Anotación con EAGLES e introducción a sintaxis

Fernando Carranza
fernandocarranza86@gmail.com

Clase 10
Sábado 24/05/2025

- Unidad 1: Introducción
- Unidad 2: Los algoritmos supervisados
- Unidad 3: Anotación morfológica y de clase de palabra
  - a. Análisis morfológico
  - b. **Clases de palabras**
    - i) etiquetas de BNC y de Universal. Práctica de anotación.
    - **ii) etiquetamiento en CONLL-U de clase de palabra con EAGLES y de rasgos morfológicos.**
    - iii) Postag NLTK con tagsets BNC y Universal y Postag de Spacy y Stanza.
- Unidad 4: Anotación sintáctica
  - **i) Análisis basado en constituyentes y gramáticas basadas en dependencias.**
  - **ii) Parser basado en constituyentes BLLIP** y parsers basados en dependencias Spacy y Stanza.
  - **iii) Penn Treebank. Anotación sintáctica basada en constituyentes.**
  - iv) Análisis sintáctico basado en dependencias y su anotación en CONLL-U.
- Unidad 5: Anotación para propósitos específicos

Presentación

Estructura y temas de la clase de hoy:

1. Introducción
2. 3.b.ii) etiquetamiento en CONLL-U de clase de palabra con EAGLES y de rasgos morfológicos.
3. 4.i) Análisis basado en constituyentes y gramáticas basadas en dependencias.
   - Constituyentes
   - Gramáticas Dependencias
4. Recapitulación
5. Bibliografía

Las etiquetas del tagset de EAGLES (Leech y Wilson 1996) están pensadas para manifestar información de clase de palabra y de morfología flexiva y para poder ser aplicadas en, al menos, todas las lenguas de la comunidad europea.
Se las puede encontrar en la xpos de Ancora y en Freeling (Padró y Stanilovsky 2012, Padró *et al.* 2010, `https://nlp.lsi.upc.edu/freeling/node/1`, `https://github.com/TALP-UPC/freeling`), que incluye una serie de diccionarios de español que parea cada forma de palabra con su respectivo lema y etiqueta de clase de palabra (*pos-tag*). .

Algunos ejemplos de formas de palabra, lema y anotación morfosintáctica con el tagset de EAGLES para algunos nombres:

```
perro perro NCMS000
perros perro NCMP000
persa persa NCCS000
persas persa NCCP000
persecuciones persecución NCFP000
persecución persecución NCFS000
```

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
5-6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

https://www.cs.upc.edu/~nlp/tools/parole-sp.html

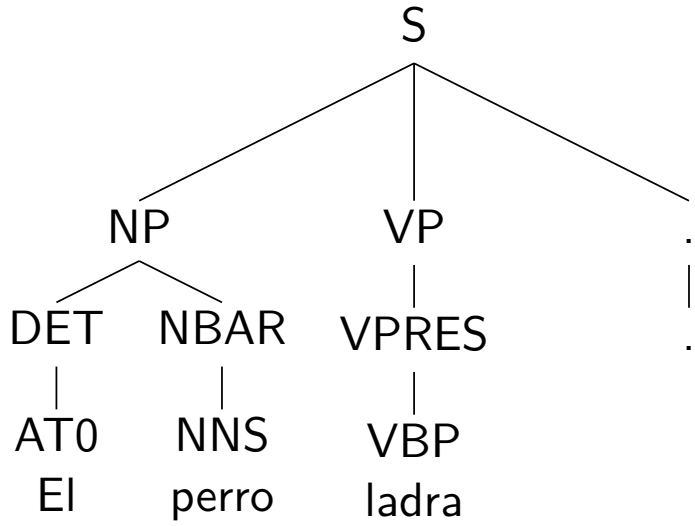
- Las gramáticas basadas en constituyentes asumen que la estructura oracional es pasible de ser analizada en términos de agrupamientos de palabras jerárquicamente organizados.
- Estos agrupamientos se denominan constituyentes.
- Los constituyentes se reconocen por una serie de pruebas que incluyen la sustitución por proformas, la interrogación, el movimiento y la coordinación.
- Estos constituyentes se etiquetan en función de su categoría y/o de su función sintáctica.
- La mayor parte de las teorías asumen que las funciones sintácticas se derivan de las relaciones estructurales y que, por lo tanto, no son primitivos de la descripción gramatical.

La estructura de constituyentes se puede representar mediante distintos recursos visuales. Entre otros, se utilizan los siguientes:

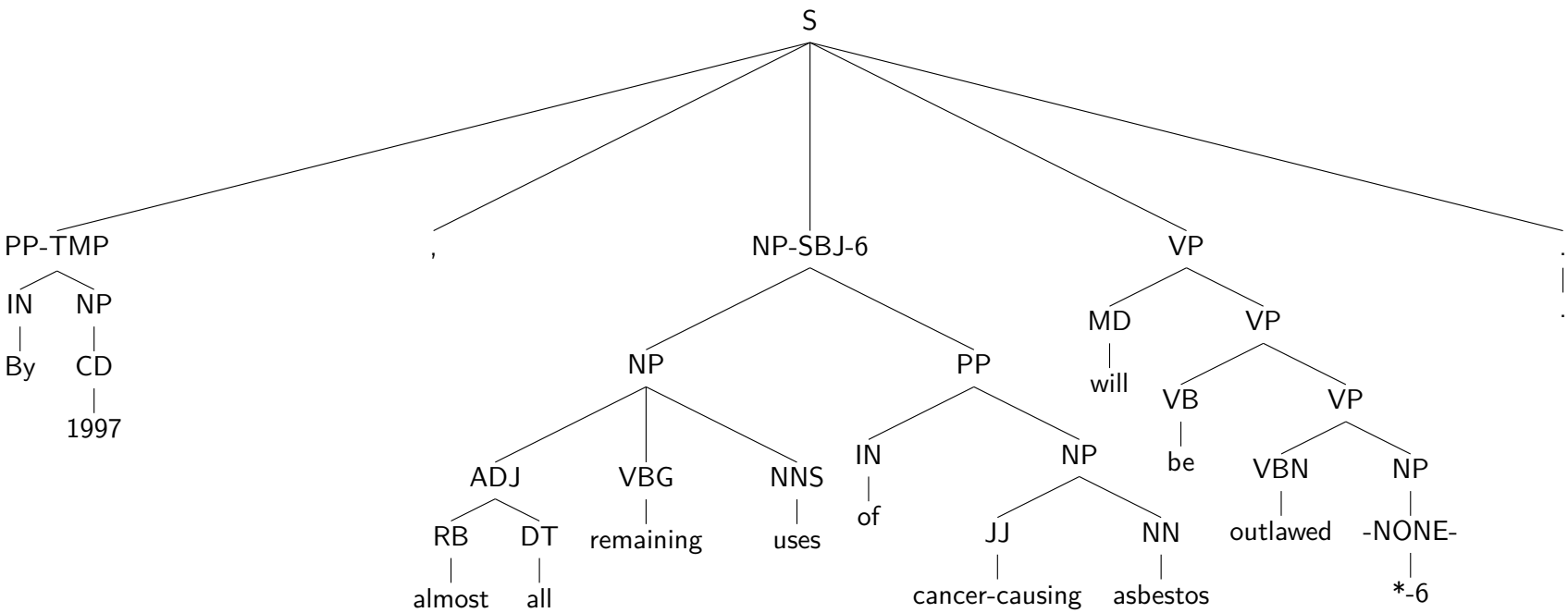
- Mediante corchetes o paréntesis.

```
(S
     (NP (DET "el/AT0")
      (NBAR "perro/NNS"))
     (VP (VPRES "ladra/VBP"))
     (? (FIN "./.")))
```

- Mediante árboles

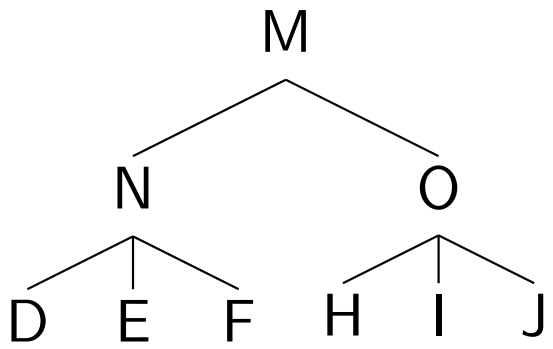


(1)


S
PP-TMP
,
NP-SBJ-6
VP
.
IN
NP
By
CD
1997
NP
PP
MD
VP
will
VB
VP
be
ADJ
VBG
NNS
IN
NP
VBN
NP
RB
DT
remaining
uses
of
JJ
NN
outlawed
-NONE-
almost
all
cancer-causing
asbestos
*-6

Representar los árboles en términos de grafos acíclicos dirigidos, además de proveer una visualización amigable, tiene propiedades matemáticas que hacen sencilla la formulación de axiomas y la posibilidad de evaluar si un parser funciona adecuadamente o no.

(2) a.

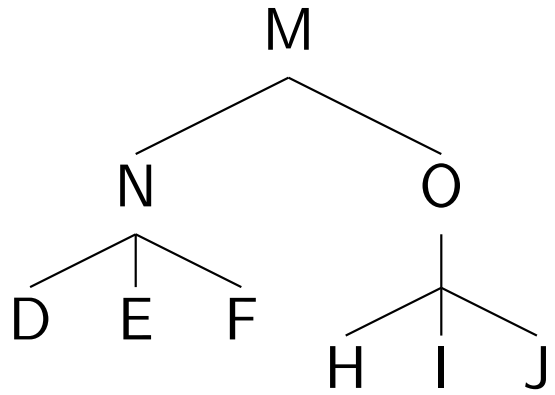


b. {<M,N>, <M,O>, <N,D>, <N,E>, <N,F>, <O,H>, <O,I>, <O,J>}.

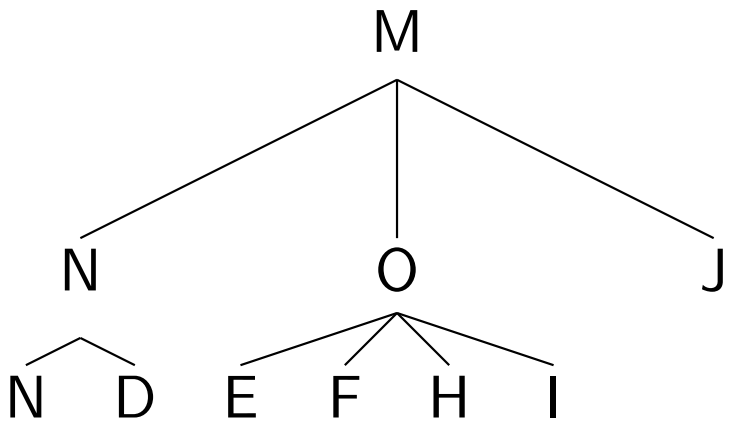
c. {(M,0,6), (N,0,3), (D,0,1), (E,1,2), (F,2,3), (O,3,6), (H,3,4), (I,4,5), (J,5,6)}

Métrica PARSEVAL (Black *et al.* 1991) para determinar la precisión de un parser:

(3) a. **Resultado esperado:**



b. **Resultado obtenido:**



- Se compara el output del parser con el parseo esperado en términos de listas de categorías y spans:

(4) a. Resultado esperado: {(M,0,6), (N,0,3), (D,0,1), (E,1,2), (F,2,3), (O,3,6), (H,3,4), (I,4,5), (J,5,6)}

b. Resultado obtenido: {(M,0,6), (N,0,2), (D,0,1), (E,1,2), (O,2,5), (F,2,3), (H,3,4), (I,4,5), (J,5,6)}

- Se cuenta la cantidad de coincidencias entre el resultado obtenido y el esperado y se la divide por la cantidad de spans que contiene el resultado obtenido:

(5) Precisión: $\frac{7}{9}$

Axiomas de los árboles sintácticos (ver Wall 1972, Partee *et al.* 2012, Carnie 2010):

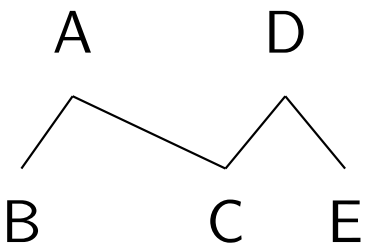
Axiomas de dominancia

- **A1. Reflexividad**: Todo nodo se domina a sí mismo.

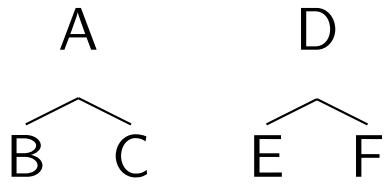
- **A2 Single root condition**: Solo puede haber un nodo inicial.

Este axioma excluye árboles como los siguientes:

(6) a.



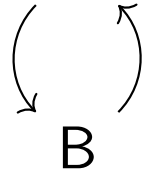
b.



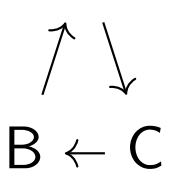
- **A3 Transitividad:** Si un nodo x domina al nodo y e y domina al nodo z, x domina a z.
- **A4. Antisimetría:** Dos nodos no pueden dominarse mutuamente.

Estos dos axiomas excluyen árboles como los siguientes:

(7) a.



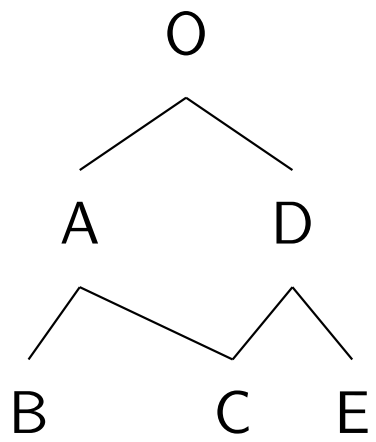
b.



- **A5. No multidominancia.** Un nodo no puede estar inmediatamente dominado por más de un nodo.

Este axioma excluye árboles como el siguiente:

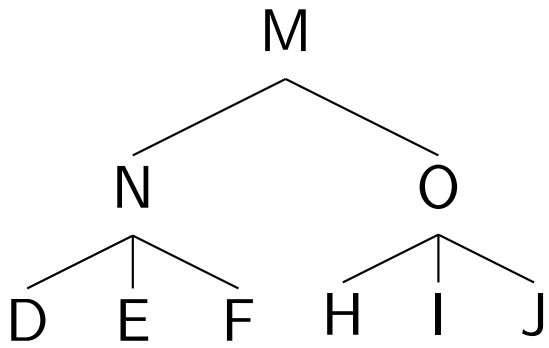
(8)



Los nodos se clasifican en los siguientes tipos:

- **Raíz**: nodo solo dominado por sí mismo.
- **Nodos no terminales**: Nodos que dominan al menos a otro nodo, además de a sí mismos.
- **Nodos intermedios**: Nodos no terminales que no sean solo dominados por sí mismos.
- **Nodos terminales**: Nodos que no dominan a ningún otro nodo, más allá de a sí mismos.

(1)



¿Cuál es el nodo raíz?

¿Cuáles son los nodos intermedios?

¿Cuáles son los nodos terminales?

Axiomas de precedencia

- **A6. Transitividad**: La precedencia es transitiva.
- **A7. Asimetría**: La precedencia es asimétrica.
- **T1. Irreflexividad**: Un nodo no puede precederse ni seguirse a sí mismo.

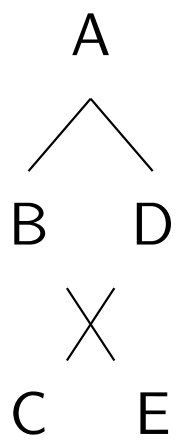
- **A8. Condición de exclusividad**: Si x precede a y, x no puede dominar a y.

La condición de exclusividad es la que le da a los símbolos no terminales su carácter abstracto y metalingüístico, es decir, la que los excluye de la cadena final.

- **A9. Non-tangling condition**: Si x precede a y, x domina a w e y domina a z, entonces w precede a z.

Este axioma impide que se dé una configuración como esta:

(9)



¿A cuál de los axiomas vistos anteriormente violentan los siguientes ejemplos del español?

(10) **Topicalización**

a. Romina conoció a Matías en la facultad.

b. A Matías, Romina lo conoció en la facultad.

c. En la facultad, Romina conoció a Matías.

- Para dar cuenta de estas violaciones a la *non-tangling condition*, las teorías lingüísticas formales usan dos grandes recursos: categorías vacías (e.g., HPSG) o transformaciones (.e.g, Gramática generativa), dependiendo de si se trata de enfoques representacionales o derivacionales.
- Como los Treebanks son por naturaleza representaciones, utilizan mayormente categorías vacías:

*	"Understood" subject of infinitive or imperative
0	Zero variant of that in subordinate clauses
T	Trace–marks position where moved wh-constituent is interpreted
NIL	Marks position where preposition is interpreted in pied-piping contexts

Cuadro: Categorías vacías en el Penn Treebank (Marcus *et al.* 1993)

A cuál de los axiomas vistos anteriormente violenta el recurso retórico de la digresión mediante una parentética oracional:

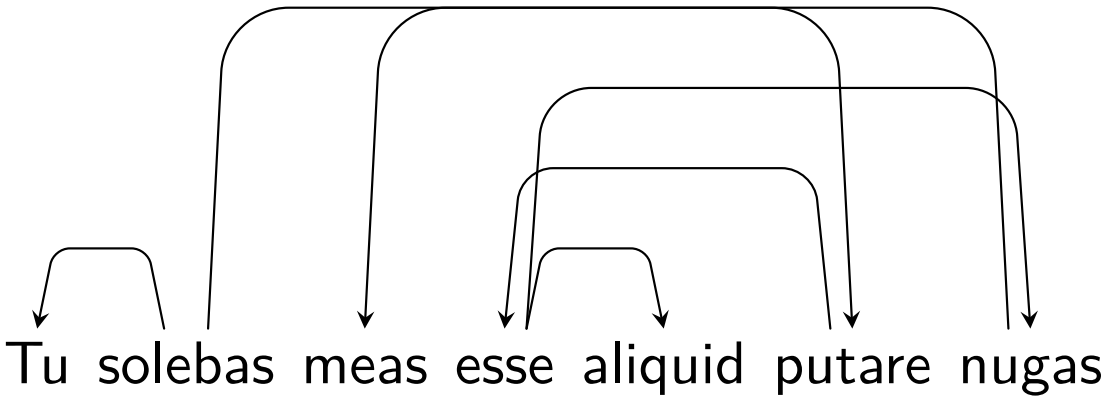
(11) El arte (Benjamin lo anotó muchas veces a propósito de los surrealistas) tiene una capacidad muy intensa de producir estos encuentros inesperados entre sentidos diferentes. (Sarlo 2007: 38, apud Müller 2023)

La teoría sintáctica en general no sabe muy bien cómo tratar este tipo de parentéticas. Tampoco en los sistemas de anotación se suele problematizar. Por ejemplo, en Universal Dependencies, la función más semejante es la de appos.

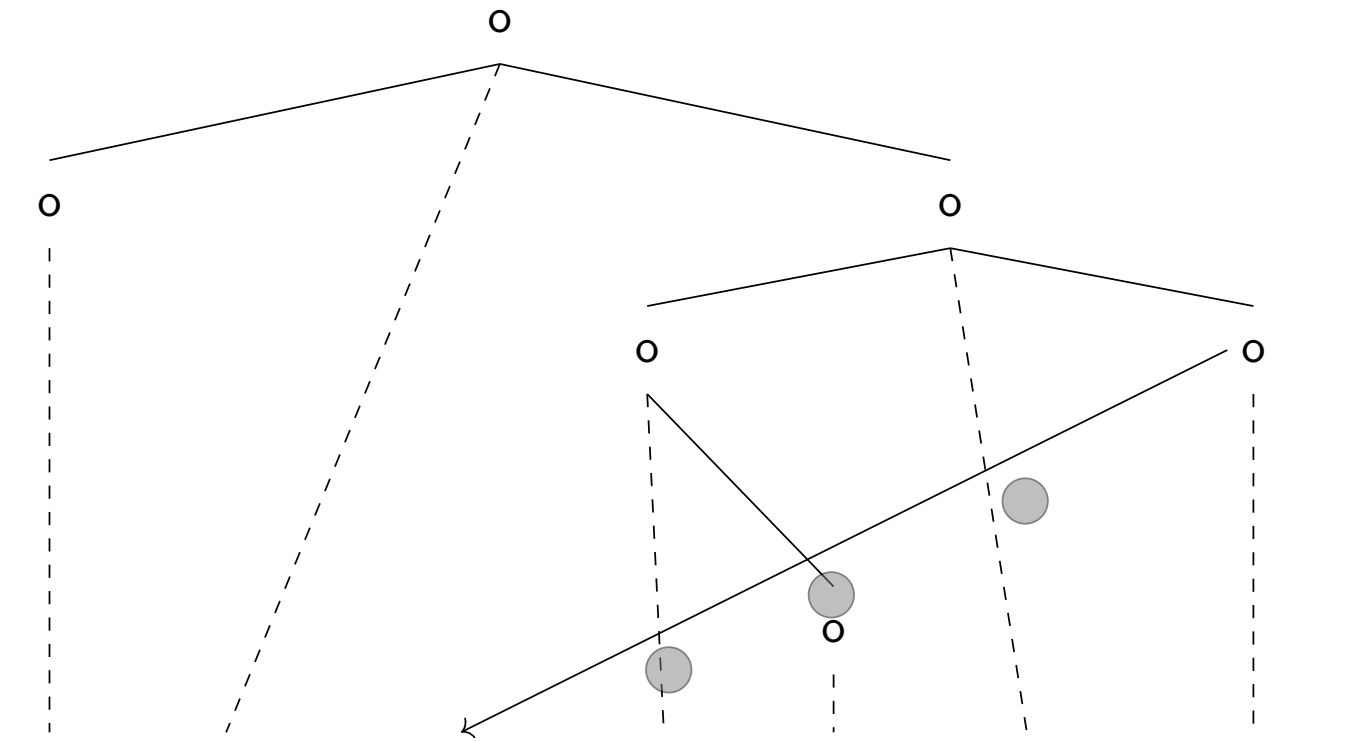
- Las gramáticas de dependencias tienen una larga historia en lingüística que se remonta a la obra póstuma de Tesnière (1959) (ver Carranza 2016 o Müller 2016 para una revisión histórica).
- En la tradición gramatical las nociones más influyentes de Tesniére fueron la de valencia argumental y la de traspositor.
- En la década del 2000 empezaron a aparecer varios recursos computacionales que usan el formalismo de las gramáticas de dependencias: SyntaxNet (Google), el Malt Parser (Nivre 2003), los parsers de Spacy y StanfordNLP/Stanza, Freeling, TXALE (Atserias Batalla *et al.* 2005), entre otros.

- La idea básica de las gramáticas de dependencias es que las palabras se vinculan entre sí no a partir de agrupaciones jerárquicas denominadas constituyentes, sino a partir de dependencias.

(2)



- Una ventaja de este enfoque es que es mucho más flexible para dar cuenta de las lenguas de orden libre de palabras.



Tu	*solebas*	*meas*	*esse*	*aliquid*	*putare*	*nugas*
tú	soler	mi	ser	algo	pensar	bagatela
nom	2sg	fem.pl.acc	inf	nom	inf	fem.pl.acc

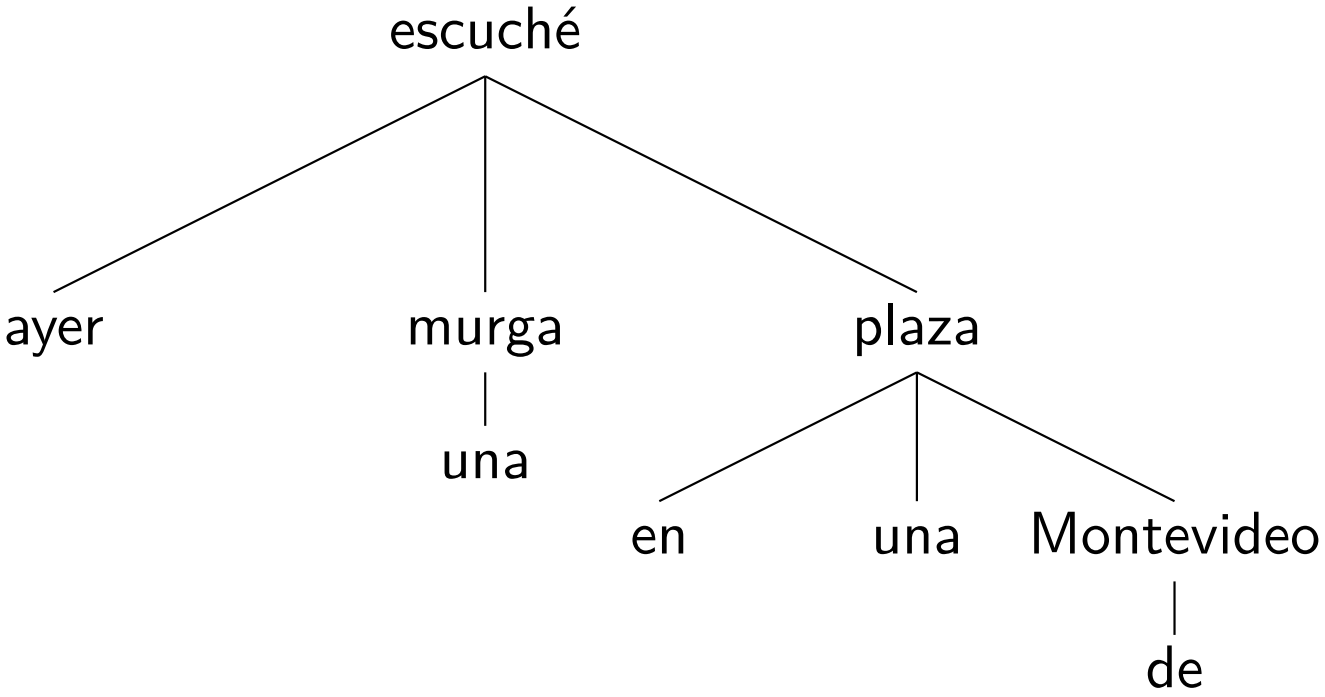
"Tú solías pensar que mis bagatelas eran algo."

- Otra ventaja es que el concepto de dependencia permite formalizar de manera más directa la noción de núcleo, algo que las gramáticas basadas en constituyentes tienen que agregar *ad hoc*.

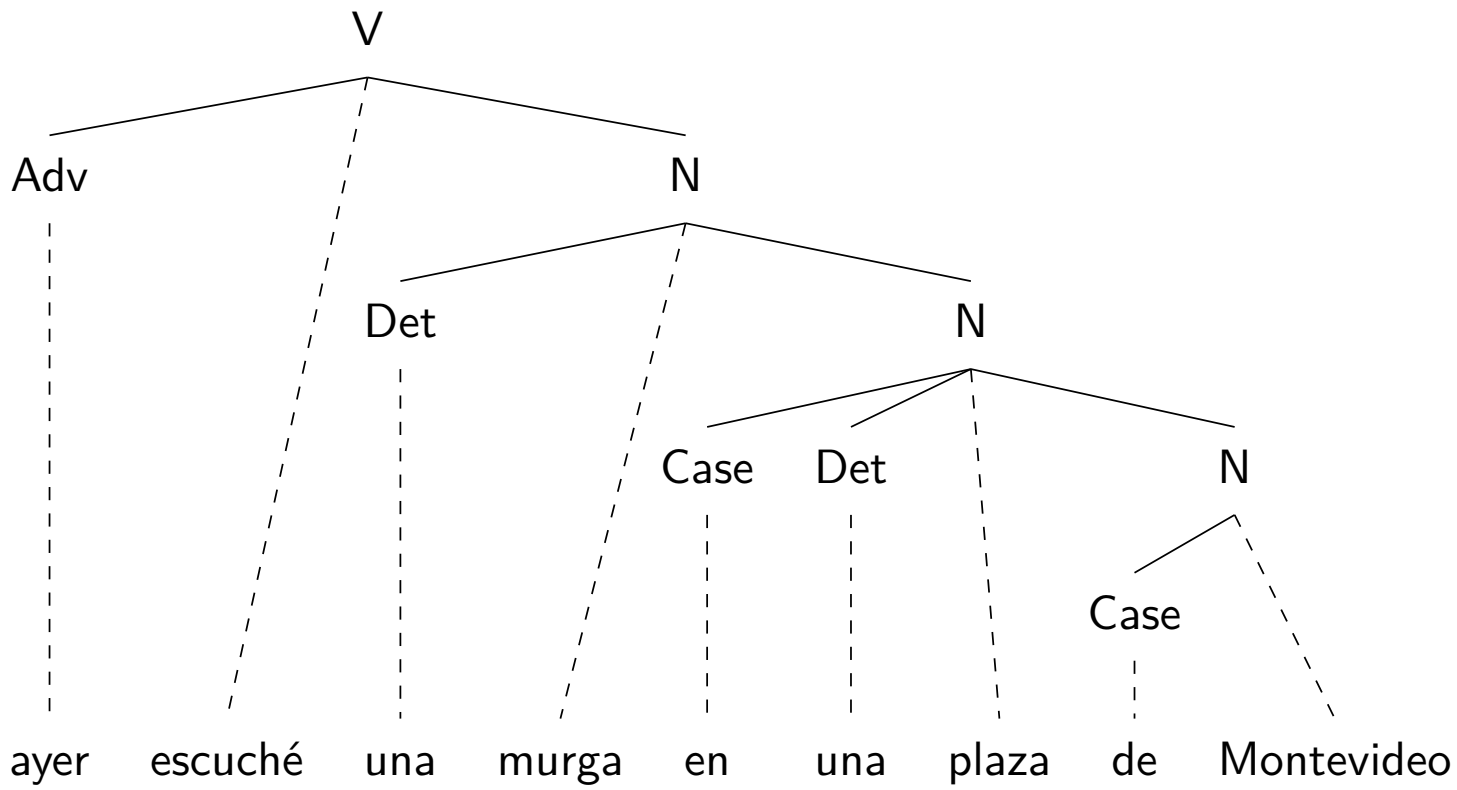
La representación arbórea del análisis sintáctico en el formato de las gramáticas de dependencias se conoce con el nombre de Stemma. Existen distintos modos de visualizar los stemmas:

- Stemma al estilo de Tesnière (1959)

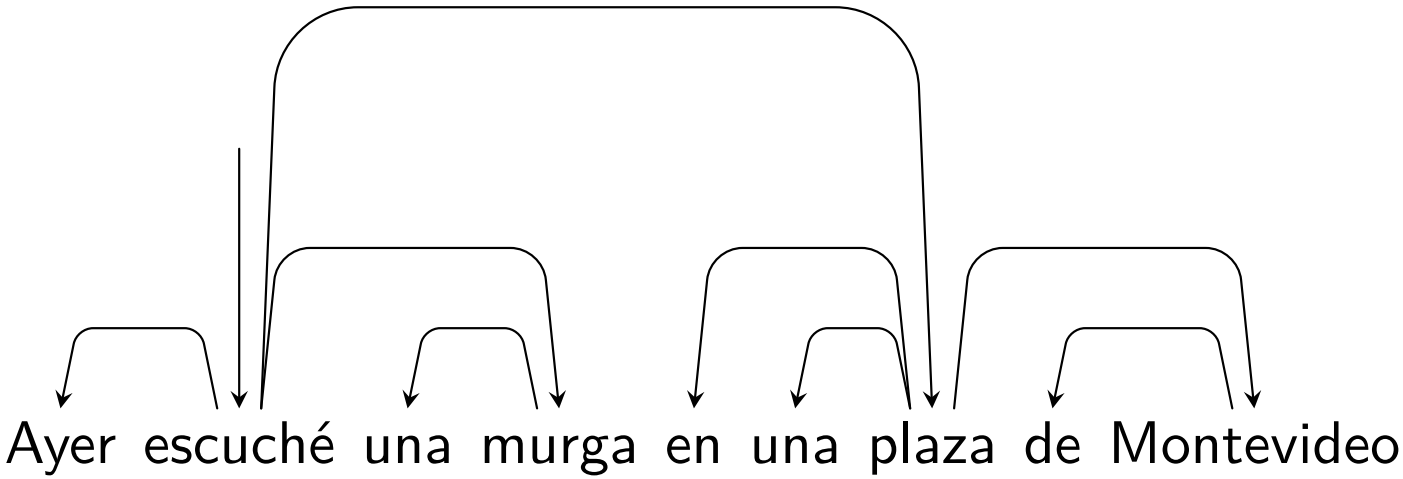
(3)



- Stemma al estilo de Hays (1964) y Gaifman (1965):



- Stemma al estilo de la Word Grammar (Hudson 1984, 2010)



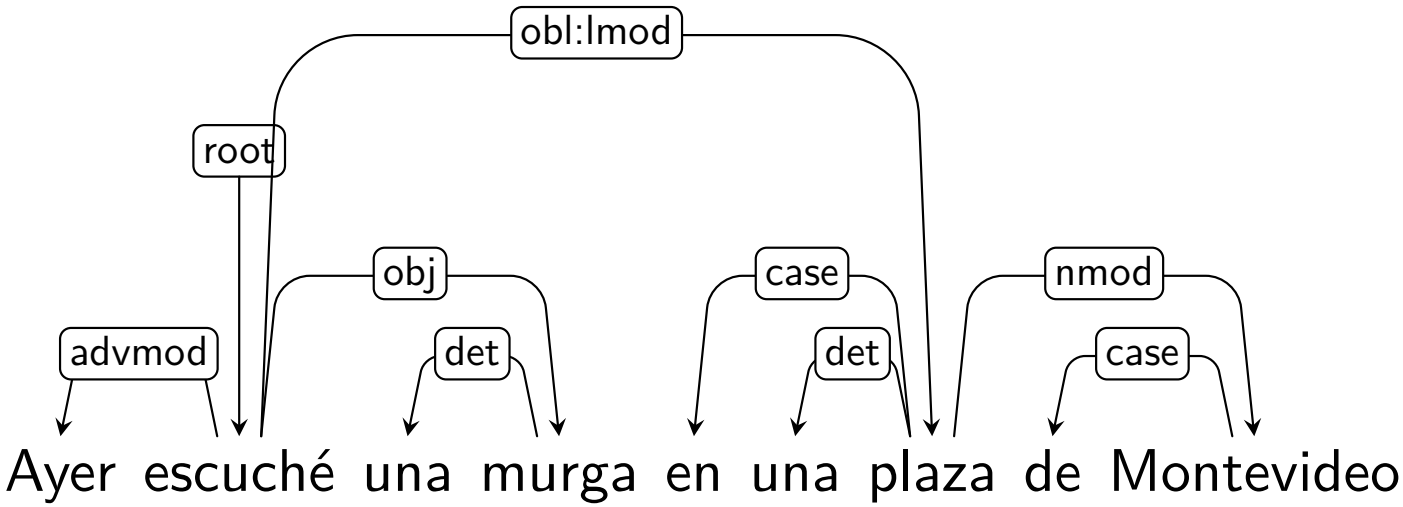
Es el estilo más usual.

- Las dependencias se tipifican en función de una ontología de funciones. Lo más común es que se trate de funciones sintácticas.

Mel'čuk (2011) propone una extensión de la gramática de dependencias en tres estratos:

- **Dependencias semánticas**: Encargadas de establecer relaciones entre las unidades mínimas del estrato semántico (semantemas) y corresponderían a una relación de predicado–argumento, donde el regidor es el predicado.
- **Dependencias sintácticas**: Establecen las relaciones entre la red representada por el estrato semántico y la cadena de morfemas del estrato morfológico en la forma de un árbol de dependencias.
- **Dependencias morfológicas**: Se corresponden con las relaciones de *Government* y *Agreement*.

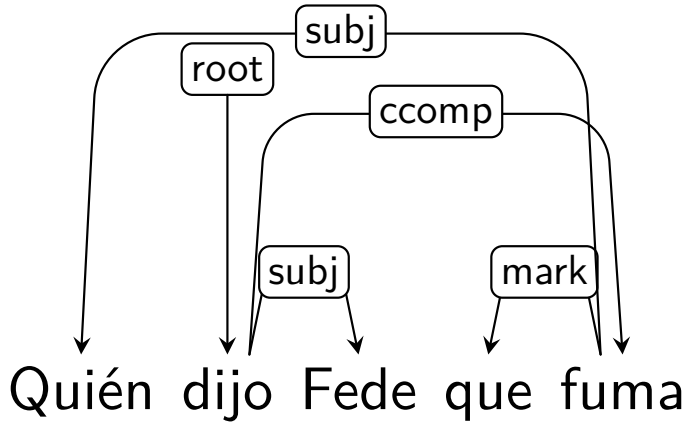
En esta unidad nos vamos a limitar tan solo a las dependencias sintácticas (e.g., sujeto, objeto, etc.)



Robinson (1970) formuló cuatro axiomas que rigen la formación de las estructuras de dependencias.

- **Raíz única**: Existe un único elemento que no depende de ningún otro elemento. Ese elemento es la raíz (*root*). Suele representarse por una flecha que no tiene nodo de procedencia.
- **Conectividad**: Todos los elementos de la oración dependen directamente de otro elemento. No habrá elementos intermedios en las estructuras de dependencias, por lo que únicamente habrá tantos nodos como palabras en el grafo correspondiente.
- **No multidominancia**: Ningún elemento depende de más de un elemento. Un regidor podrá regir varios dependientes, pero cada dependiente puede tener únicamente un regidor.
- **Proyectividad**: Si un elemento A depende directamente de un elemento B y un elemento C interviene en el orden lineal entre ellos, ese elemento C depende directamente de A o de B. La proyectividad expresa una condición similar a la *non-tangling condition*.

Existen oraciones en las lenguas naturales que no cumplen con la proyectividad:



El problema es que la proyectividad acarrea un gran costo de procesamiento. Por esta razón, para dar cuenta de la proyectividad muchos parsers tienen que agregar estrategias de posprocesamiento o algún tipo de estrategia especial (ver discusión en Nivre 2007).

- A estos axiomas se suele agregar la aciclicidad.
- Una diferencia importante entre las gramáticas basadas en constituyentes y las de dependencias es la condición de exclusividad.
- Desde un punto de vista estrictamente matemático, las gramáticas de dependencias y las gramáticas basadas en constituyentes son débilmente equivalentes.

Las gramáticas de dependencia en la práctica computacional suelen evitar el agregado de categorías nulas, aunque esto no es una restricción que aplique necesariamente a la teoría (de hecho, autores como Hudson, Mel'čuk, Starosta y otros asumen categorías nulas, contra otros como Groß y Osborne 2009)

En esta clase tratamos de introducir los siguientes temas:

- En qué consiste la inteligencia artificial y algunas nociones generales de cómo funcionan las computadoras.
- En qué consiste la programación clásica y el aprendizaje automático.
- Algunas nociones básicas de programación en Python

Bibliografía I

Atserias Batalla, J., Comelles Pujadas, E., y Mayor Martínez, A. (2005). Txala un analizador libre de dependencias para el castellano. *Procesamiento del lenguaje natural, nº 35 (sept. 2005); pp. 455-456*.

Black, E., Abney, S., Flickinger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J. L., Liberman, M., Marcus, M., Roukos, S., Santorini, B., y Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of english grammars. En *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Carnie, A. (2010). *Constituent structure*. Oxford University Press, Oxford.

Carranza, F. (2016). Tesnière y su gramática de dependencias: continuidades y discontinuidades. *RAHL: Revista argentina de historiografía lingüística*, 8(2):59–78.

Bibliografía II

Gaifman, H. (1965). Dependency systems and phrase structure systems. *Information and Control*, (8):304–337.

Groß, T. y Osborne, T. (2009). Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics*, 22:43–90.

Hays, D. G. (1964). Dependency theory: A formalism and some observations. *Language*, 40(4):511–525.

Hudson, R. (1984). *Word Grammar*. Blackwell, Oxford.

Hudson, R. (2010). *An introduction to Word Grammar*. Cambridge University Press, Cambridge.

Leech, G. y Wilson, A. (1996). *Recommendations for the Morphosyntactic Annotation of Corpora*.

Marcus, M., Santorini, B., y Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Bibliografía III

Mel'čuk, I. (2011). Dependency in language-2011. En *Proceedings of International Conference on Dependency Linguistics*, pp. 1–16. Citeseer.

Müller, G. E. (2023). Parentéticas: Elaboración construccional y cuestiones limítrofes. *Cuadernos de la ALFAL*, 15(2):140–158.

Müller, S. (2016). *Grammatical Theory: From transformational grammar to constraint-based approaches*. Language Science Press, Berlin.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. En *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)*, pp. 149–160, Nancy, France.

Bibliografía IV

Nivre, J. (2007). Incremental non-projective dependency parsing. En Sidner, C., Schultz, T., Stone, M., y Zhai, C., editores, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 396–403, Rochester, New York. Association for Computational Linguistics.

Padró, L., Collado, M., Reese, S., Lloberes, M., y Castellón, I. (2010). Freeling 2.1: Five years of open-source language processing tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, La Valletta. ELRA.

Padró, L. y Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul. ELRA.

Bibliografía V

Partee, B., Meulen, A., y Wall, R. (2012). *Mathematical methods in linguistics*. Kluwer Academics, Dordrecht.

Robinson, J. J. (1970). Dependency structures and transformational rules. *Language*, pp. 259–285.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Klincksieck, Paris.

Wall, R. (1972). *Introduction to Mathematical Linguistics*. Prentice-Hall, New Jersey.