Inteligencia Artificial
Los diferentes niveles del procesamiento del lenguaje natural

Primavera 2007

profesor: Luigi Ceccaroni

Objetivos generales

- Conocer el ámbito del PLN y sus principales aplicaciones
- Comprender la problemática asociada a la comprensión del LN y los niveles de análisis sintáctico y semántico
- Conocer las bases de la programación de la análisis con gramáticas de cláusulas definidas (DCGs)

Ámbitos del PLN

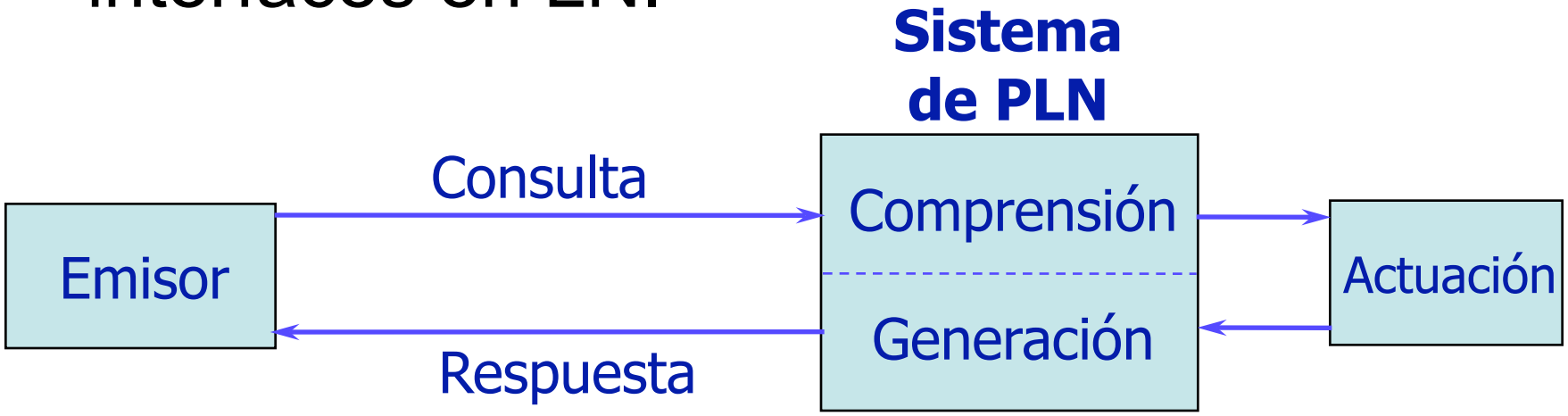
- El PLN consiste en construir sistemas computacionales capaces de **comprender** i **generar** lenguaje **humano** en todas sus formas.
- Para esto se necesita:
  - Saber cómo las personas **generan** expresiones correctas y comprensibles
  - Conocer cómo las personas **comprenden** expresiones de otras personas
  - Ser capaces de **formalizar** el conocimiento y los procesos necesarios de manera que sean tratables por un sistema computacional

Interdisciplinariedad

- Disciplinas asociadas al PLN:
  - Inteligencia artificial
    - Representación del conocimiento
    - Razonamiento
    - Aprendizaje
  - Lingüística computacional
  - Teoría de lenguajes formales
    - Compiladores

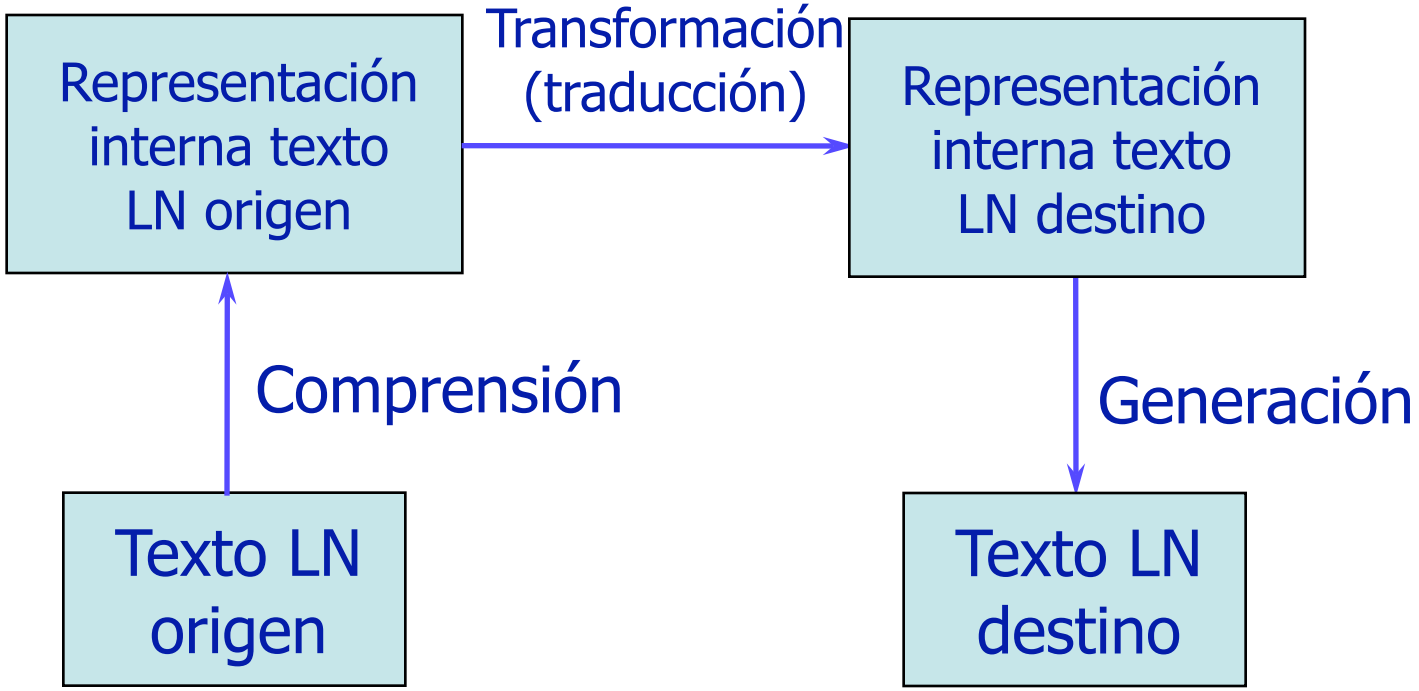
Comprensión y generación

- Son las dos operaciones básicas de las interfaces en LN.



- La consulta y la respuesta pueden ser en lenguaje oral: *speech recognition and synthesis*

Comprensión y traducción



- En lugar de texto puede haber una intervención oral.

Comprender el LN

- La comprensión exige:
  - Extraer el significado individual de las palabras
  - Comprender las relaciones entre las palabras
  - Referir el significado literal al contexto de actuación del sistema
- Todo esto se alcanza a través de un análisis de los componentes del lenguaje a diferentes niveles.

Aplicaciones

- Traducción y resumen automáticos
- Extracción de información a partir de textos
- Interfaces y sistemas de diálogo
- Sistemas de consulta telefónica
- Clasificación y filtro de documentos, email
- *Question answering*
- Web semántica
- Búsqueda de información en Internet

Ejemplo de análisis

"Em parlarà sens dubte de la reestructuració urbana a Barcelona"

- Ejemplos de cosas que hay que detectar:
  - Palabras individuales: em, parlarà, sens...
  - El papel (categoría) de las palabras en la frase: nombre, nombre propio, nombre compuesto, verbo, artículo...
  - La relación entre categorías (papel sintáctico) para establecer el significado global: sujeto, objeto directo...

Niveles de análisis

- Fonológico
- Textual
- Morfológico
- Léxico
- Sintáctico
- Lógico
- Semántico
- Pragmático
- Ilocutivo

Niveles de análisis

- Fonológico
  - Se aplica en el procesamiento del lenguaje oral.
  - Es el tratamiento de los sonidos para detectar unidades de expresión (palabras).

Niveles de análisis

- Textual
  - **Segmentación**: el texto debe ser segmentado en fragmentos que puedan tratarse de forma hasta cierto punto independiente (párrafos, oraciones, intervenciones de diversos interlocutores...).
  - **Filtrado de información no relevante**: los textos a tratar vienen a menudo acompañados de otros materiales que deben ser eliminados o extraídos (por ejemplo, si la fuente de información es una página de Internet, diferentes tipos de marcas que definen las características de visualización de la página).
  - **Localización de unidades tratables**: las unidades básicas de tratamiento son las palabras; localizar las palabras ortográficas es sencillo si el espacio o los signos de puntuación actúan como separadores.

Niveles de análisis

- Textual
  - Detecció d'unitats tractables: **paràgrafs** i **oracions**
    - Mètodes simples
      - basats en localitzar marques de puntuació: ".", "?", "!", "..."
      - Problemes: sigles, inicials
    - Mètodes basats en tècniques d'Aprenentatge Automàtic (classificació)
      - Tenen en compte més informació contextual

Niveles de análisis

- Morfológico
  - La morfología estudia la estructura de las palabras y su relación con las categorías del lenguaje.
  - El objetivo del análisis morfológico automático es llevar a cabo una clasificación morfológica de las palabras.
  - Por ejemplo, el análisis de la palabra *gatos* resulta en

    gato+Noun+Masc+Pl,

    que nos indica que se trata de un sustantivo plural con género masculino y que su forma normalizada (lema) es gato.

Niveles de análisis

- Morfológico
  - Versió simple: utilització de **formaris** (llista de formes amb informació morfològica i els lexemes corresponents)
  - Morfema = lexema (o arrel) o gramema

Lexema	Gramema
cant	o es a em en

Niveles de análisis

- Morfológico
  - **Analitzadors morfològics**:
    - Diccionaris de morfemes:
      - diccionari d'arrels (lexemes), de sufixes, prefixes, infixes
    - Morfotàctica: regles de combinació de morfemes
    - Variacions fonològiques: canvis al combinar els morfemes (ex., ploure, plovisquejar)
  - Tipus d'**analitzadors**
    - FSA (finite state automaton)
    - FST (finite state transducer)
    - cascada de FSTs

Niveles de análisis

- Léxico
  - Distingue entre palabras ortográficas y palabras gramaticales.
  - Obtiene información léxica de diccionarios, ontologías...

Niveles de análisis

- Léxico
  - Detectar unitats de significat
    - Requereix ser capaç de reconèixer i fragmentar adequadament les paraules: "***/Parlarà/ /sens dubte/ /de/ /les/ / reestructuracions/ /urbanes/ /a/ /Sant Cugat/***"
  - Recollir informacions útils i aplicar coneixements per a facilitar les fases d'anàlisi posteriors
    - Associar categories gramaticals
    - Associar informació semàntica a les unitats lèxiques (ús d'ontologies, diccionaris)
    - Reconeixement i classificació de noms propis i entitats

Niveles de análisis

- Léxico
  - **Correspondència paraules ortogràfiques /gramaticals**
    - Necessitat de coneixement o informació per a detectar els casos següents:

      “dóna-m’ho”, “*dímelo*” (1 p. ortogràfica, 3 p. gramaticals)

      “sens dubte”, “*sin embargo*” (2 p. ortogràfiques, 1 p. gramatical)
  - **Homonímia**
    - Mateixa forma i diverses categories gramaticals

      “roda” (verb, 3a persona), **“roda”** (nom) -> connexió sintaxis
  - **Polisèmia**
    - Mateixa forma i categoria, diversos significats (p.ex, **“banc”**)

Niveles de análisis

- Léxico
  - **Sigles**
    - **“Un cop s’ha generat un PCB es pot enviar a una cua FIFO”**
    - **“*The cell’s DNA sample was identified by PRC, a process approved by the official UBI*”**
  - **Abreviatures**
    - **“El Dr. Pirvo va parlar del Tract. del Lleng. Natural…”**
  - **Fórmules i mesures**
    - **“Afegir dos mg de DM-oxano i guardar dins d’un vial de PVC”**
    - **“Si tenim en compte que *x*=*y**2 + *k*, on *k* és una constant...”**
  - **Volum d’informació**

Niveles de análisis

- Ambigüedad léxica
  - "*Pinchó la rueda de delante*"
    - "rueda" puede ser nombre o verbo (*part-of-speech tagging* - *POS-tagging*)
  - "*Vio el banco*"
    - "banco" puede ser el mueble para sentarse, la entidad financiera o un grupo de peces (*word sense disambiguation* - WSD)

Niveles de análisis

- Utilització de **lexicons**
  - "Diccionaris lèxics"
  - Apleguen informació útil per a reconèixer i categoritzar paraules i la seva ubicació al text

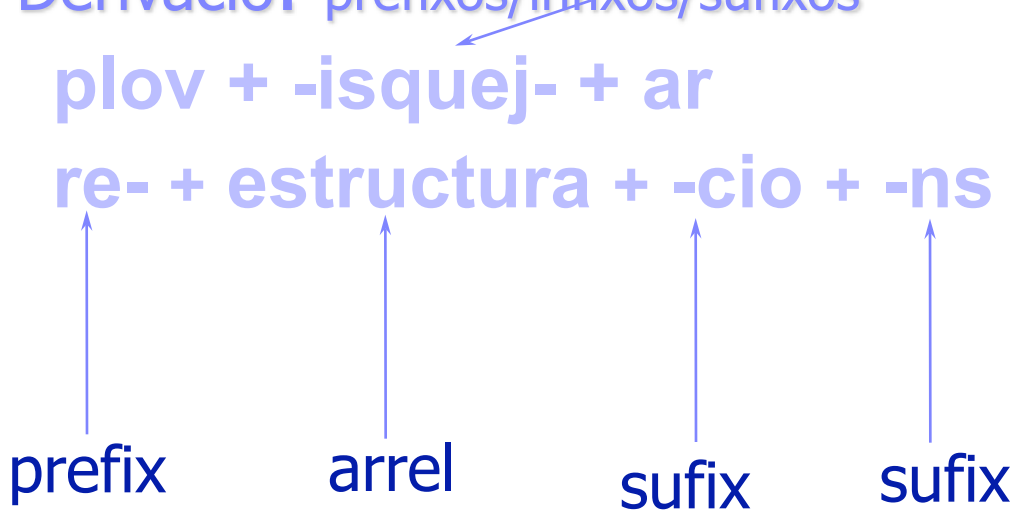
Lexema	Informació
cant-	cantar V / Infinitiu -o/-es/-a/-em/-eu/-en

Problemàtica: **representació (1)**

- Decidir el tipus d'informació que ha de contenir:
  - Categoria sintàctica
    - determinant, proposició, nom propi, substantiu, verb, etc.
    - Problema de la granularitat (verb -> transitiu/intransitiu)
  - Propietats sintàctiques de concordança
    - gènere (masculí/femení)
    - nombre (singular/plural)
    - persona (primera, segona...)
    - cas (acusatiu,datiu..)

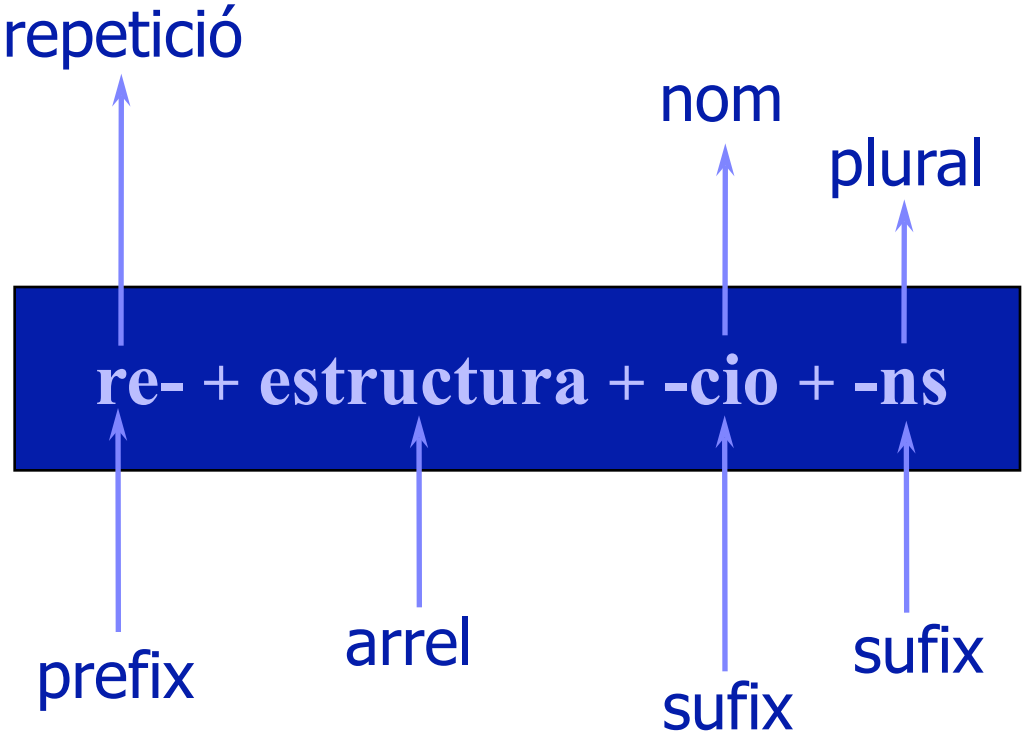
Problemàtica: **representació** (2)

- Altres propietats sintàctiques:
  - Tipus de complement del verb
  - Preposicions que accepta una paraula
- Categoria semàntica
- Informació morfològica
  - Derivació: prefixos/infixos/sufixos



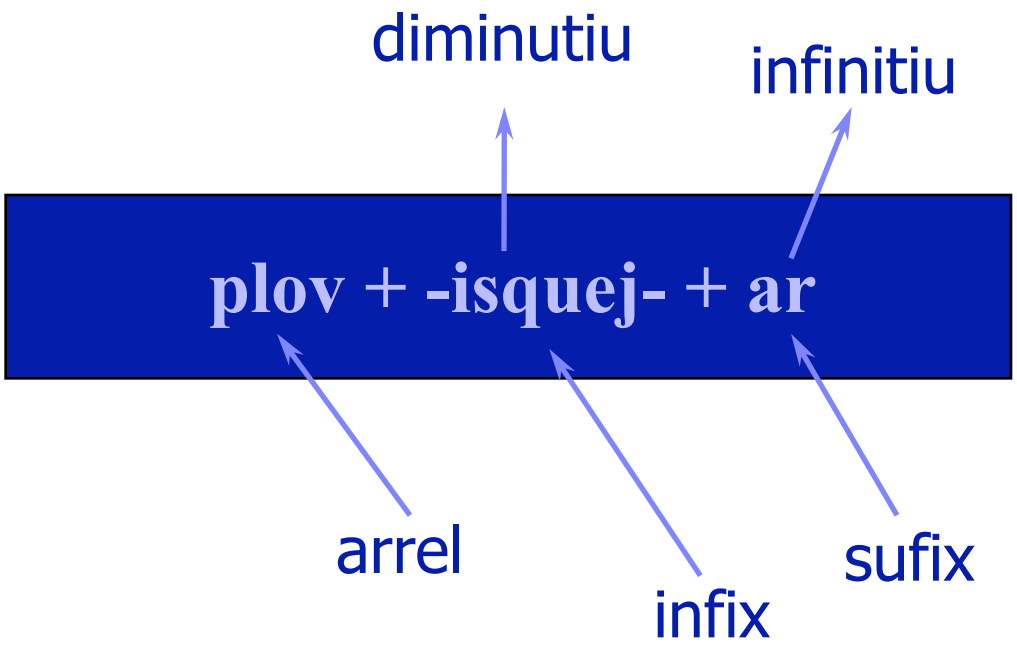
Problemàtica: **representació (3)**

- Informació lèxica



Problemàtica: **representació (4)**

- Informació lèxica

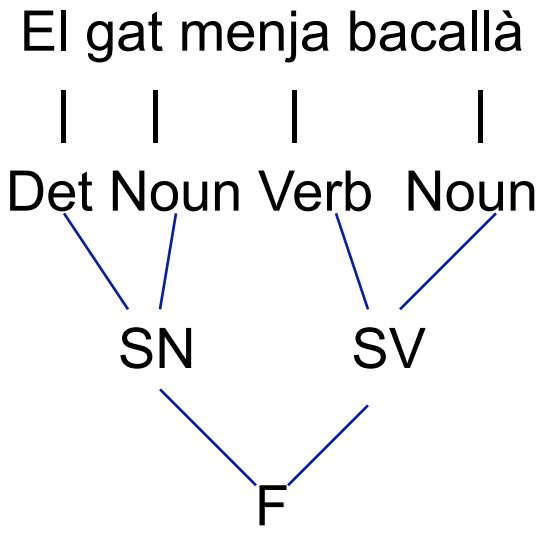


Niveles de análisis

- Sintáctico
  - Reconoce, extrae y representa estructuras sintácticamente válidas (o inválidas):

Els gat vell menja bacallà

El gata menja bacallà



Niveles de análisis

- Ambigüedad sintáctica
  - "*El vendedor de diarios del barrio*" (*prepositional-phrase attachment - PP-attachment*)
  - "*Vio un hombre con unos prismáticos*"

Niveles de análisis

- Lógico
  - Extrae y representa el significado literal de una oración a través de un lenguaje formal: cálculo de predicados de primer orden (CP1), ontologías, mapas conceptuales...
  - En el caso de CP1, expresiones en términos de predicados, variables, funciones, constantes, conectivas lógicas...

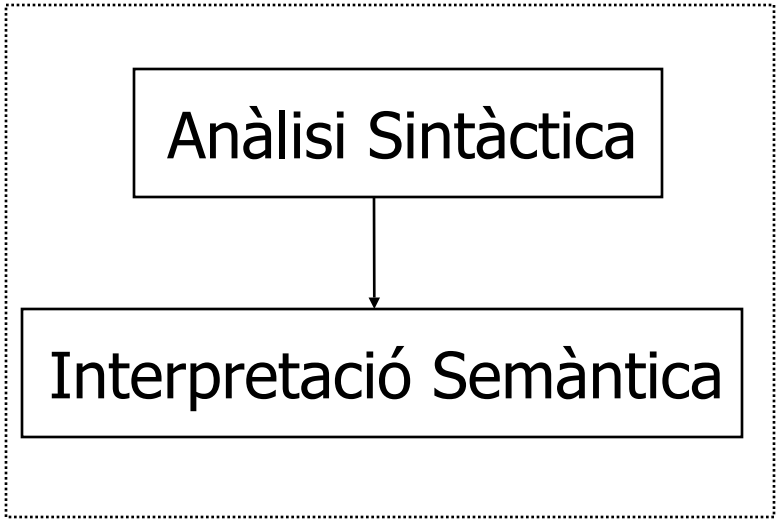
"**El gat menja bacallà**"

existen x, y (Gat(x) & Bacallà(y) & Menja(x,y))

Niveles de análisis

- Semàntic
  - Interpretació de la forma lògica: Relació de les entitats lògiques (constants, variables, termes, etc.) amb el món real (o la seva representació): objectes del domini
  - El gat és un felí, el bacallà és un peix comestible, l'actor de menjar ha de ser un ésser viu, etc.
  - Extreure significat global a partir de significats individuals i relacions
- Ambigüedad semántica
  - "*Dio un pastel a los niños*"
    - Puede ser 1 a todos o 1 a cada niño
  - "*Las ideas verdes duermen furiosamente*"

Análisis sintáctico y semántico



- sense sintaxi
- sense semàntica
- procés en cascada (1)
  - sintaxi | semàntica
- procés en cascada (2)
  - {sintaxi + filtre semàntic} | semàntica
- procés en paral·lel
  - {sintaxi, semàntica}

Preprocés

- Segmentació
- Localització d'unitats (paraules)
- Lematització, anàlisi morfològica
- Desambiguació morfosintàctica (*POS-tagging*)
- Etiquetat semàntic
- Desambiguació semàntica (WSD)
- Detecció i classificació d'entitats amb nom (*named entity recognition*, NER)

Exemple

Quina es la capital de França?

resultat de l'anàlisi morfològica

quina	quin	DT0FS00	quina	NCFS000		
és	ésser	VMIP3S0				
la	el	TDFS0	ell	PP3FSO00	la	I
capital	capital	AQPCS00	capital	NCFS000	capital	NCMS000
de	de	SPS00				
França	França	NP00000-loc				
?	?	Fit				

resultat del POS-tagging

quina	quin	DT0FS00
és	ésser	VMIP3S0
la	el	TDFS0
capital	capital	NCFS000
de	de	SPS00
França	França	NP00000-loc
?	?	Fit

Postprocés

- Anàlisi semàntica - pragmàtica
- Anàlisi il·locutiva
  - Reconeixement d'intencions

Anàlisi sintàctica (1)

- Objectius
  - Determinar que l'oració (la unitat textual) es sintàcticament correcta
  - Crear una estructura sintàctica amb informació que pugui ser utilitzada per a l'anàlisi semàntica i d'altres

Anàlisi sintàctica (i 2)

- Alfabet (vocabulari) Σ
- Operació de concatenació
- Σ^*: conjunt de totes les cadenes amb símbols de Σ (monoide lliure)
- llenguatge $L \subseteq \Sigma^*$
- donada una cadena de Σ^*, $w_1^n = w_1, \ldots w_n$, $w_i \in \Sigma$, hem de determinar si $w_1^n \in L$

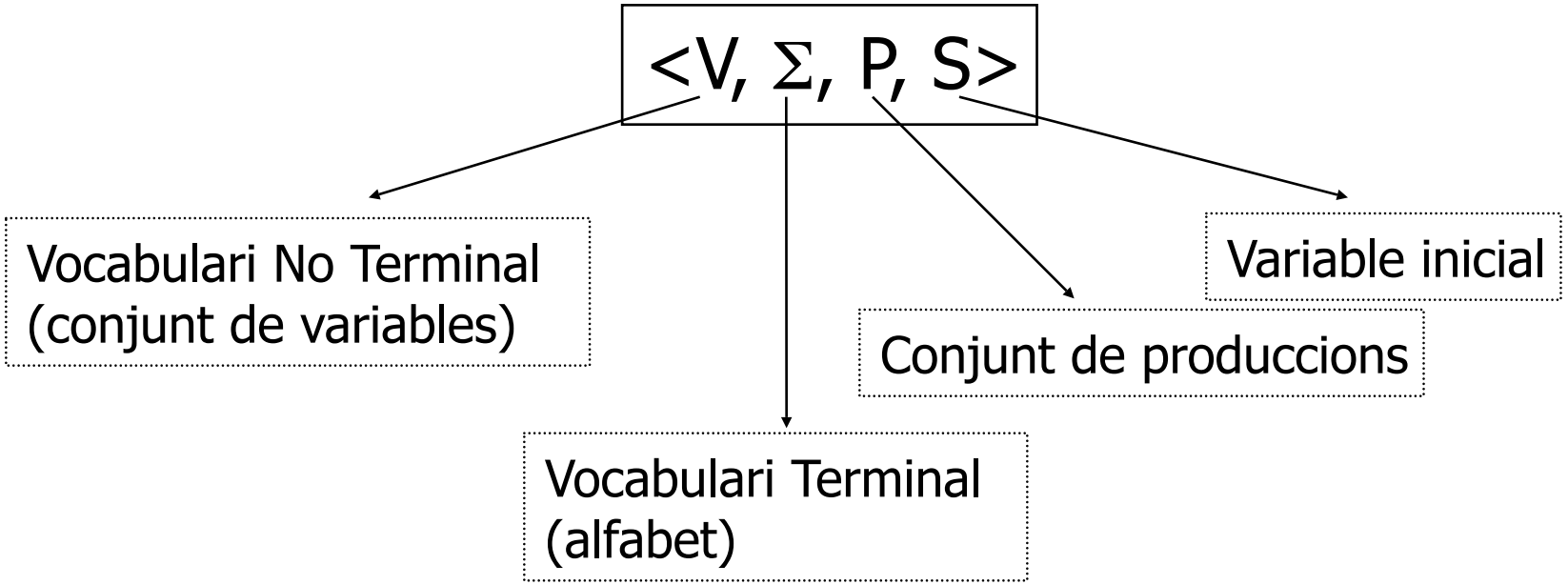
Formes de definir la pertinència

- Gramàtica
  - $G \Rightarrow L(G)$
  - $w_1^n \in L(G)$?
- Model del llenguatge
  - $P(w_1^n)$
  - si $P(w_1^n) > 0 \Rightarrow w_1^n \in L$
- Corpus (oracions, patrons) que defineix les oracions correctes
  - diccionari sintàctic
  - regles de composició
- Regles de bona formació
  - filtres, gramàtiques negatives

Forma més habitual: gramàtica

- Exemples:
  - Gramàtiques d'estructura sintagmàtica
  - Gramàtiques de constituents
    - Arbres de derivació
  - Gramàtiques de dependències
    - Esquemes de dependències
  - Gramàtiques de casos
    - Models d'actants => Xarxes semàntiques
  - Gramàtiques transformacionals
  - Gramàtiques sistèmiques

Gramàtiques d'estructura sintagmàtica



$\Sigma \cap V = \emptyset$

$\Sigma \cup V =$ Vocabulari

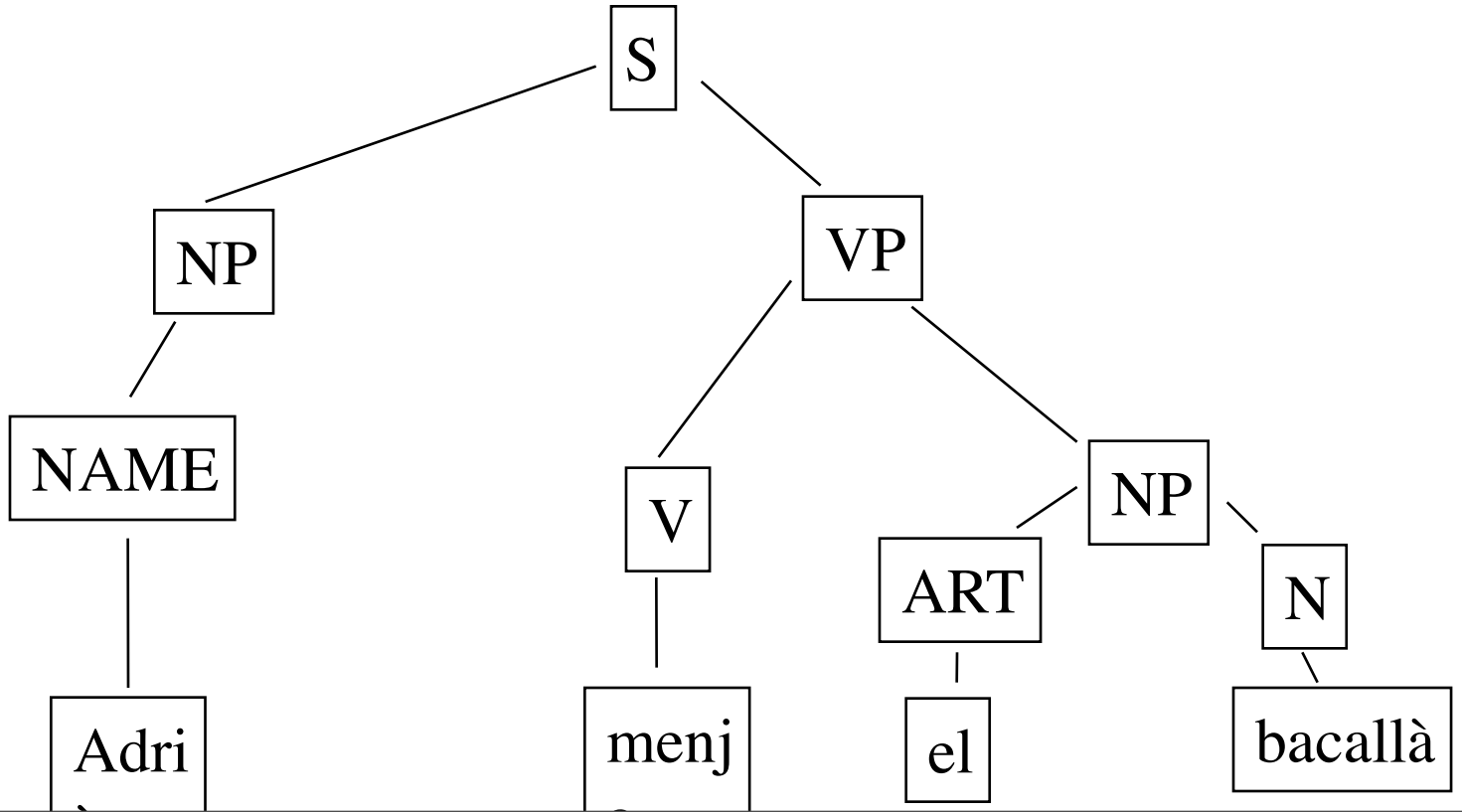
$S \in V$

Grammars and parsing

- To examine how the syntactic structure of a sentence can be computed:
  - **Grammar**, a formal specification of the structures allowable in the language
  - **Parsing technique**, the method of analyzing a sentence to determine its structure according to the grammar
- The most common way of representing how a sentence is broken into its major subparts (**constituents**), and how those subparts are broken up in turn, is a **tree**.

Grammars and sentence structure (1)

- Tree representation for the sentence *Adrià menja el bacallà*:



Grammars and sentence structure (2)

- The sentence (S) consists of an initial noun phrase (NP) and a verb phrase (VP).
- The initial noun phrase is made of the simple NAME *Adrià.*
- The verb phrase is composed of a verb (V) *menja* and an NP, which consists of an article (ART) *el* and a common noun (N) *bacallà.*
- In list notation this same structure could be represented as:

(S (NP (NAME Adrià))
(VP (V menja)
(NP (ART el)
(N bacallà))))

Grammars and sentence structure (3)

- To construct a tree structure for a sentence, you must know what structures are legal.
- A set of **rewrite rules** describes what tree structures are allowable.
- These rules say that a certain symbol may be expanded in the tree by a sequence of other symbols.
- A set of rules constitutes a grammar:

1. S → NP VP
2. VP → V NP
3. NP → NAME
4. NP → ART N
5. NAME → Adrià
6. V → menja
7. ART → el
8. N → bacallà

Grammars and sentence structure (4)

- Rule 1 says that an S may consist of an NP followed by a VP.
- Rule 2 says that a VP may consist of a V followed by an NP.
- Rules 3 and 4 say that an NP may consist of a NAME or may consist of an ART followed by an N.
- Rules 5 - 8 define possible words for the categories.
- Grammars consisting entirely of rules with a single symbol on the left-hand side, called the **mother**, are called **context-free grammars** (CFGs).

Grammars and sentence structure (5)

- Context-free grammars (CFGs) are a very important class of grammars because:
  - the formalism is powerful enough to describe most of the structure in natural languages,
  - yet is restricted enough so that efficient parsers can be built to analyze sentences.
- Symbols that cannot be further decomposed in a grammar (the words *Adrià*, *menja*…) are called **terminal symbols**.
- The other symbols, such as NP and VP, are called **nonterminal symbols**.
- The grammatical symbols such as N and V that describe word categories are called **lexical symbols**.
- Many words will be listed under multiple categories. For example, *poder* would be listed under V and N.
- Grammars have a special symbol called the *start symbol*. Usually, the start symbol is S (meaning *sentence*).

Grammars and sentence structure (6)

- A grammar is said to **derive** a sentence if there is a sequence of rules that allow you to rewrite the start symbol into the sentence, for instance, *Adrià menja el bacallà.*
- This can be seen by showing the sequence of rewrites starting from the S symbol, as follows:

S	
=> NP VP	**(rewriting S)**
=> NAME VP	**(rewriting NP)**
=> Adrià VP	**(rewriting NAME)**
=> Adrià V NP	**(rewriting VP)**
=> Adrià menja NP	**(rewriting V)**
=> Adrià menja ART N	**(rewriting NP)**
=> Adrià menja el N	**(rewriting ART)**
=> Adrià menja el bacallà	**(rewriting N)**

Grammars and sentence structure (7)

- Two important processes are based on derivations:
  - The first is **sentence generation**, which uses derivations to construct legal sentences. A simple generator could be implemented by randomly choosing rewrite rules, starting from the S symbol, until you have a sequence of words. The preceding example shows that the sentence *Adrià menja el bacallà* can be generated from the grammar.
  - The second process based on derivations is **parsing**, which identifies the structure of sentences given a grammar.

Grammars and sentence structure (8)

- There are two basic methods of searching:
  - A **top-down strategy** starts with the S symbol and then searches through different ways to rewrite the symbols until the input sentence is generated, or until all possibilities have been explored. The preceding example demonstrates that *Adrià menja el bacallà* is a legal sentence by showing the derivation that could be found by this process.

Grammars and sentence structure (and 9)

- In a **bottom-up strategy**, you start with the words in the sentence and use the rewrite rules backward to reduce the sequence of symbols until it consists solely of S. The left-hand side of each rule is used to rewrite the symbol on the right-hand side. A possible bottom-up parse of the sentence *Adrià menja el bacallà* is:

  => NAME menja el bacallà (rewriting Adrià)
  => NAME V el bacallà (rewriting menja)
  => NAME V ART bacallà (rewriting el)
  => NAME V ART N (rewriting bacallà)
  => NP V ART N (rewriting NAME)
  => NP V NP (rewriting ART N)
  => NP VP (rewriting V NP)
  => S (rewriting NP VP)

- A tree representation can be viewed as a record of the CFG rules that account for the structure of the sentence.

What makes a good grammar

- In constructing a grammar for a language, you are interested in:
  - **generality**, the range of sentences the grammar analyzes correctly;
  - **selectivity**, the range of non-sentences it identifies as problematic;
  - **understandability**, the simplicity of the grammar itself.

Generative capacity (1)

- Grammatical formalisms based on rewrite rules can be compared according to their **generative capacity**, which is the range of languages that each formalism can describe.
- It turns out that no natural language can be characterized precisely enough to define generative capacity.
- Formal languages, however, allow a precise mathematical characterization.

Generative capacity (2)

- Consider a formal language consisting of the symbols *a, b,* c and *d* (think of these as words).
- Then consider a language L1 that allows any sequence of letters in alphabetical order. For example, *abd, ad, bcd, b,* and *abcd* are all legal sentences. To describe this language, we can write a grammar in which the right-hand side of every rule consists of one terminal symbol possibly followed by one nonterminal.
- Such a grammar is called a **regular** grammar. For L1 the grammar would be:

S -> a S1	**S -> d**	**S1 -> d**	**S3 -> d**
S -> b S2	**S1 -> b S2**	**S2 -> c S3**	
S -> c S3	**S1 -> c S3**	**S2 -> d**	

Generative capacity (3)

- Consider another language, L2, that consists only of sentences that have a sequence of *a's* followed by an equal number of b's—that is, *ab, aabb, aaabbb,* and so on. You cannot write a regular grammar that can generate L2 exactly.
- A **context-free** grammar to generate L2, however, is simple:

S -> a b S -> a S b

Generative capacity (4)

- Some languages cannot be generated by a CFG.
- One example is the language that consists of a sequence of a's, followed by the same number of b's, followed by the same number of c's - that is, *abc, aabbcc, aaabbbccc,* and so on.
- Similarly, no context-free grammar can generate the language that consists of any sequence of letters repeated in the same order twice, such as *abab, abcabc, acdabacdab,* and so on.
- There are more general grammatical systems that can generate such sequences, however. One important class is the **context-sensitive** grammar, which consists of rules of the form:

$$\alpha\, A\, \beta \rightarrow \alpha\, \psi\, \beta$$

where A is a symbol, α and β are (possibly empty) sequences of symbols, and ψ is a nonempty sequence of symbols.

Generative capacity (and 5)

- Even more general are the **type 0** grammars, which allow arbitrary rewrite rules.
- Work in formal language theory began with Chomsky (1956). Since the languages generated by regular grammars are a subset of those generated by context-free grammars, which in turn are a subset of those generated by context -sensitive grammars, which in turn are a subset of those generated by type 0 languages, they form a hierarchy of languages (called the **Chomsky Hierarchy**).

Condició de gramaticalitat

Una frase w (un mot de Σ^*) pertany al llenguatge generat per la gramàtica G, si la gramàtica G pot derivar el mot w utilitzant les produccions a partir de S.

Llenguatges associats a les gramàtiques de la jerarquia de Chomsky

Gramàtica	**Reconeixedor**	**Llenguatge**
Tipus 0	màquines de Turing	llenguatges enumerables recursivament
Tipus 1	linear-bounded automata (LBA)	llenguatges contextuals
Tipus 2	autòmats a pila no deterministes	llenguatges incontextuals
Tipus 3	autòmats finits (FSA)	llenguatges regulars

Obtenció de la gramàtica

Obtenció de la gramàtica

- Definició de l'etiquetari terminal (*tagset*, Σ)

Obtenció de la gramàtica

- Definició de l'etiquetari terminal (*tagset*, Σ)
- Definició del etiquetari no terminal (V)

Obtenció de la gramàtica

- Definició de l'etiquetari terminal (*tagset*, Σ)
- Definició del etiquetari no terminal (V)
- Regles de la gramàtica (P)
  - construcció manual
  - construcció automàtica
    - inferència (inducció) gramatical
  - construcció semiautomàtica

Gramàtiques per al tractament de la llengua

- Mínim: gramàtiques incontextuals
- És el LN un llenguatge incontextual?
- Suficient? NO (normalment)
- Solució
  - CFG + {adició procedimental del context}
  - Gramàtiques lògiques i d'unificació
  - Gramàtiques enriquides amb informació estadística
    - SCFG
  - Gramàtiques lexicalitzades

Exemple de gramàtica incontextual (G1)

(1) Oracio ® GN, GV
(2) GN ® det, n
(3) GN ® n
(4) GV ® vi
(5) GV ® vt, GN
(6) det ® el | un | ...
(7) n ® gat | peix | ...
(8) vt ® menja | ...
(9) vi ® menja | ...

Exemple de CFG + {addició procedimental del context}

intervencio	®	pregunta | ordre | ...
ordre	®	v, sn {imperatiu(1), ordre(1)}
sn	®	snbase, [snmods] | np {concordancia (1,2)}
snbase	®	[det], n, [adjs] {concordancia (1,2,3)}
adjs	®	adj, [adjs]
snmods	®	snmod, [snmods]
snmod	®	sp | ...
sp	®	prep, sn
np	®	"barcelona" | "valencia" | ...
n	®	"billet" | "euromed" ...
v	®	"donim" | ...
det	®	"un" | "el" | ...

Analitzadors sintàctics

- Factors que influeixen l'anàlisi sintàctica
- Analitzadors per a CFG
- Analitzadors per a gramàtiques d'unificació ("unificació" com a mecanisme bàsic de composició entre constituents)
- Resultat de l'anàlisi sintàctica

Factors que incideixen en el procés d'anàlisi sintàctica

- Expressivitat de la gramàtica
- Àmbit (*coverage*)
- Fonts de coneixement implicades
- Estratègia de l'anàlisi
- Direcció de l'anàlisi
- Ordre d'aplicació de les regles
- Gestió de l'ambigüitat
- (In)determinisme
- Enginyeria dels analitzadors

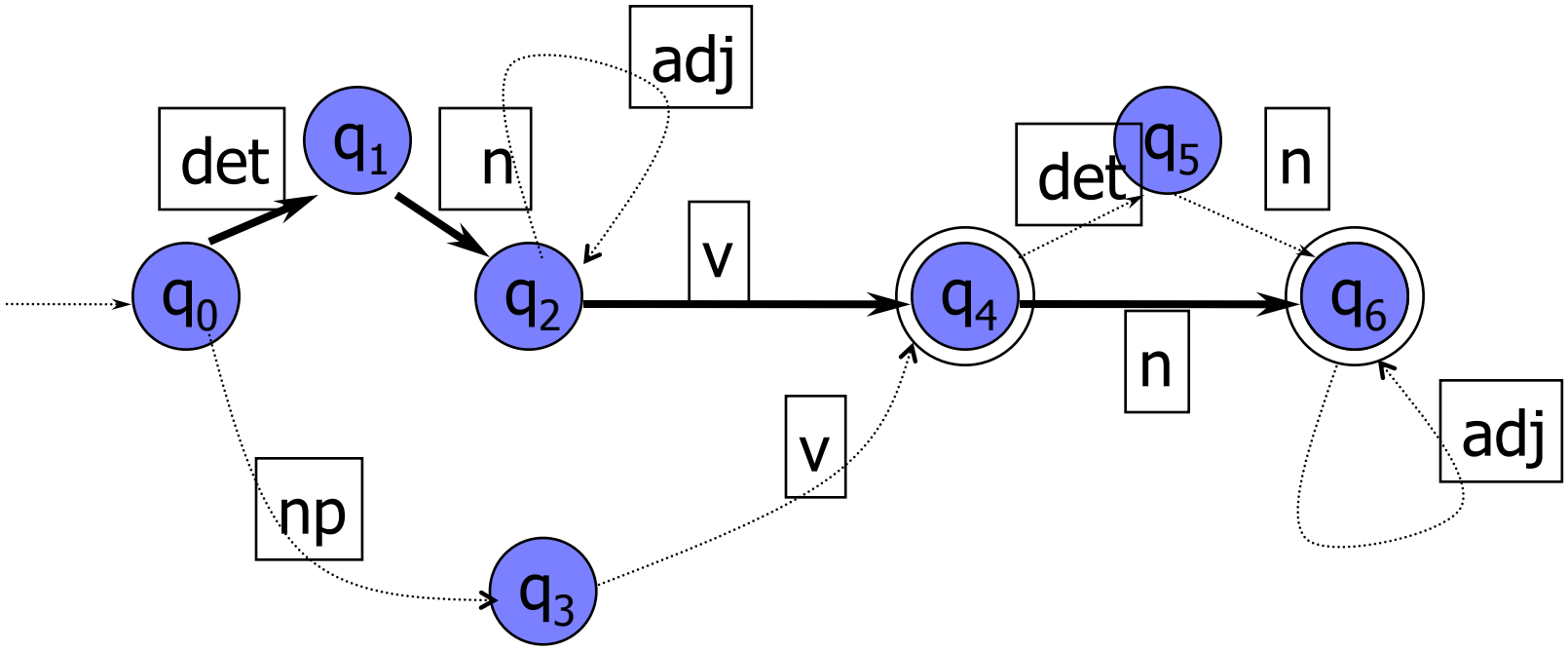
Analitzadors per CFG i extensions

- Simplificacions de CFG
  - CFG ⇒ RG
    - Tècniques d'estats finits: FSA
- Extensions dels FSA
  - TN ⇒ RTN (*Recursive Transition Network*) ⇒ ATN (*Augmented Transition Network*) (Woods, 1970)
- WFST (*Well-Formed String Tables*), Charts (M. Kay, 1980)
- Gramàtiques d'estructura de frase: LSP (N. Sager, 1981), Diagram (A. Robinson, 1981)
- Parsifal (M. Marcus, 1980)

Xarxes de Transició (TNs)

- Autòmat finit
  - Estats associats a parts de la frase
  - Transicions
    - Etiquetes que fan referència a categories morfosintàctiques
      - Una transició és acceptable si la paraula té la mateixa categoria que apareix etiquetada a l'arc
  - No determinisme
    - Més d'un estat inicial
    - Una paraula amb més d'una categoria possible
    - Més d'un arc per la mateixa categoria

TN: exemple



El	gat	menja	bacallà
det	n	v	n

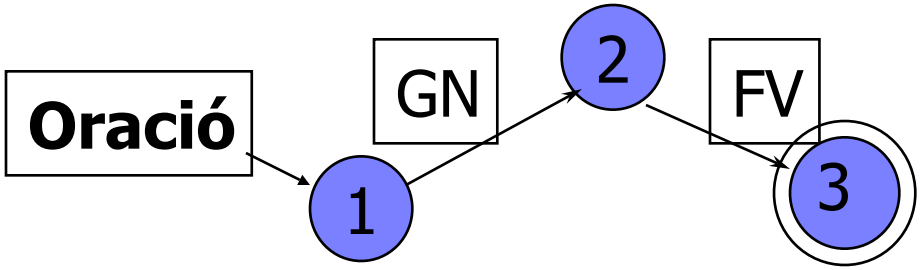
TN: limitacions

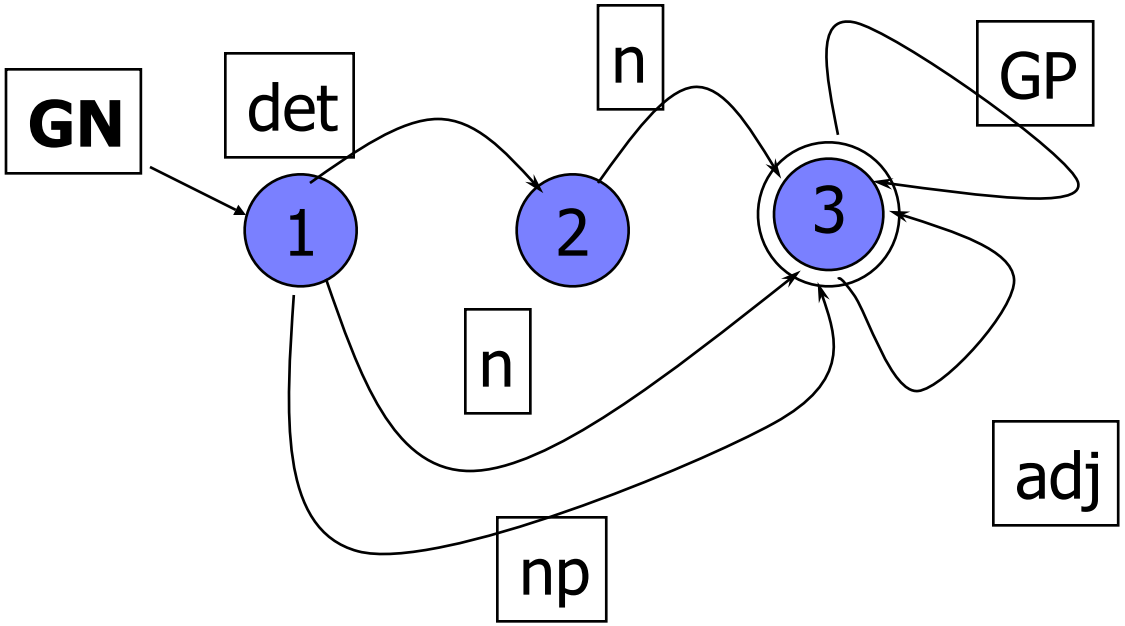
- Limitat a llenguatges regulars
- No es pot dir que analitzi
  - Reconeix
- No-determinisme $\Rightarrow$ *backtracking*
  - Ineficiència
- No separació entre gramàtica i analitzador
  - gramàtica $\Rightarrow$ descripció del model sintàctic
  - analitzador (*parser*) $\Rightarrow$ control

TNs Recurrents (RTNs)

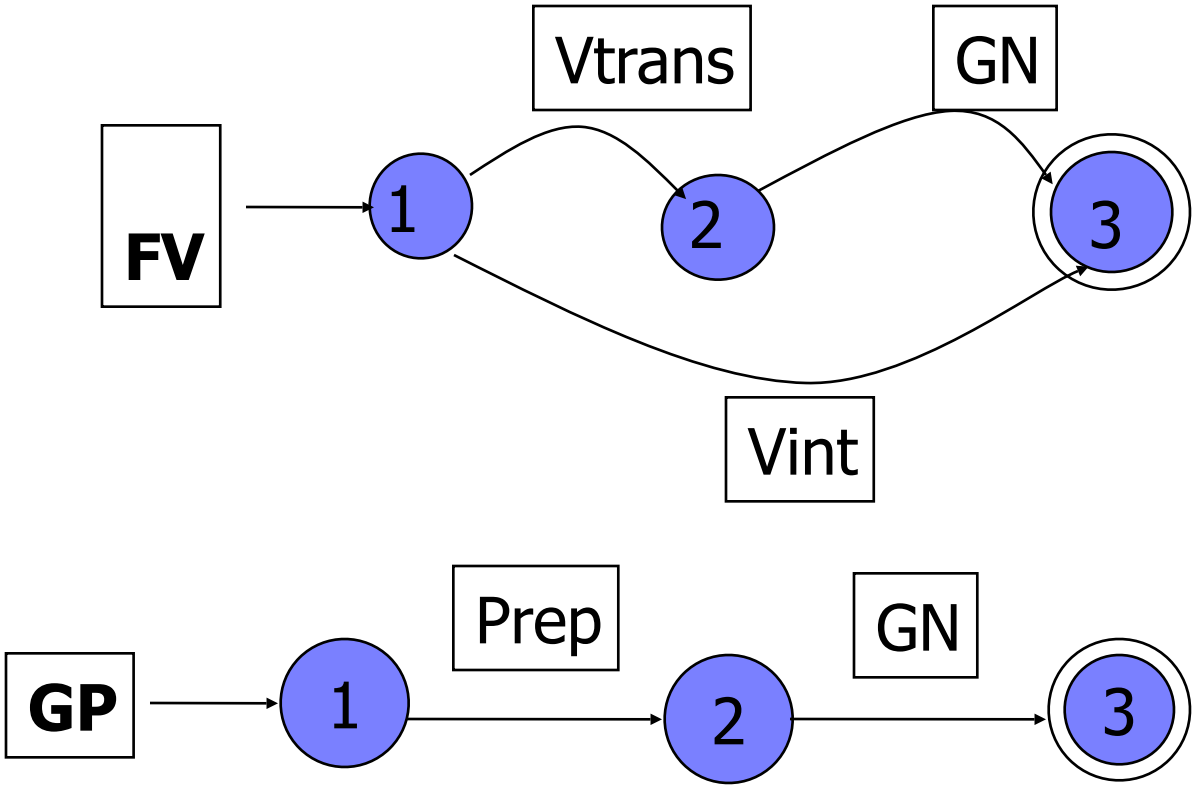
- Col·lecció de xarxes de transició (TNs) etiquetades amb un nom
  - Arcs
    - Etiquetats amb categories → com xarxes normals
      - Etiquetes terminals
    - Etiquetats amb identificadors de xarxes de transició (TNs)
      - Etiquetes no terminals: els estats finals de les TNs causen el retorn a l'estat destí de la transició que ha causat la crida
- Les RTN son dèbilment equivalents a les CFG

RTN: exemples (1)





RTN: exemples (i 2)



RTN: limitacions

- Transicions només depenen de les categories (poc expressiu)
  - Llenguatge de context lliure
- Reconeixen però no analitzen
- Ineficiència inherent al *backtracking*

Redes de transición aumentadas (ATNs)

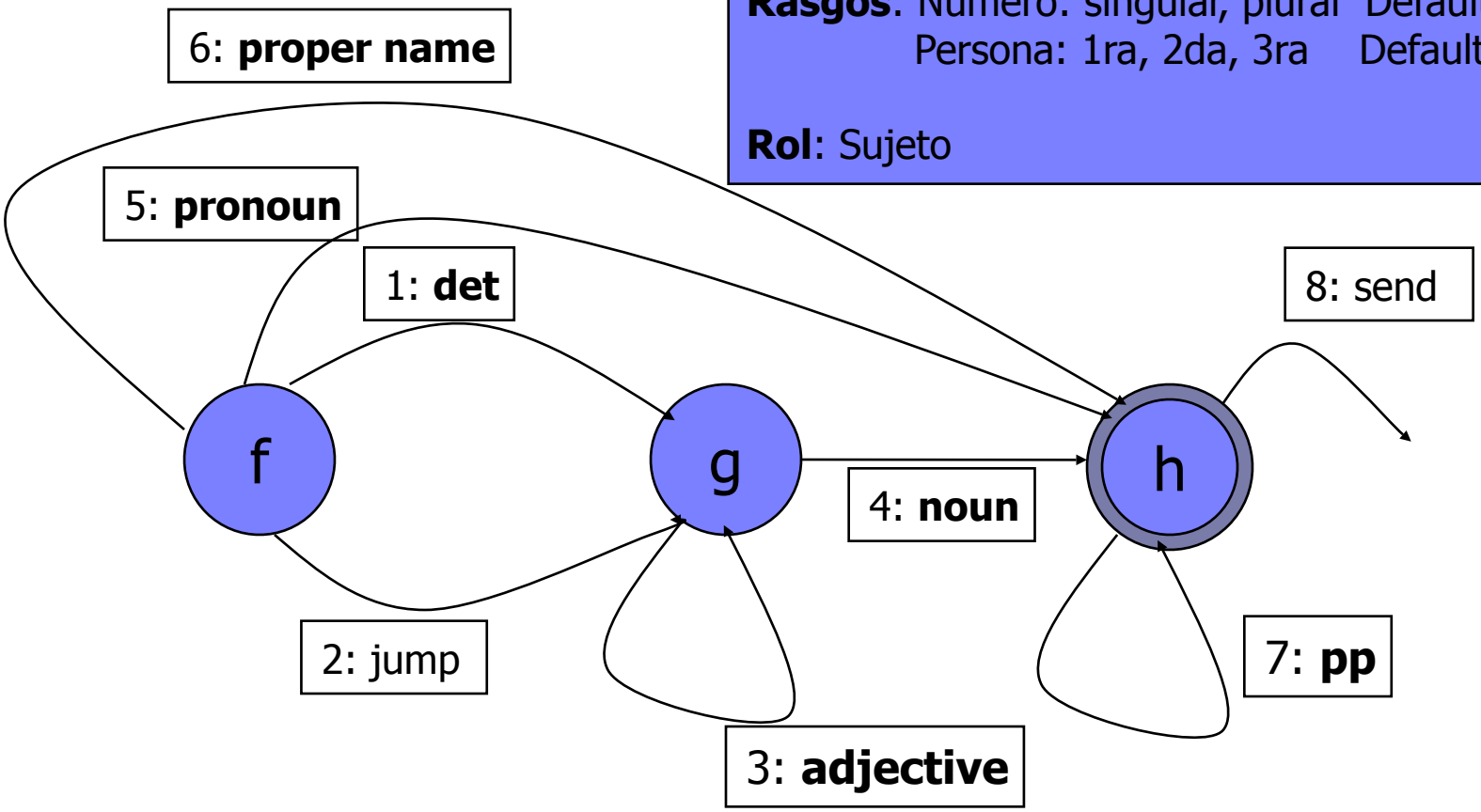
- ATNs = RTNs con operaciones añadidas a los arcos y uso de registros
- Operaciones:
  - Condiciones: filtrar transiciones entre estados
  - Acciones: construir estructuras de salidas y convertir el reconocimiento en análisis
  - Inicializaciones
- Permiten expresar las restricciones del contexto.

ATN: ejemplo

Red para reconocer grupos nominales (NP)

Rasgos: Número: singular, plural Default: vacío
Persona: 1ra, 2da, 3ra Default: 3ra

Rol: Sujeto



ATN: ejemplo

- Inicializaciones, condiciones y acciones

NP-1: $_{f}$Determiner$_{g}$
A: Set Number to the number of *
NP-4: $_{g}$Noun$_{h}$
C: Number is empty or number is the number of *
A: Set Number to the number of *
Set Nucli-Subjecte to *
NP-5: $_{f}$Pronoun$_{h}$
A: Set Number to the number of *
Set Person to the Person of *
Set Nucli-Subjecte to *
NP-6: $_{f}$Proper$_{h}$
A: Set Number to the number of *
Set Nucli-Subjecte to *

ATNs: limitaciones

- No resulta fácil implementar un análisis ascendiente o híbrido.
- Hay redundancia en las operaciones de vuelta atrás:
  - ineficiencia
- Problemas de expresividad de la notación:
  - La gramática se mezcla con las acciones.

Charts (1)

- Intenten eliminar redundàncies en l'anàlisi (alleugeriment del cost del *backtracking*) memoritzant estructures parcials ja construïdes.
- No afecten l'estratègia de l'anàlisi
- Inconvenients: espai, temps de construcció, només guarden components ben formats

Charts (2)

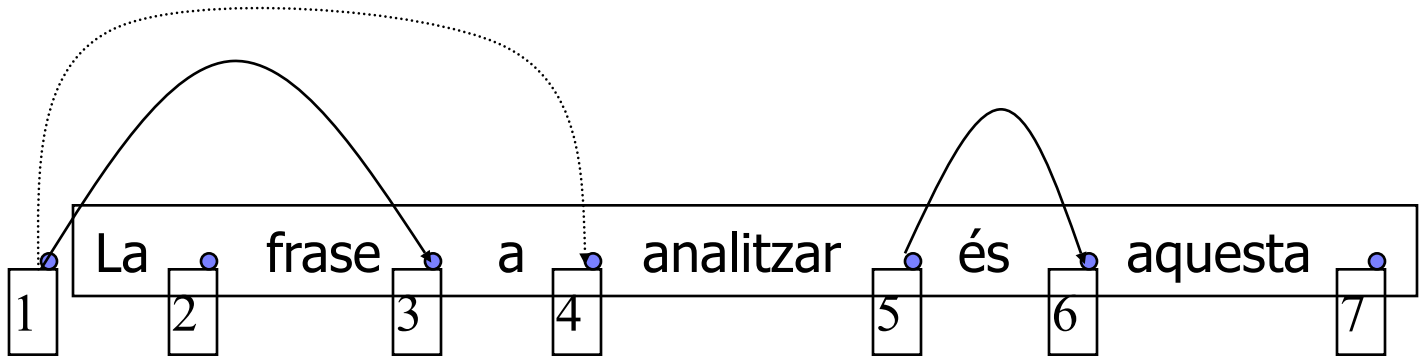
- **Chart** = graf dirigit que es construeix de manera dinàmica i incremental a mesura que es realitza l'anàlisi.

- Els **nodes** corresponen al principi i final de la frase i a les separacions entre paraules (N+1 nodes)

Charts (i 3)

- Els **arcs** es creen dinàmicament. Un arc de la posició i a la j (j ≥ i) engloba totes les paraules que estan entre la posició i i la j.

- Els arcs poden ser
  - **actius** = objectius o hipòtesis per completar
  - **inactius** = components completament analitzades



Charts: notació (1)

- ***Regla puntejada*** (DR, “dotted rule”): producció de la gramàtica que conté algun punt en la seva part dreta.

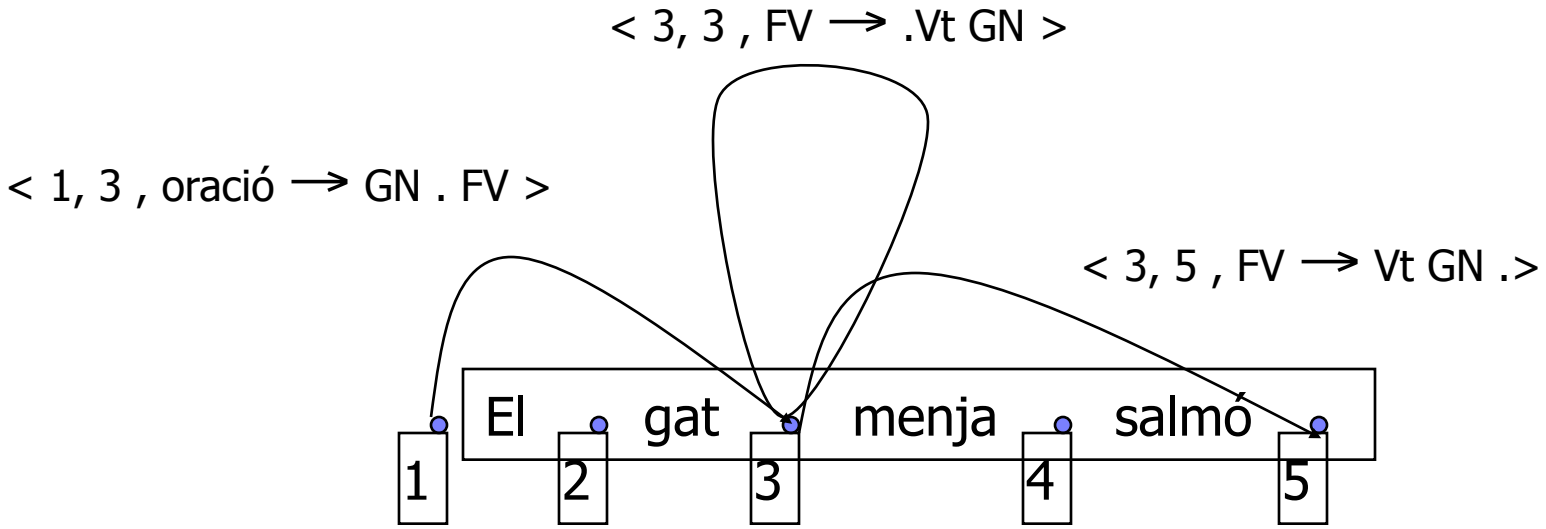
Per exemple, de la regla A –> BCD es poden derivar les següents regles puntejades:

A –> . B C D (corresponent a un arc actiu)
A –> B . C D ”
A –> B C . D ”
A –> B C D . (corresponent a un arc inactiu)

Charts: notació (2)

Arc d'un chart: $< i , j , X \rightarrow a.b >$

i , j :	nodes origen i destí
$X \rightarrow ab$	producció de la gramàtica
$X \rightarrow a.b$	DR



Regla bàsica de combinació

Arc actiu: $\langle i, j, A \rightarrow a.Bb \rangle$

Arc inactiu: $\langle j, k, B \rightarrow g. \rangle$

Resultat: $\langle i, k, A \rightarrow aB.\ b \rangle$

Estratègia ascendent

Regla bàsica: Cada vegada que s'afegeix un arc inactiu al Chart <i, j, A → a.>, aleshores s'ha d'afegir al seu extrem esquerre un arc actiu <i, i, B → .Ab > per cada regla B → Ab de la gramàtica

Inicialització: afegir els **arcs inactius** que corresponen a les categories lèxiques (terminals). Ex: <1, 2, Det → el.>

Estratègia descendent

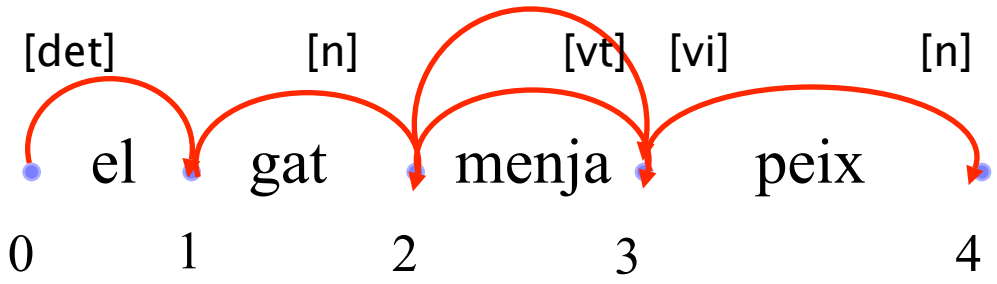
Regla bàsica: Cada vegada que s'afegeix un arc actiu al Chart < i, j, A → a.Bb >, aleshores, per cada regla B → b de la gramàtica, s'ha d'afegir un arc actiu al seu extrem dret < j, j, B →.b >

Inicialització: Igual que abans però a més cal afegir l'**arc actiu** que correspon a l'objectiu d'obtenir una frase.
Ex: <1, 1, oració →.SN SV>

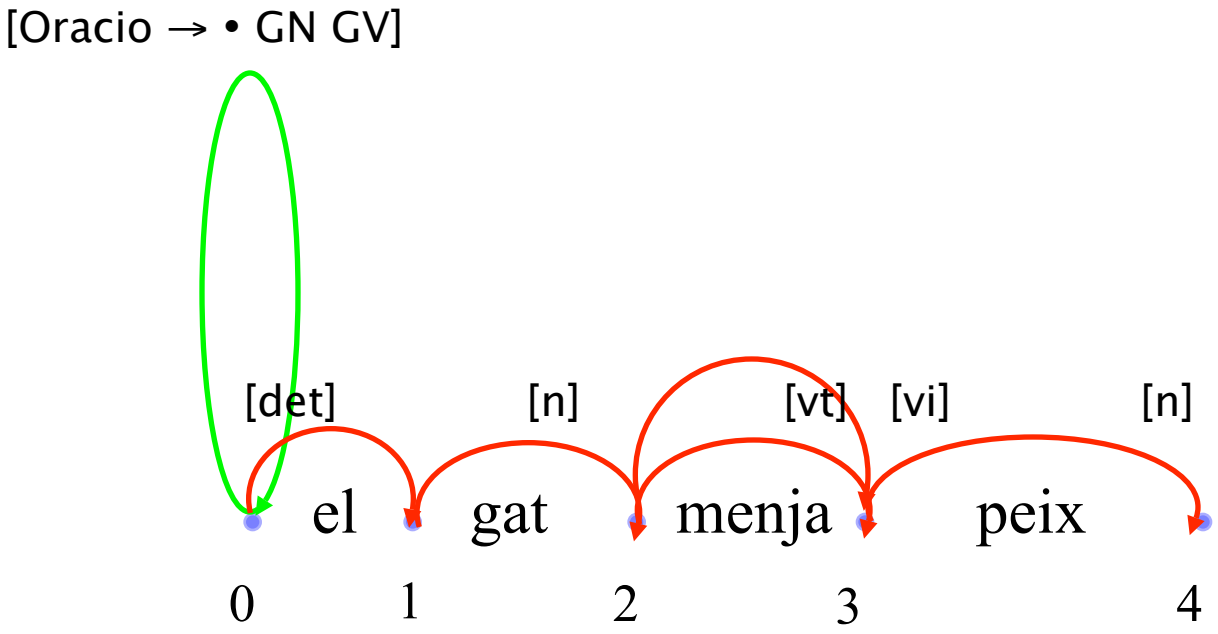
La regla bàsica de combinació amb l'estratègia ascendent o descendent (o una combinació de les dues) és el que ens proporciona el mètode d'anàlisi

Charts: exemple

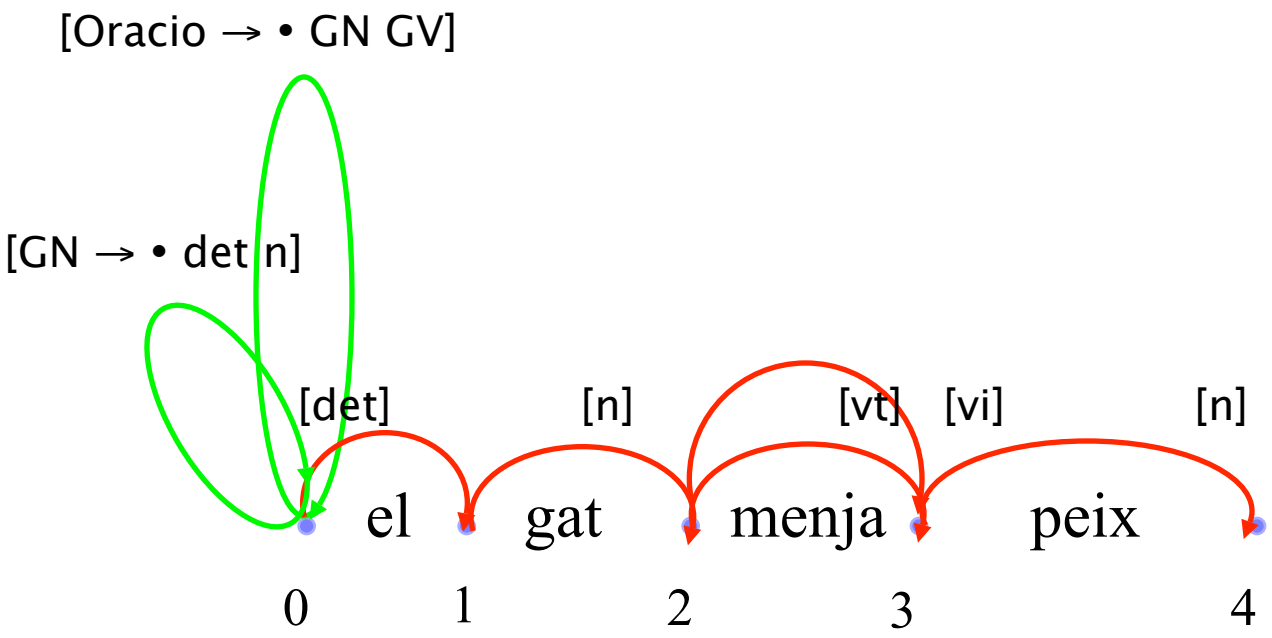
Charts: exemple



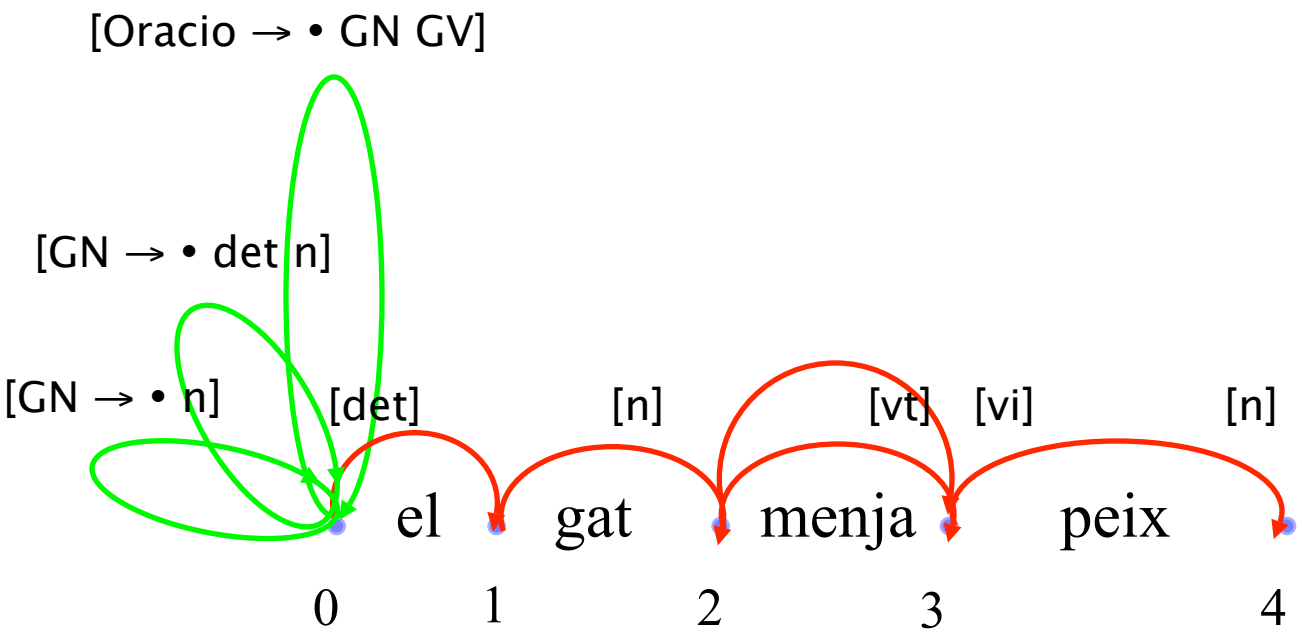
Charts: exemple



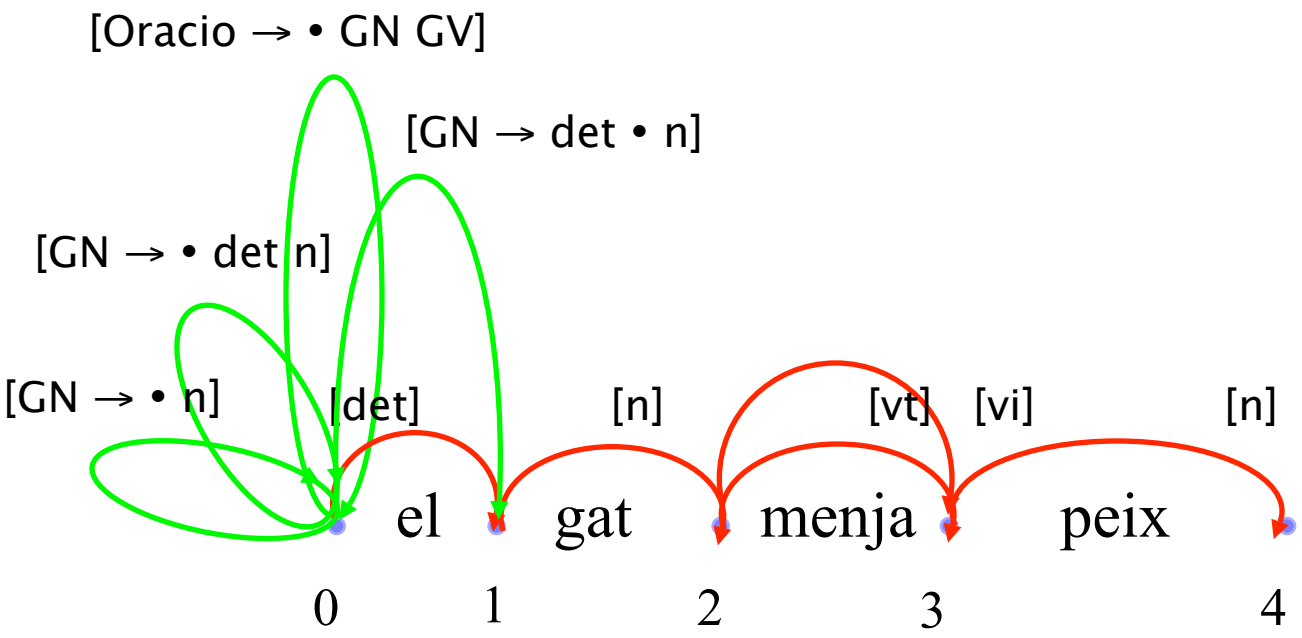
Charts: exemple



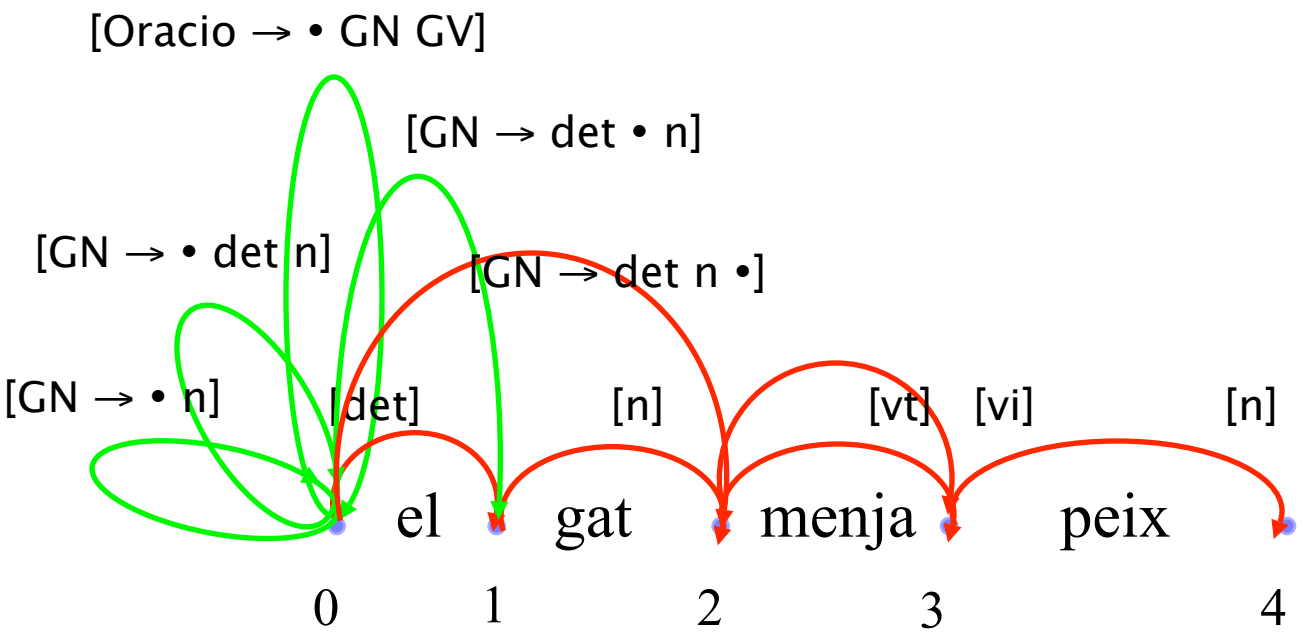
Charts: exemple



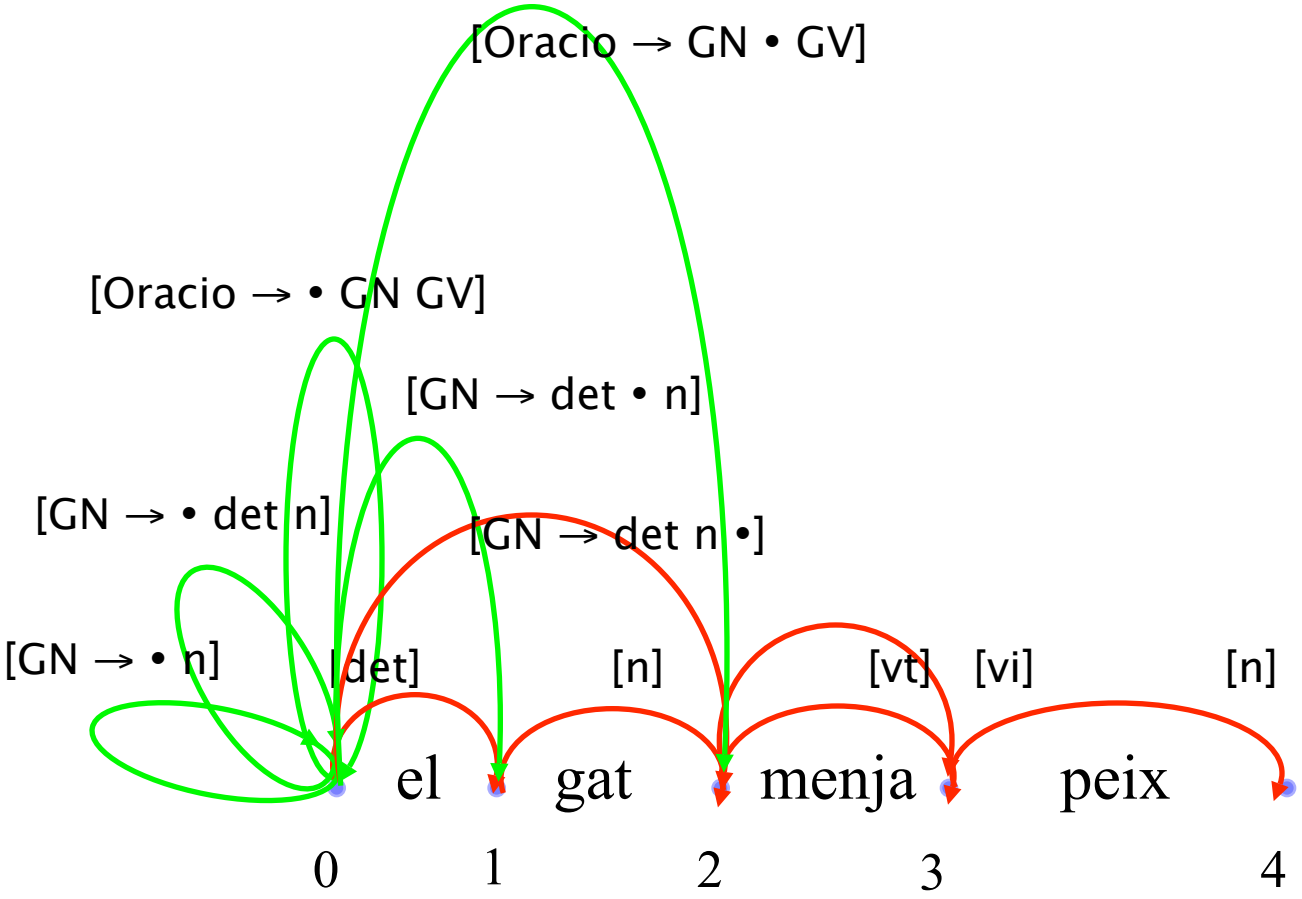
Charts: exemple



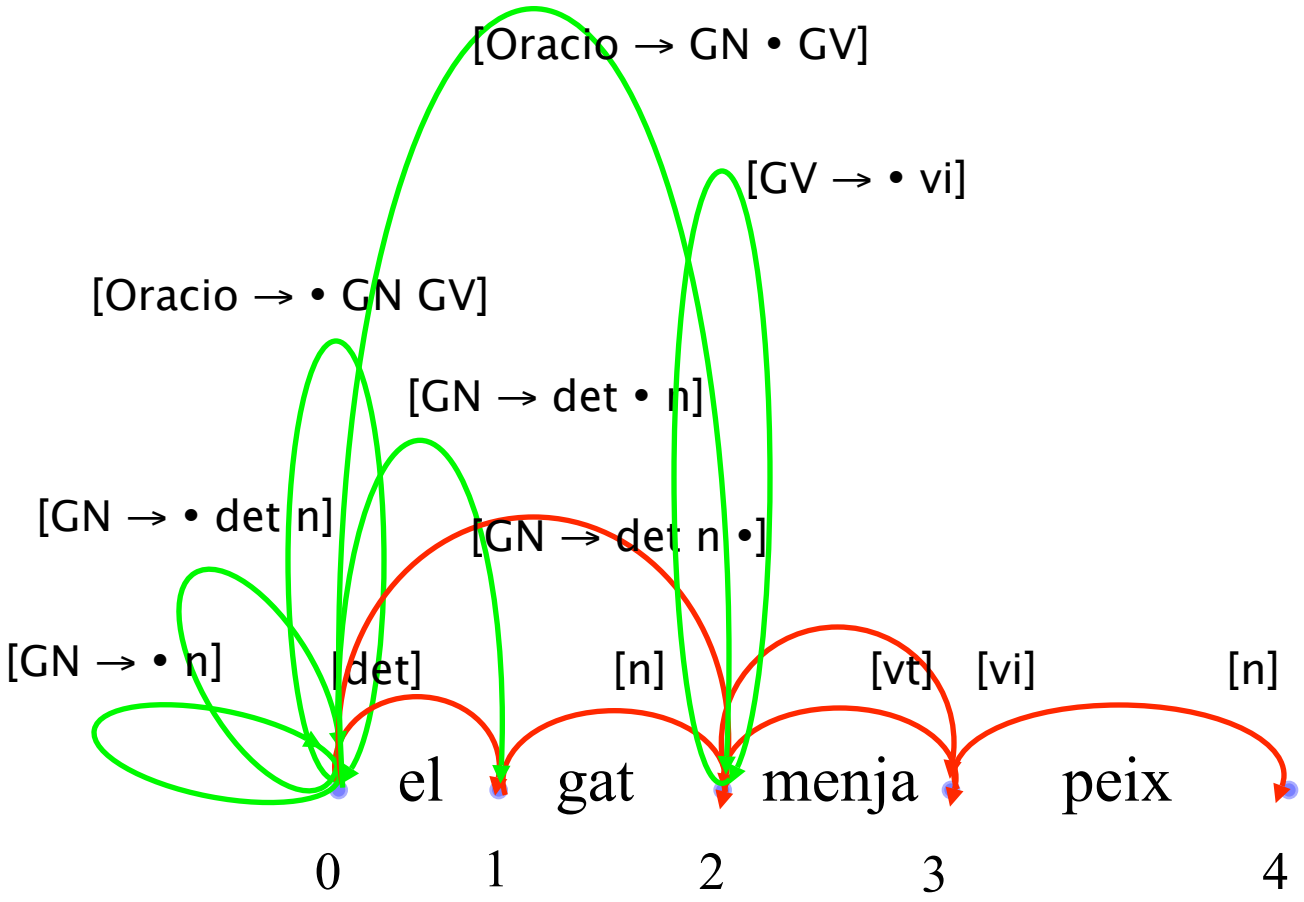
Charts: exemple



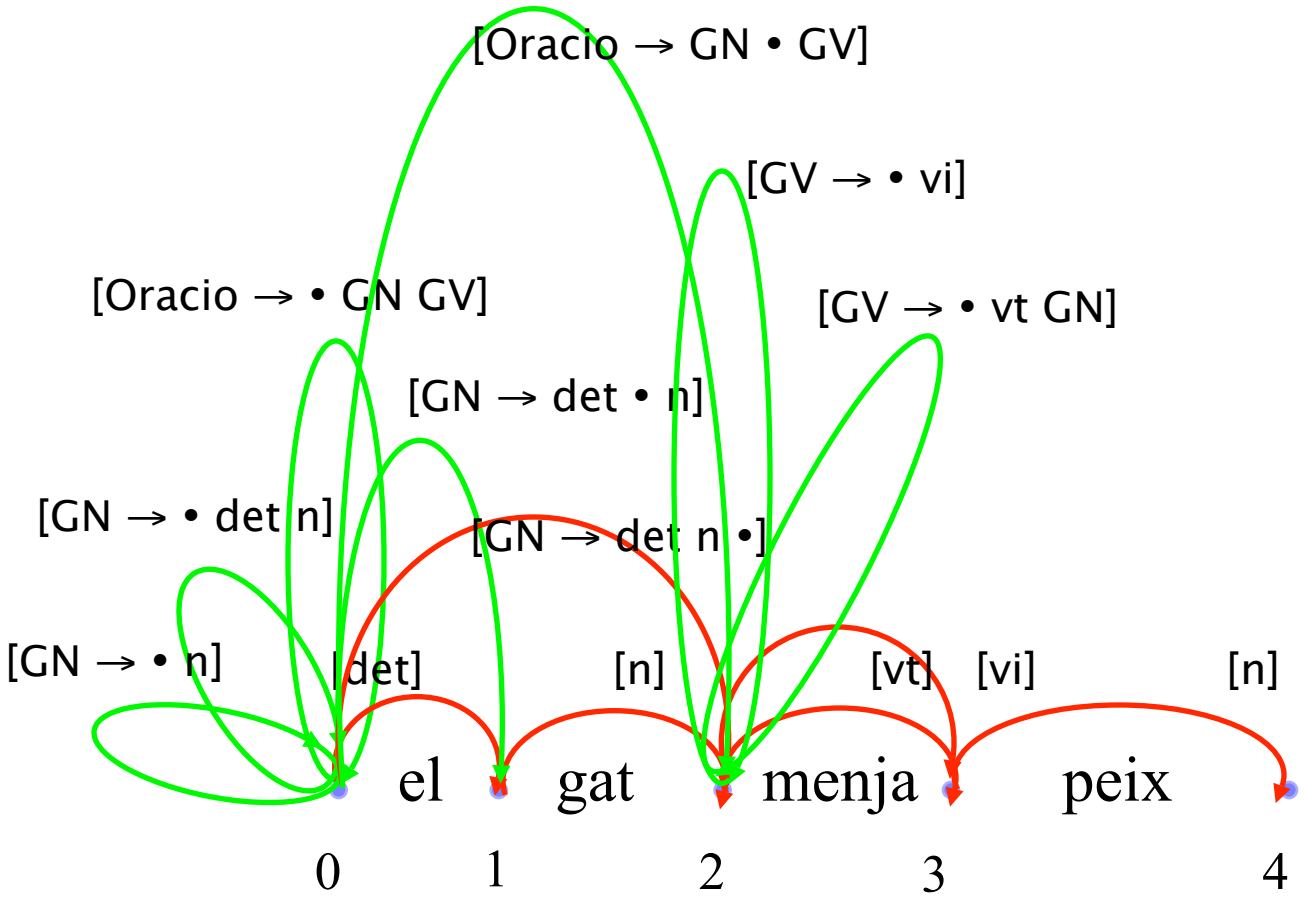
Charts: exemple



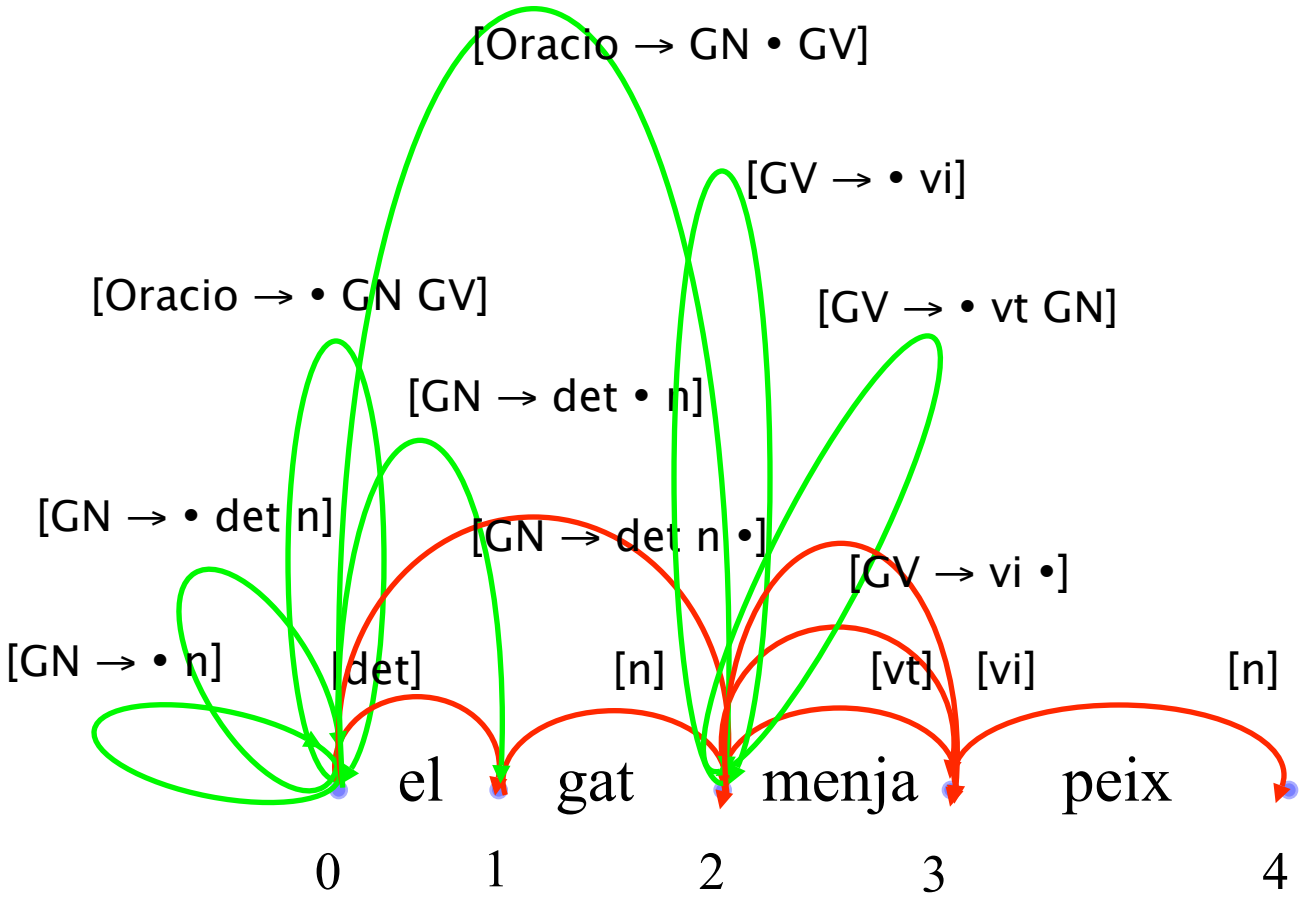
Charts: exemple



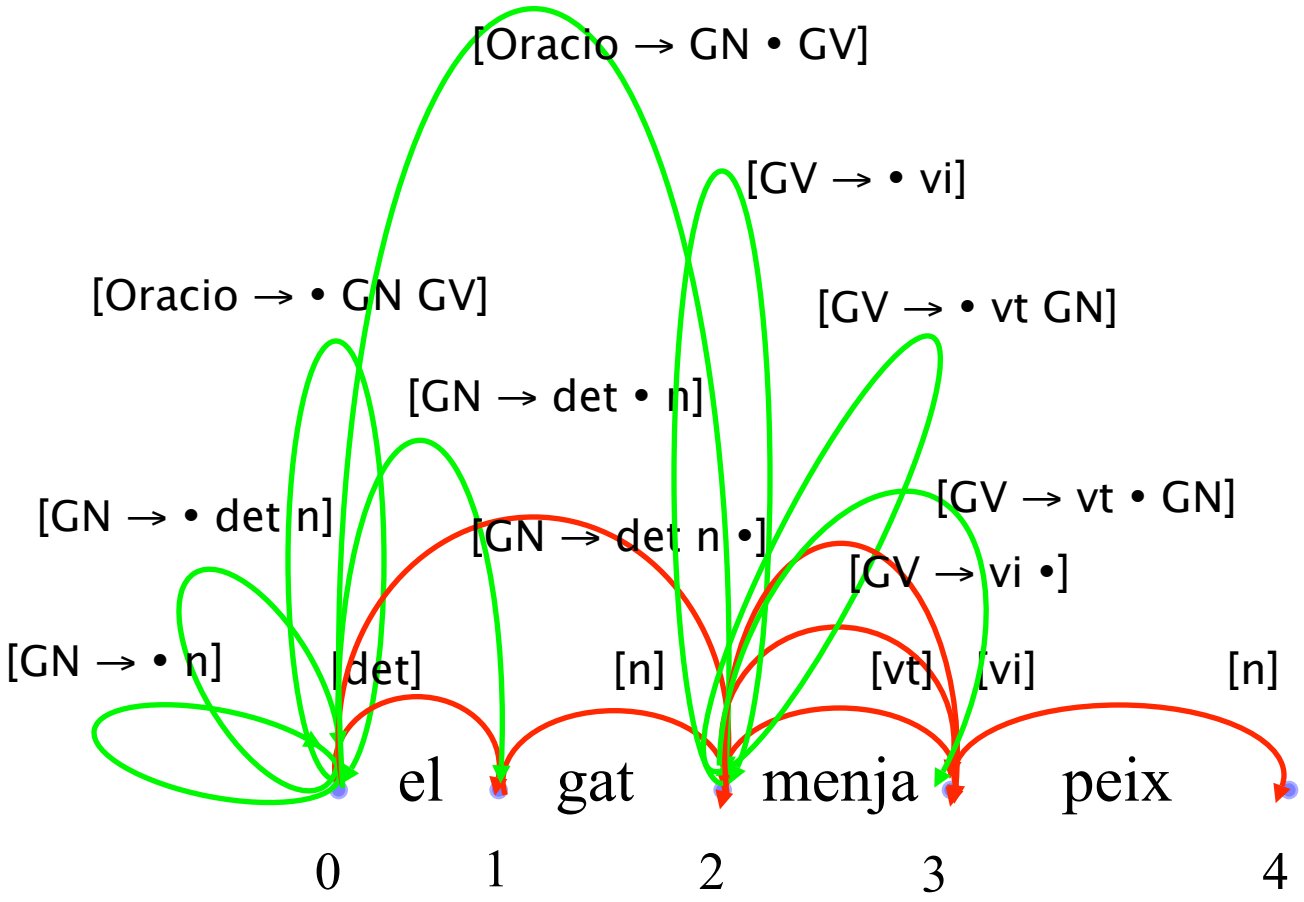
Charts: exemple



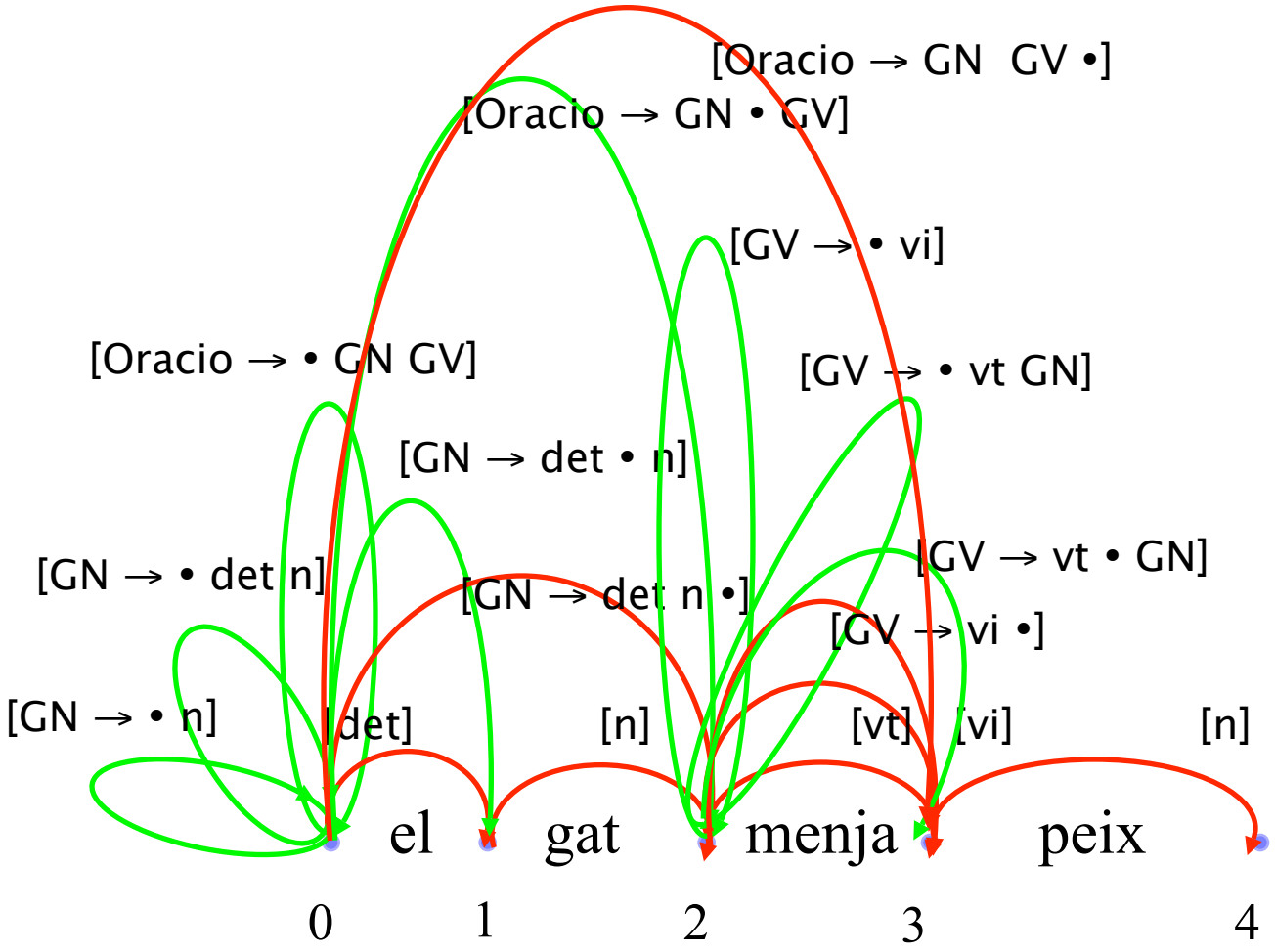
Charts: exemple



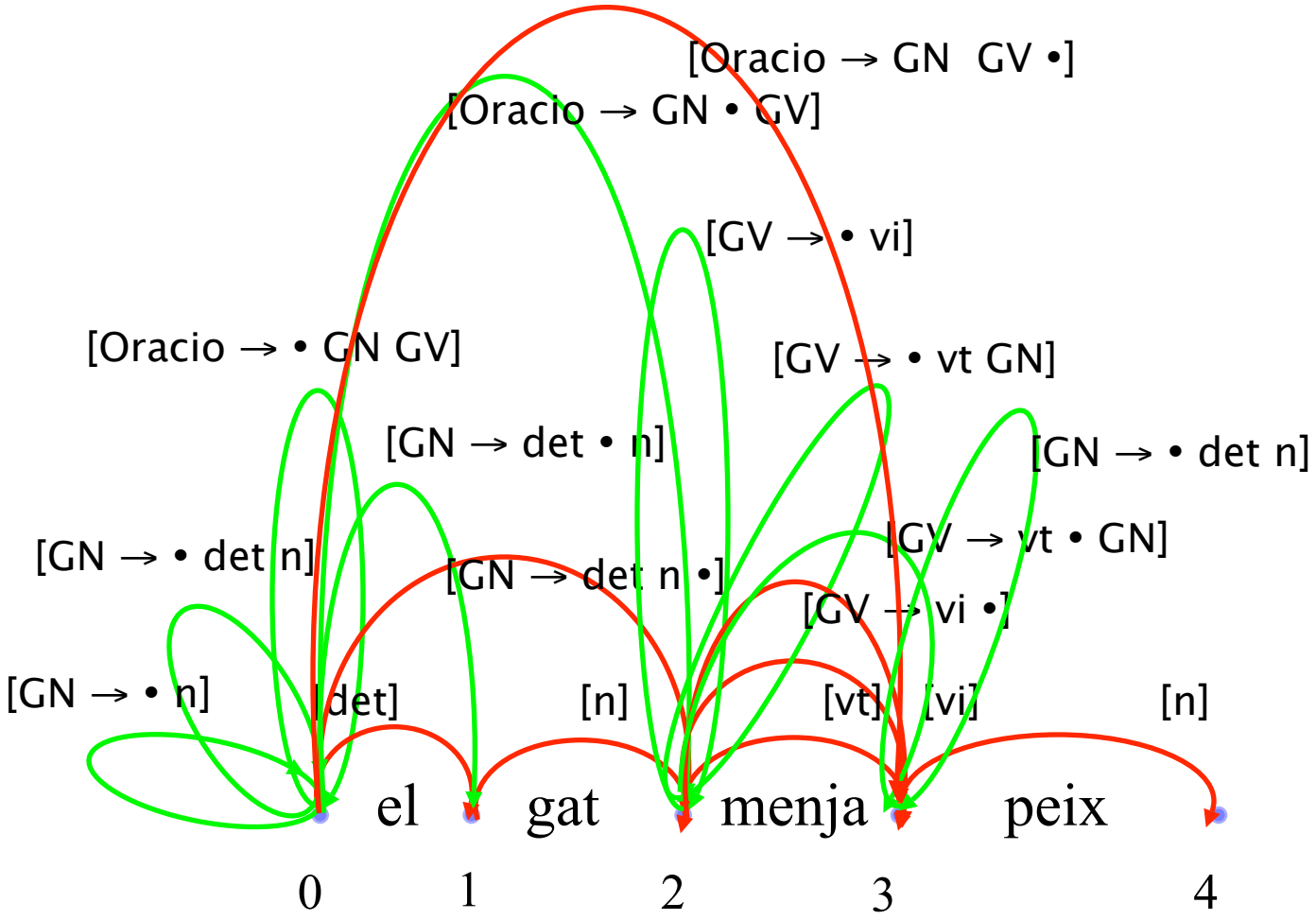
Charts: exemple



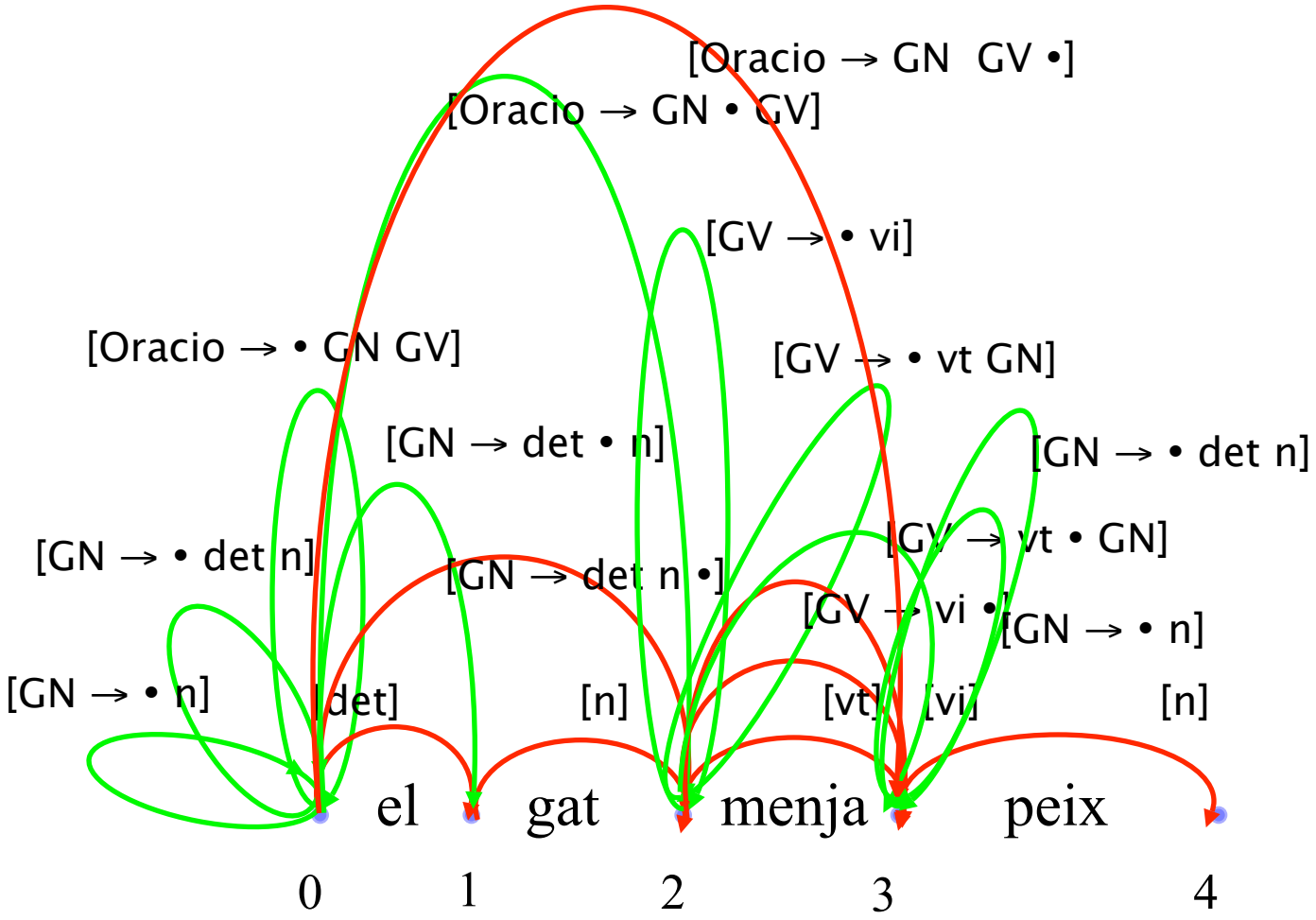
Charts: exemple



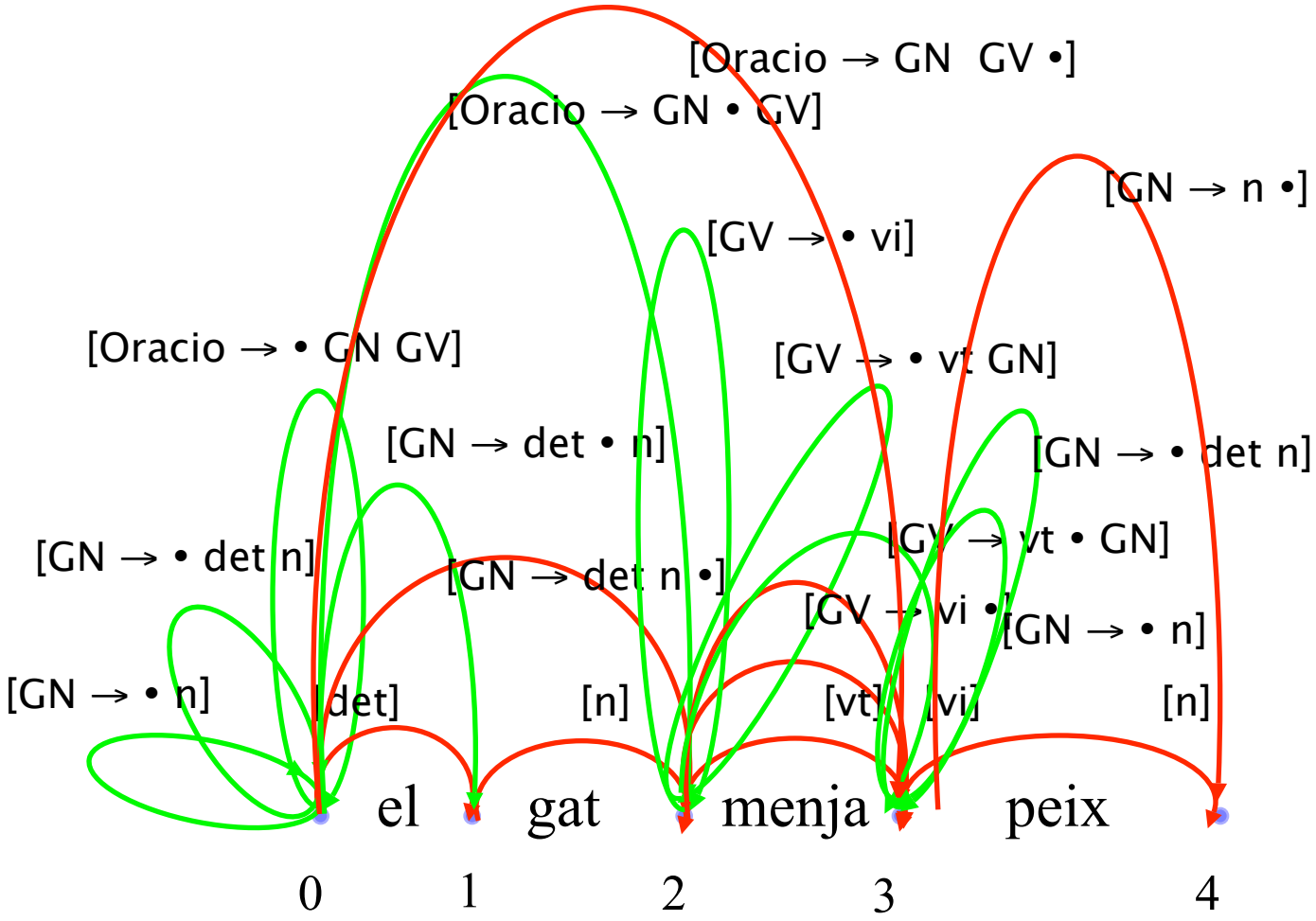
Charts: exemple



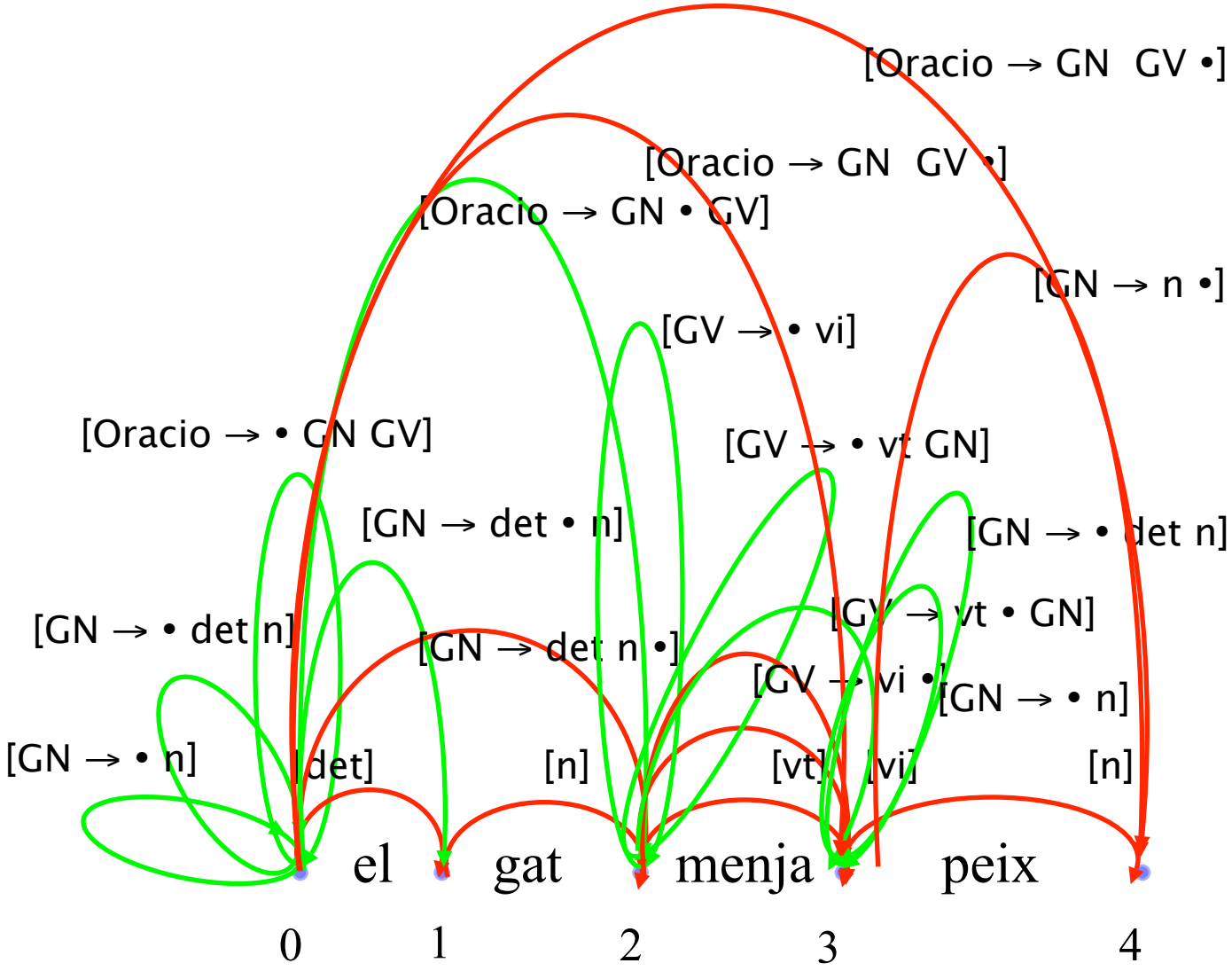
Charts: exemple



Charts: exemple



Charts: exemple



Rasgos (features)

- Context-free grammars provide the basis for most of the computational parsing mechanisms developed to date.
- As they have been described, they would be inconvenient for capturing natural language.
- The basic context-free mechanism can be extended defining constituents by a set of **features**.
- This extension allows aspects of natural language such as agreement and subcategorization to be handled in an intuitive and concise way.

Sistemas de rasgos y gramáticas aumentadas

- In natural language there are often agreement restrictions between words and between phrases.
- For example, the noun phrase (NP) *un hombres* is not correct because the article *un* indicates a single object while the noun *hombres* indicates a plural object.
- The NP does not satisfy the **number agreement** restriction.

Sistemas de rasgos y gramáticas aumentadas

- There are many other forms of agreement, including:
  - subject-verb agreement
  - gender agreement for pronouns
  - restrictions between the head of the phrase and the form of its complement.
- Features are introduced to handle such phenomena.
- A feature NUMBER might be defined that takes a **value** of either *s* (for singular) or *p* (for plural) and we then might write an augmented CFG rule such as

$$NP \rightarrow ART\ \ N \text{ only when } NUMBER_1 \text{ agrees with } NUMBER_2$$

- This rule says that a legal NP consists of an article (ART) followed by a noun (N), but only when the number feature of the first word agrees with the number feature of the second.

Sistemas de rasgos y gramáticas aumentadas

NP → ART N only when NUMBER_1 agrees with NUMBER_2

- This one rule is equivalent to two CFG rules that would use different terminal symbols for encoding singular and plural forms of all NPs:

  NP-SING → ART-SING N-SING

  NP-PLURAL → ART-PLURAL N-PLURAL

- The two approaches seem similar in ease-of-use in this one example.
- Though, in the second case, all rules in the grammar that use an NP on the right-hand side would need to be duplicated to include a rule for NP-SING and a rule for NP-PLURAL, effectively doubling the size of the grammar.

Sistemas de rasgos y gramáticas aumentadas

- Handling additional features, such as person agreement, would double the size of the grammar again, and so on.
- Using features, the size of the augmented grammar remains the same as the original one, yet accounts for agreement constraints.
- A constituent is defined as a feature structure (FS), a mapping from features to values that defines the relevant properties of the constituent.

Sistemas de rasgos y gramáticas aumentadas

- Example: a FS for a constituent ART1 that represents a particular use of the word *un*:

ART1: (**CAT** ART

ROOT un

NUMBER s)

- It is a constituent in the category ART that has its root in the word *un* and is singular.
- Usually an abbreviation is used that gives the CAT value more prominence:

ART1: (ART **ROOT** un **NUMBER** s)

Sistemas de rasgos y gramáticas aumentadas

- FSs can be used to represent larger constituents as well.
- FSs themselves can occur as values.
- Special features based on the integers (1, 2, 3…) stand for the subconstituents (first, second…).
- The representation of the NP constituent for the phrase *un pez* could be:

NP1: (NP **NUMBER** s

1 (ART **ROOT** un

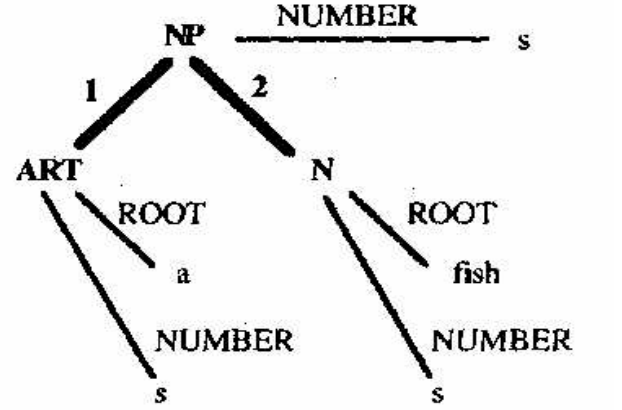
NUMBER s)

2 (N **ROOT** pez

NUMBER s))

Sistemas de rasgos y gramáticas aumentadas

- The previous one can also be viewed as a representation of a parse tree:



- Subconstituent features 1 and 2 correspond to the subconstituent links in the tree.

Sistemas de rasgos y gramáticas aumentadas

- The rules in an augmented grammar are stated in terms of FSs rather than simple categories.
- Variables are allowed as feature values so that a rule can apply to a wide range of situations.
- For example, a rule for simple NPs would be as follows:

  (NP **NUMBER** ?n): (ART **NUMBER** ?n) (N **NUMBER** ?n)

- This says that an NP constituent can consist of two subconstituents, the first being an ART and the second being an N, in which the NUMBER feature in all three constituents is identical.

Sistemas de rasgos y gramáticas aumentadas

(NP **NUMBER** ?n): (ART **NUMBER** ?n) (N **NUMBER** ?n)

- According to this rule, constituent NP1 given previously is a legal constituent.
- The constituent

*(NP **1** (ART **NUMBER** s)

2 (N **NUMBER** s))

is not allowed by this rule because...

Sistemas de rasgos y gramáticas aumentadas

(NP **NUMBER** ?n): (ART **NUMBER** ?n) (N **NUMBER** ?n)

- According to this rule, constituent NP1 given previously is a legal constituent.
- The constituent

*(NP **1** (ART **NUMBER** s)

2 (N **NUMBER** s))

is not allowed by this rule because there is no NUMBER feature in the NP.

Sistemas de rasgos y gramáticas aumentadas

(NP **NUMBER** ?n): (ART **NUMBER** ?n) (N **NUMBER** ?n)

- The constituent

*(NP **NUMBER** s

1 (ART **NUMBER** s)

2 (N **NUMBER** p))

is not allowed because...

Sistemas de rasgos y gramáticas aumentadas

(NP **NUMBER** ?n): (ART **NUMBER** ?n) (N **NUMBER** ?n)

- The constituent

*(NP **NUMBER** s

1 (ART **NUMBER** s)

2 (N **NUMBER** p))

is not allowed because the NUMBER feature of the N constituent is not identical to the other two NUMBER features.

Sistemas de rasgos y gramáticas aumentadas

- Variables are also useful in specifying ambiguity in a constituent.
- The word *crisis* is ambiguous between a singular and a plural reading.
- The word might have two entries in the lexicon that differ only by the value of the NUMBER feature.
- Alternatively, we could define a single entry that uses a variable as the value of the NUMBER feature:

  (N **ROOT** crisis **NUMBER** ?n)

- This works because any value of the NUMBER feature is allowed for the word *crisis*.

Sistemas de rasgos y gramáticas aumentadas

- In many cases not just any value would work, but a range of values is possible.
- We introduce **constrained variables**, which are variables that can only take a value out of a specified list.
- For example, the variable ?n{s p} would be a variable that can take the value s or the value p.
- When we write such variables, we will drop the variable name altogether and just list the possible values.
- The word *crisis* might be represented by the constituent:

(N **ROOT** crisis **NUMBER** ?n{s p})

or more simply as

(N **ROOT** crisis **NUMBER** {s p})

Sistemas de rasgos y gramáticas aumentadas

- There is an interesting issue of whether an **augmented CFG** can describe languages that cannot be described by a simple CFG.
- The answer depends on the constraints on what can be a feature value.
- If the set of feature values is finite, then it would always be possible to create new constituent categories for every combination of features. Thus it is expressively equivalent to a CFG.
- If the set of feature values is unconstrained then such grammars have arbitrary computational power.
- In practice, even when the set of values is not explicitly restricted, this power is not used, and the standard parsing algorithms can be used on grammars that include features.

Formalismos de unificación

- La unificación se usa como mecanismo básico de composición entre constituyentes en **gramáticas lógicas**.
- Història:
  - Q-Systems (Colmerauer, 1972)
  - Prolog (Colmerauer, 1973)
  - Gramàtiques de Metamorfosi (Colmerauer, 1975)
  - Gramàtiques de Clàusules Definides (DCGs) (Pereira, Warren, 1980)

Análisis gramatical con unificación

- Gramática

  (1) oració (X,Y) :- gnom(X,Z), gver(Z,Y)

  (2) gnom(X,Y) :- art(X,Z), nom(Z,Y)

  (3) gver(X,Y) :- ver(X,Y)

- Lexicón

  (4) art(X,Y) :- el(X,Y)

  (5) nom(X,Y) :- gos(X,Y)

  (6) ver(X,Y) :- borda(X,Y)

Análisis gramatical con unificación

(7) el(1,2)

(8) gos(2,3)

(9) borda(3,4)

Frase a analizar: **oració(1,4)**

Análisis gramatical con unificación

oració(1,4)

(R1) (X↔1,Y↔4) por unificación

gnom (1,Z), gver(Z,4)

(R2) aplicada a gnom(1,Z)

art(1,U), nom(U,Z), gver(Z,4)

Análisis gramatical con unificación

art(1,U), nom(U,Z), gver(Z,4)

(R4) aplicada a art(1,U)

el(1,U), nom(U,Z), gver(Z,4)

(R7) (U ↔ 2)

el(1,2), nom(2,Z), gver(Z,4)

nom(2,Z), gver(Z,4)

Análisis gramatical con unificación

nom(2,Z), gver(Z,4)

(R5)

gos(2,Z), gver(Z,4)

(R8) (Z ↔ 3)

gos(2,3), gver(3,4)

gver(3,4)

Análisis gramatical con unificación

gver(3,4)

(R3)

ver(3,4)

(R6)

borda(3,4)

(R9): Fin

Análisis semántico

- Consiste en construir una representación de las frases en algún sistema formal.
- En general es un problema intratable.
- Se simplifica suponiendo que la semántica de una frase se pueda construir a partir de la semántica de sus partes: **semántica compositiva**.
- Algunas características del lenguaje se tienen que tratar a parte: **referencias, omisiones, contexto**…

Estrategias de análisis

- Dos maneras de plantear la interpretación semántica:
  - Secuencial (sintáctica → semántica)
  - Paralela (sintáctica + semántica)

Interpretación secuencial

- Problemas de ambigüedad
  - Es posible que haya más de una interpretación sintáctica.
  - Hay que considerarlas todas, para poder comprobar sucesivamente cuáles son las semánticamente posibles.
- Principal ventaja
  - El análisis semántico parte de un análisis sintáctico correcto.

Interpretación paralela

- Principal problema
  - No se sabe si la interpretación sintáctica es correcta hasta el final.
- Principal ventaja
  - Poder descartar interpretaciones sintácticas correctas (o parcialmente correctas) que no tengan interpretación semántica asociada.
- Las reglas sintácticas incluyen **la información semántica asociada**
  - Se obtiene como resultado un árbol de análisis y una o varias interpretaciones.

Sistema de representación

- El sistema de representación tiene que permitir:
  - Manejar cuantificación, predicación, negación, modalidad (creencias)
  - Resolver la ambigüedad tanto léxica (polisemia) como sintáctica
  - Manejar inferencias (herencia, razonamiento por omisión)
    - Importante a la hora de resolver problemas que involucren el contexto o el conocimiento del dominio.

Sistema de representación

- Los sistemas que no se basan en la sintaxis para la interpretación suelen utilizar sistemas tipo ontologías.
- Por lo general, se utiliza una variedad de la lógica de primer orden adecuada al dominio de aplicación.
- El elemento básico de representación es el lexema: raíz de un grupo de palabras que son diferentes formas de "la misma palabra" (ej.: ir, ido, yendo).

Sistema de representación

- Ejemplos
  - Los nombres propios corresponden a constantes.
  - Los verbos intransitivos a predicados unarios:

    Juan ríe ríe(juan)
  - Los verbos transitivos a predicados de aridad superior:

    Juan lee el Quijote lee(juan, quijote)
  - Los nombres genéricos a predicados sobre variables:

    El hombre hombre(X)
  - Los adjetivos a predicados unarios:

    La casa grande grande(X) $\wedge$ casa(X)

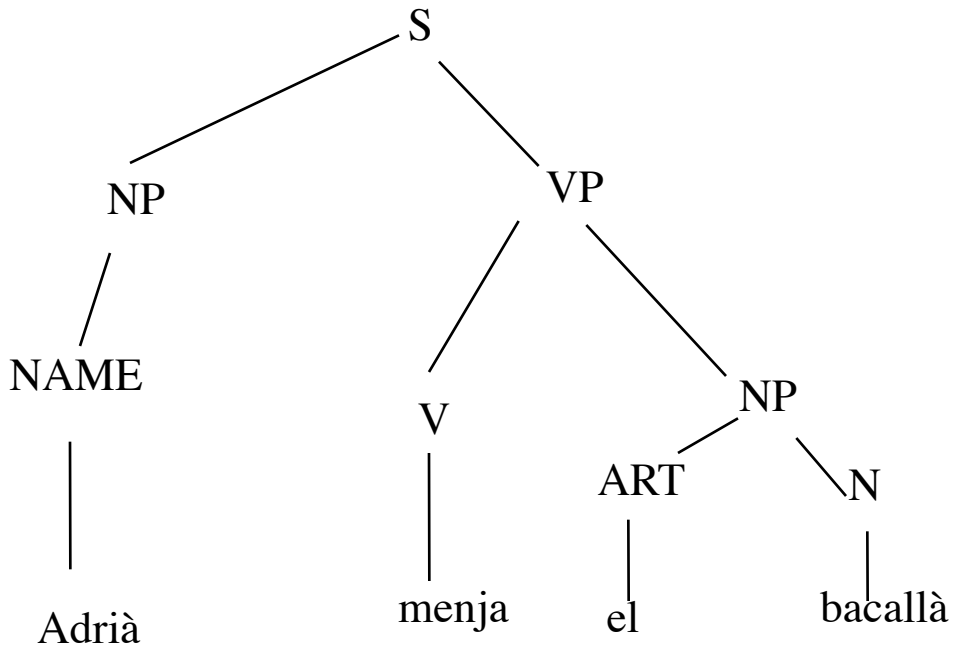
Sistema de representación

La representación consiste normalmente en un árbol de análisis y una función de composición que construye la interpretación de los elementos asociados.

$sem_S = f_S (sem_{NP} , sem_{VP})$

$sem_{NP} = f_{SN} (sem_{ART} , sem_N)$

$sem_{ART} = f_{ART} (sem\ el)$



La representación del significado: ejemplo

entrada:

Qui dirigeix el PSOE?

forma lògica:

```
(pregunta
   (referent (X))
   (X instancia (X, persona)
        (el1 (Y instancia(Y, partit_polític)
                   nom(Y, "PSOE"))
              (Z  instancia(Z, dirigir)
                     present(Z)
                     valor_prop(Z, agent, X)
                     valor_prop(Z, pacient,Y)))))
```

Niveles de análisis

- Pragmàtic
  - Interpretació dins un context (incorpora informació implícita)
  - Relacionar amb la resta del discurs
  - ***"L'avió va detectar el banc"***
  - ***"El gat vell"***
- Referencias implícitas (nivel pragmático)
  - "*Le dio un libro*"
  - "*No les gustó*"

Niveles de análisis

- Il·locutiu
  - Detecció de les intencions de qui profereix la frase
  - ***"Els plats estan bruts"***
    - Es tracta d'una frase declarativa neutra?
    - És una invitació a l'acció?
      ("renta'ls!")
    - És un retret?
      ("sempre els deixes bruts i em toca rentar-los a mi")
- Problemas de asignación de intenciones (nivel ilocutivo)
  - "*Los platos están sucios*" (por tanto, ¡lávalos!)