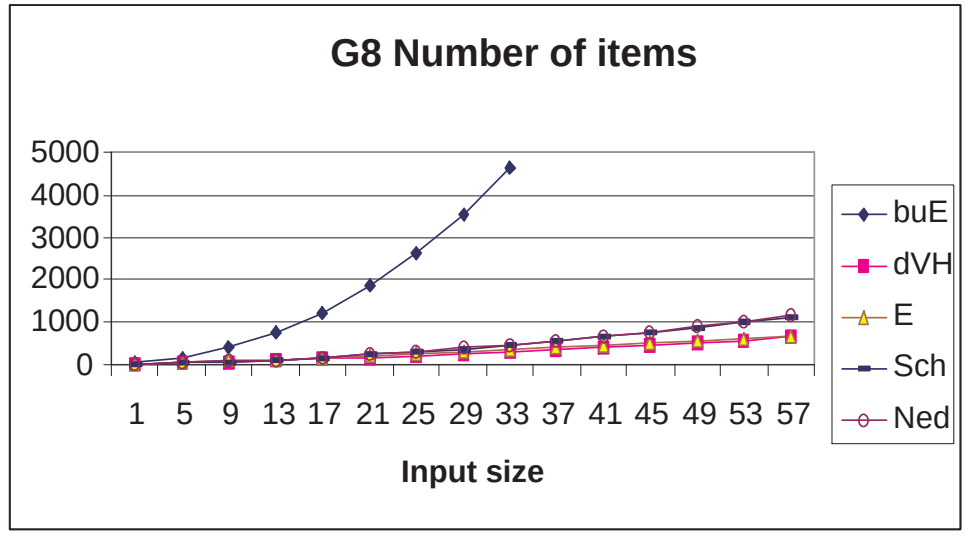


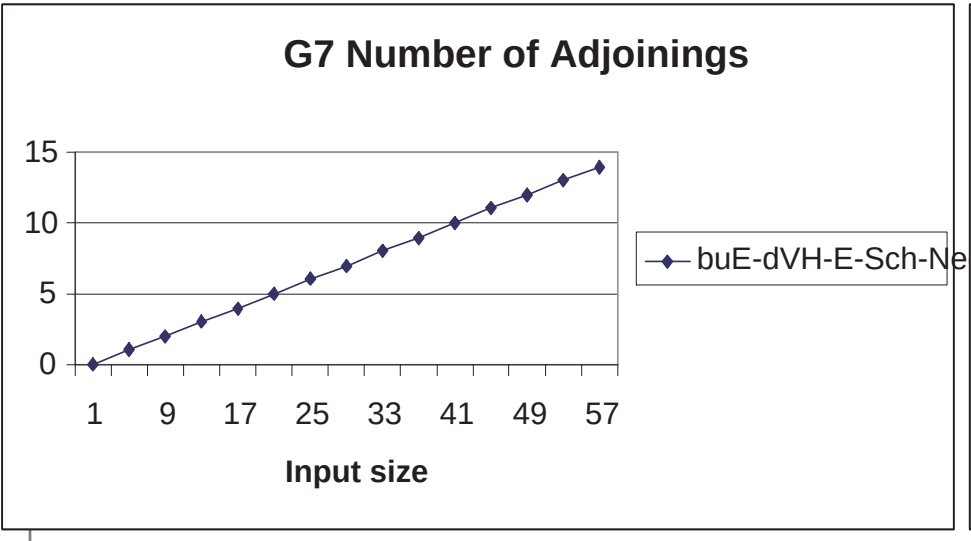


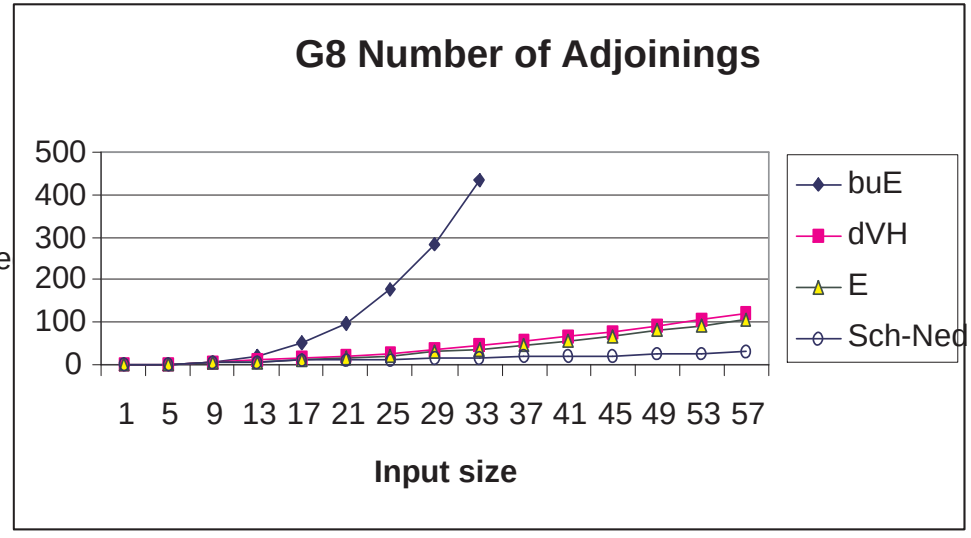


$a^n b^n e\, c^n d^n$ **y** $wcw, w \in \{a,b\}^*$









XTAG

	dVH	dVH'	E	Ned
Srini bought Beth a Book	0.27	0.05	0.38	0.44
Srini bought a book at the bookstore	0.38	0.11	0.49	0.55
Elmo borrowed a book	0.16	0.05	0.33	0.33
He hopes that murriel wins	0.22	0.05	0.66	0.49
The man who Muriel likes bought a book	0.60	0.16	0.77	0.88
The music should have been being played for the president	0.60	0.22	0.66	0.77
What did Clove catch?	0.16	0.05	0.38	0.33
Who did the elephant think the panda heard the emu said smell terrible?	0.82	0.16	1.54	1.26
Herbert is more livid and furious than angry	0.33	0.05	0.33	0.33

XTAG, con optimizaciones

	TAG	TAG+TIG	Reducción
Srini bought Beth a book	0.77	0.71	7.79 %
Srini bought a book at the bookstore	0.94	0.93	1.06 %
Elmo borrowed a book	0.55	0.49	10.91 %
he hopes that Muriel wins	1.26	1.16	7.94 %
the man who Muriel likes bought a book	2.14	1.48	30.84 %
the music should have been being played for the president	1.27	1.26	0.79 %
what did Clove catch?	0.60	0.55	8.33 %
who did the elephant think the panda heard the emu said smells terrible	3.13	2.36	24.60 %
Herbert is more livid and furious than angry	0.50	0.50	0.00 %

Análisis sintáctico de TAG

- Conceptos previos
- Gramáticas de adjunción de árboles
- Analizadores sintácticos para TAG
- Definición de los ítems
- Definición de los pasos deductivos
- Resultados experimentales

Volver al inicio