

Introducció al Processament del Llenguatge Natural (PLN)

L. Màrquez i H. Rodríguez

Objectius del temari

- Conèixer l'àmbit del PLN i les seves principals aplicacions
- Comprendre la problemàtica associada a la comprensió del llenguatge natural i els nivells d'anàlisi sintàctic i semàntic
- A nivell pràctic, coneixement bàsic de la programació d'analitzadors amb DCG's

Objectius i àmbits del PLN

- L'ús del llenguatge per expressar-nos i comunicar-nos amb els altres és una de les capacitats més importants dels éssers humans.
- L'**objectiu** del Processament del Llenguatge Natural (PLN) és construir sistemes computacionals capaços de **comprendre** i/o **generar** llenguatge humà en totes les seves formes

Objectius i àmbits del PLN

- Per arribar a aquest objectiu cal:
 - Saber com els humans **generen** expressions correctes i comprensibles pels altres
 - Conèixer com els humans **comprenen** expressions d'altres humans
 - Ser capaços de formalitzar els coneixements i els processos necessaris de manera que siguin tractables per un sistema computacional

Multidisciplinarietat

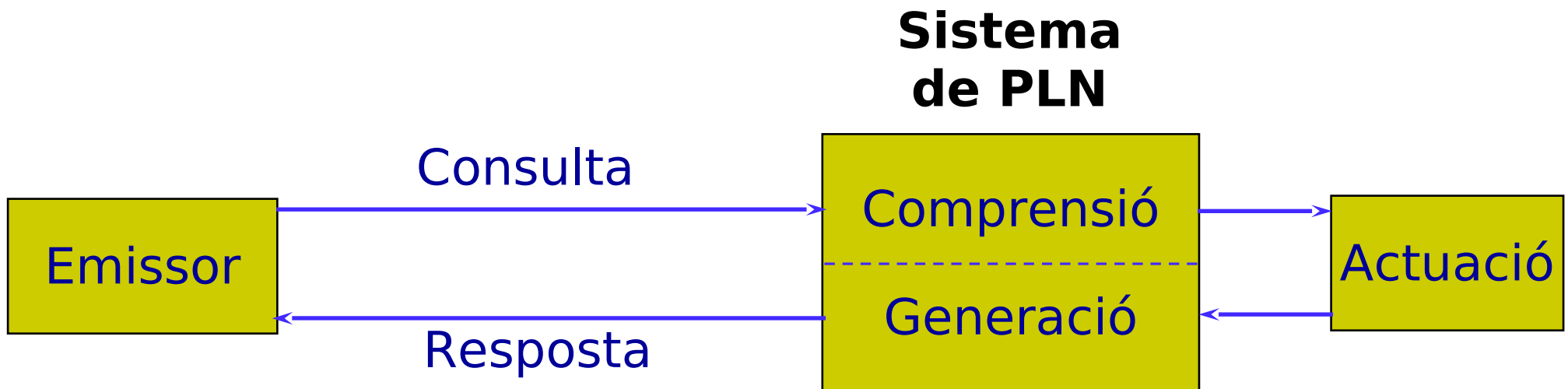
Disciplines associades al PLN:

- Lingüística
 - Lingüística computacional
- Teoria de llenguatges formals
 - Compiladors
- Intel·ligència Artificial
 - Representació del coneixement
 - Aprenentatge
 - Raonament

Comprensió/Generació

- Les dues operacions bàsiques

<Interfícies en llenguatge natural>

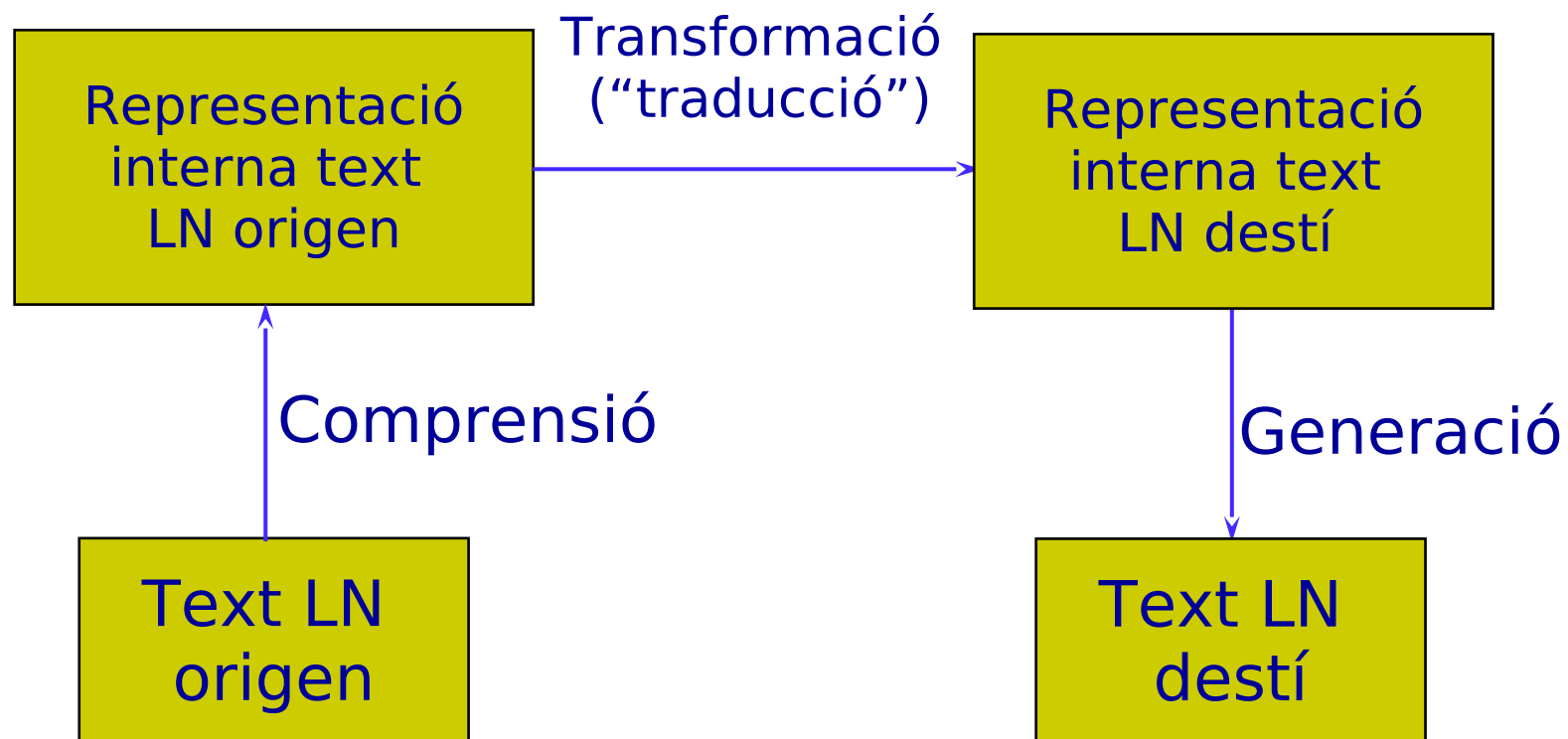


- La comprensió de la consulta i la generació de la resposta poden ser *orals: speech recognition/synthesis*

Comprensió/Generació

- Traducció automàtica

<Traducció automàtica: model de *transfer*>



- La traducció també es pot fer a partir d'intervencions orals

Què és comprendre?

“Comprendre alguna cosa consisteix a transformar-la d’una representació a una altra de tal manera que aquesta segona representació correspongui a un conjunt d’accions que es podrien efectuar i la transformació assegura que per cada element a comprendre s’efectuarà l’acció adequada”

(Rich, 1991)

Comprendre LN

- Extreure sentit d'un text per tal d'efectuar les accions corresponents
 - Consulta (*query*) a una base de dades
 - Actuadors
 - Resumir un text
 - Traduir un text

Com comprendre LN?

- La comprensió exigeix
 - Extreure significat individual de les paraules
 - Extreure significat de les relacions entre paraules
 - Referir el significat literal al context d'actuació del sistema
 - Metàfores, retòrica, ironia, entonació
- Eines
 - Anàlisi de les components del llenguatge a diversos nivells

Informació per a l'anàlisi

- Informació lèxica, sintàctica, semàntica
- Exemple:

“Em parlarà sens dubte de la reestructuració urbana a Barcelona”

- Cal detectar:
 - Paraules individuals amb significat i connectives:
Barcelona, reestructuració, urbana, de, la, parlarà, etc.
 - Acumular informació per a saber el seu paper dins la frase i establir possibles significats:
 - Categoria morfosintàctica: nom, nom propi, nom compost, verb, article, etc.
 - Informació sobre la relació entre significats per a establir el significat global
 - Paper sintàctic: subjecte, objecte directe, etc.

Nivells d'anàlisi

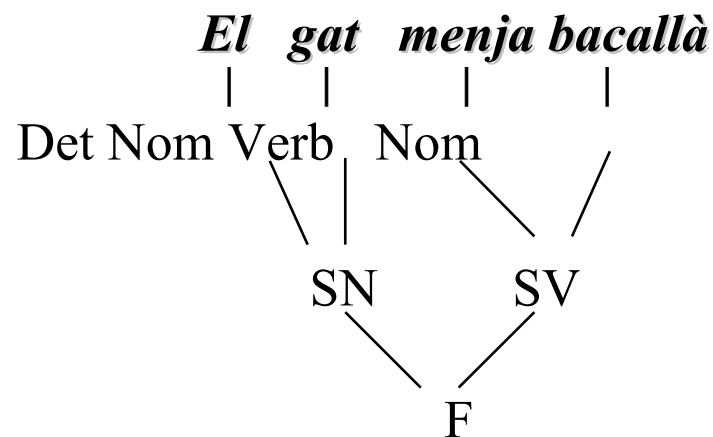
- Fonològic
 - Tractament dels sons per tal de detectar unitats d'expressió
- Textual
 - Segmentació del text en unitats tractables (paràgrafs, oracions, etc.)
 - Localització (identificació) de paraules o unitats lèxiques
- Morfològic
 - Formació de les paraules considerant la flexió, la derivació, la composició, etc.
 - L'anàlisi morfològica persegueix acumular informació per ajudar a establir les arrels i afixes (prefixes, sufixes, infixes) de les paraules
 - La paraula com a composició de morfemes
- Lèxic
 - Consideració de la paraula com a unitat de significat dins un text.
 - Obtenció d'informació lèxico-semàntica (ontologies, diccionaris semàntics)

Nivells d'anàlisi (2)

- Sintàctic (1)
 - Detecció d'estructures sintàcticament vàlides

Els gat menja bacallà
(* *Els gata menja bacallà*)

- Sintàctic (2)
 - Extracció i representació de les estructures sintàcticament vàlides



Nivells d'anàlisi (3)

- Lògic
 - Extracció del significat literal de la oració
 - Representació del significat mitjançant CP_1 , frames, xarxes semàntiques, etc.
 - En el cas de CP_1 : expressió en termes de variables, predicats, funcions, constants, connectives lògiques, quantificadors, etc.

“El gat menja bacallà”

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

Nivells d'anàlisi (5)

- 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
 - Ex:
 - El gat és un felí,
 - el bacallà és un peix comestible,
 - l'actor de menjar ha de ser un ésser viu,
 - ...

Nivells d'anàlisi (6)

- Pragmàtic
 - Interpretació dins un context (incorpora informació implícita)
 - Ex: *“L’avió va detectar el banc”*
- Il.locutiu
 - Detecció de les intencions de qui profereix la frase
 - Ex: *“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 mí”)

Problemàtica del LN (1)

- Ambigüitat lèxica
 - “*Va puxar la roda del davant*”
 - “roda” pot ser nom o verb (*POS tagging*)
 - “*Va veure el banc*”
 - Moble per a seure? Entitat financera? Banc de peixos? (*WSD*)
- Ambigüitat sintàctica
 - “*Va veure un home dalt de la muntanya amb uns prismàtics*”
 - “*El venedor de diaris del barri...*” (*PP-attachment*)

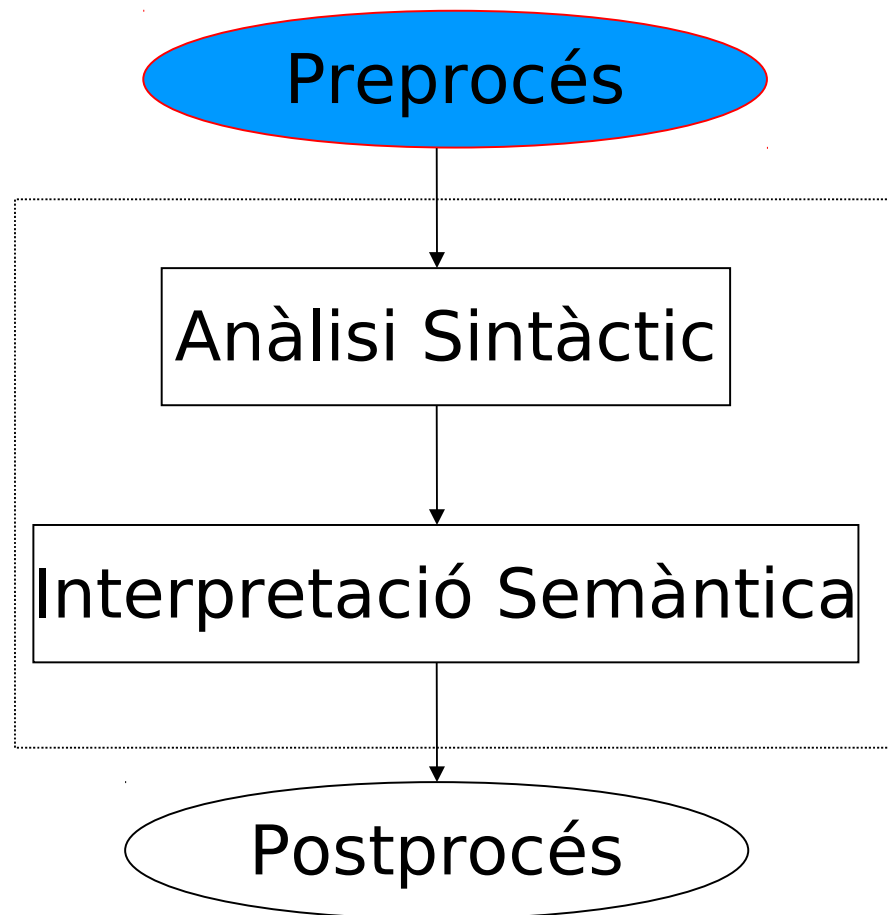
Problemàtica del LN (2)

- Ambigüitat semàntica
 - “*Li va donar un pastís als nens*”
 - 1 a tots?, 1 a cadascun? (àmbit quantificació)
 - “*Les idees verdes dormen furiosament*” (Chomsky)
- Referències, el.lipsis (nivell pragmàtic)
 - “*Li va donar un llibre*” (...) “*No li va agradar*”
- Il.locució: problema d’assignació d’intencions
 - “*Els plats estan bruts*”
 - (per tant, renta’ls?)

Plantejament de l'Anàlisi

- La resolució d'algunes ambigüitats necessita la **col.laboració entre diversos nivells** d'informació (un analitzador per nivell?): *més coneixement de context*
- Cooperació entre analitzadors:
 - Estratificat: Seqüencial / en cascada
 - Global (en paral.lel)

Processament del LN



“Quina es la capital de França ?”

...

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)

Anàlisi Textual

- Detecció d'unitats tractables: **paràgrafs i oracions**
 - Mètodes simples,
 - basats en localitzar marques d e puntuació:
 - “.”, “?”, “!”, ”...”, etc.
 - Problema: sigles, inicials, etc.
 - Mètodes basats en tècniques d'Aprenentatge Automàtic (classificació)
 - Tenen en compte més informació contextual

Anàlisi Lèxica: objectius

- Detectar paraules (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, etc.)
 - Reconeixement i classificació de noms propis i entitats

Problemàtica de l'anàlisi lèxica (1)

- **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 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**”)

Problemàtica de l'anàlisi lèxica (2)

- **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. Peris 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**

Informació necessària

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

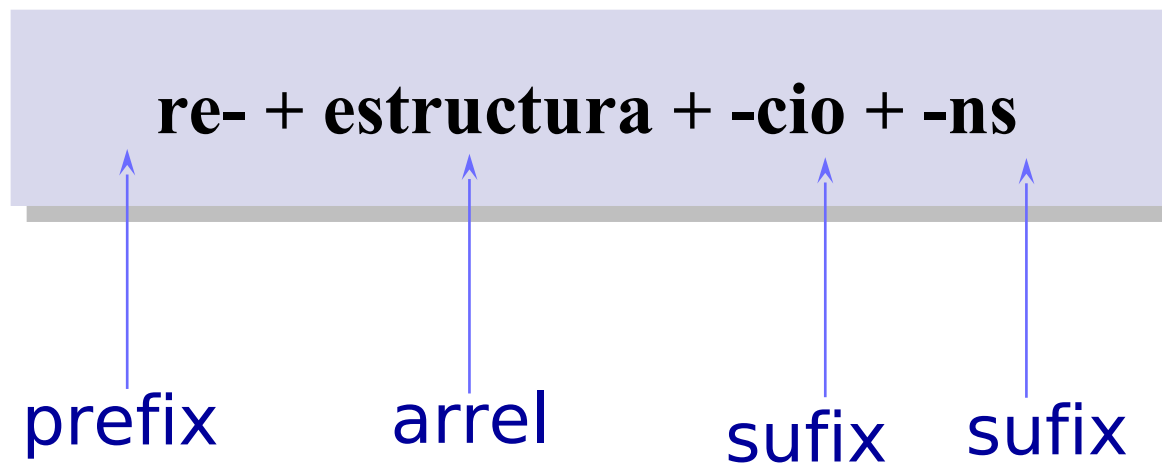
| Lexema | Informació |
|--------|--|
| cant- | cantar -o/-es/-a/-em/-eu/-en V / Infinitiu |

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..)

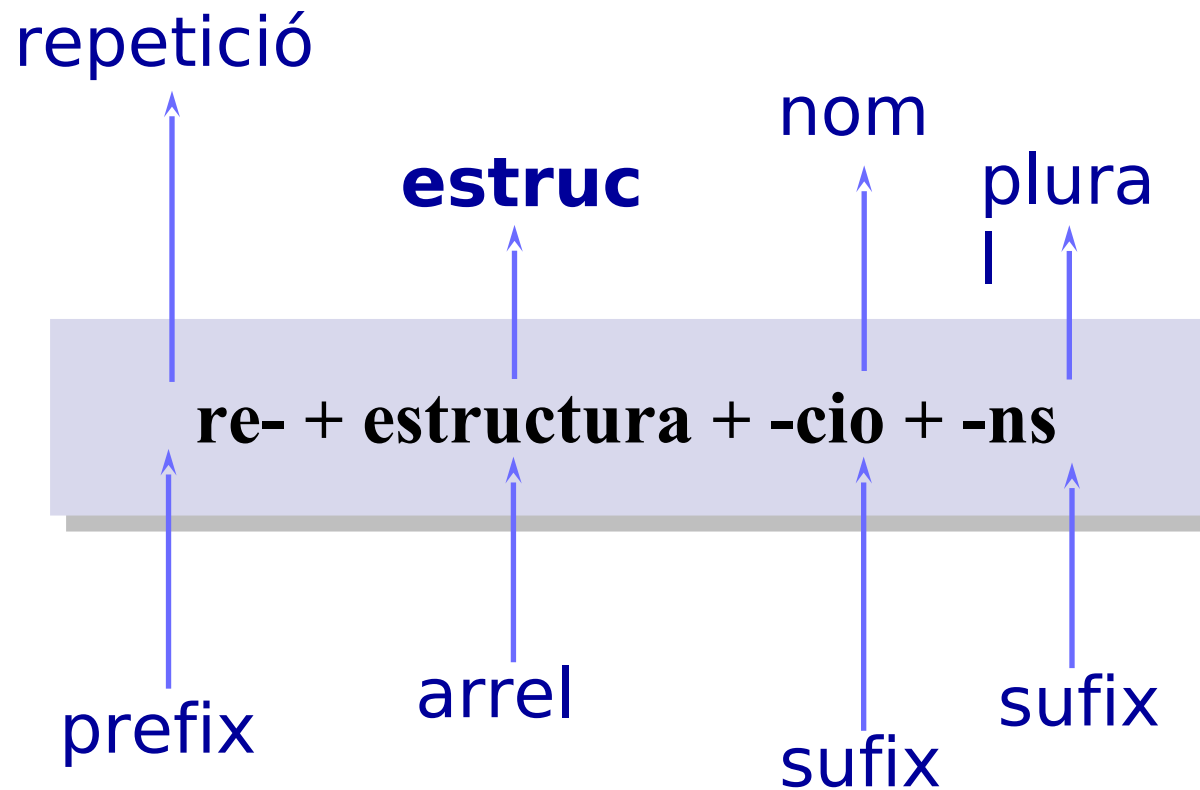
Problemàtica: representació (2)

- 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



Anàlisi Morfològica

- Versió simple: utilització de **formaris** (llista de formes amb informació morfològica i els lemes corresponents)
- **Analitzadors morfològics:**
 - Dicionaris de morfemes:
 - diccionari d'arrels, de sufixes, prefixes, etc.
 - Morfotàctica: regles de combinació de morfemes
 - Variacions fonològiques: canvis al combinar els morfemes
- Tipus d'**analitzadors**
 - 1 nivell: FSA
 - 2 nivells: FST
 - més de dos nivells: Cascada de FSTs

Resultat del Preprocés (lèxic/morfològic)

“Quina es la capital de França ?”

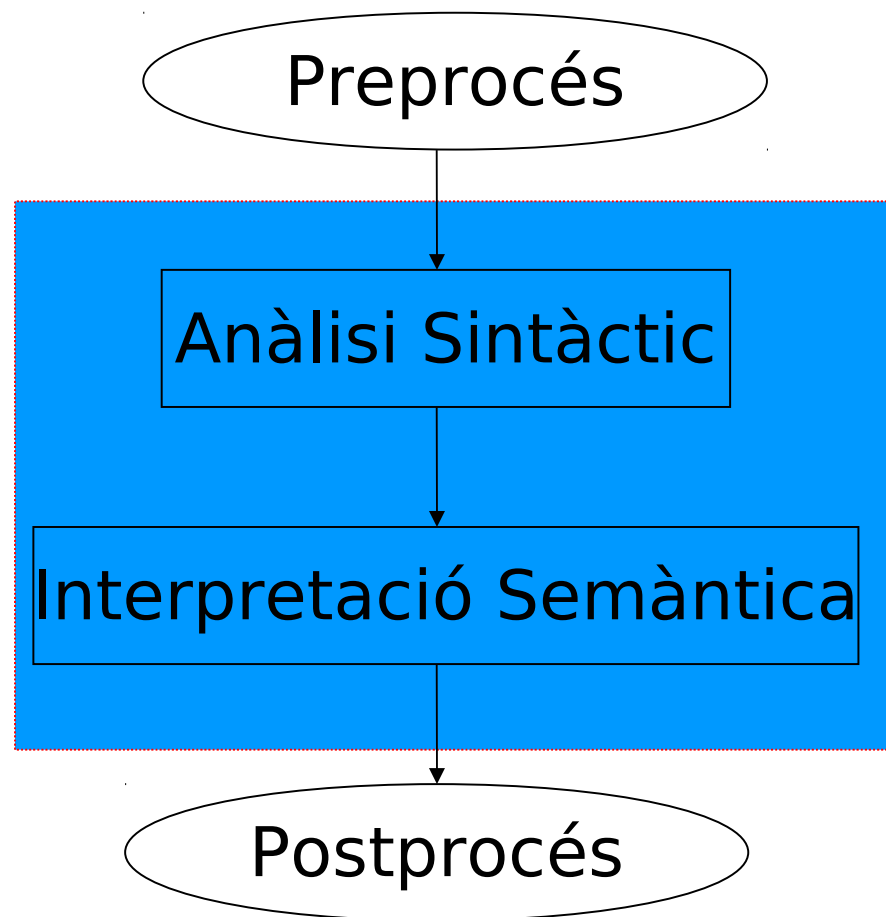
Resultat de l'anàlisi morfològica

| | | | |
|----------------|--------------------|-----------------|-----------------|
| quina | quin DT0FS00 | quina NCFS000 | |
| és | ésser VMIP3S0 | | |
| la | el TDFS0 | ell PP3FS000 | 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 |

Processament del LN



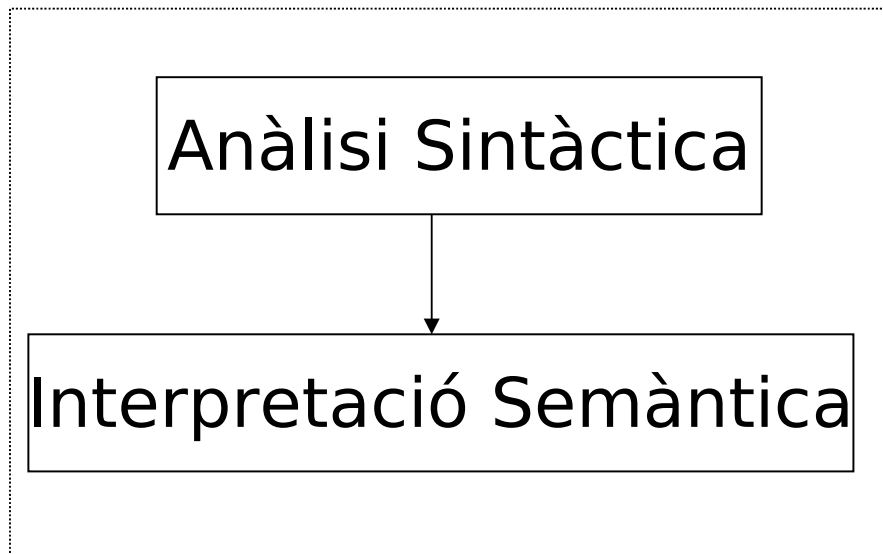
“Quina es la capital de França ?”

...

```
(oracio
  (oracio_interrogativa
    (pron_interrogatiu (quina))
    (verb (es)
      (sn (... la capital de França)))
    )
  )
)
```

...

Formes de col·laboració



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

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 (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^* , $w_1^n = w_1, \dots, w_n$, $w_i \in \Sigma$, hem de determinar si $w_1^n \in L$

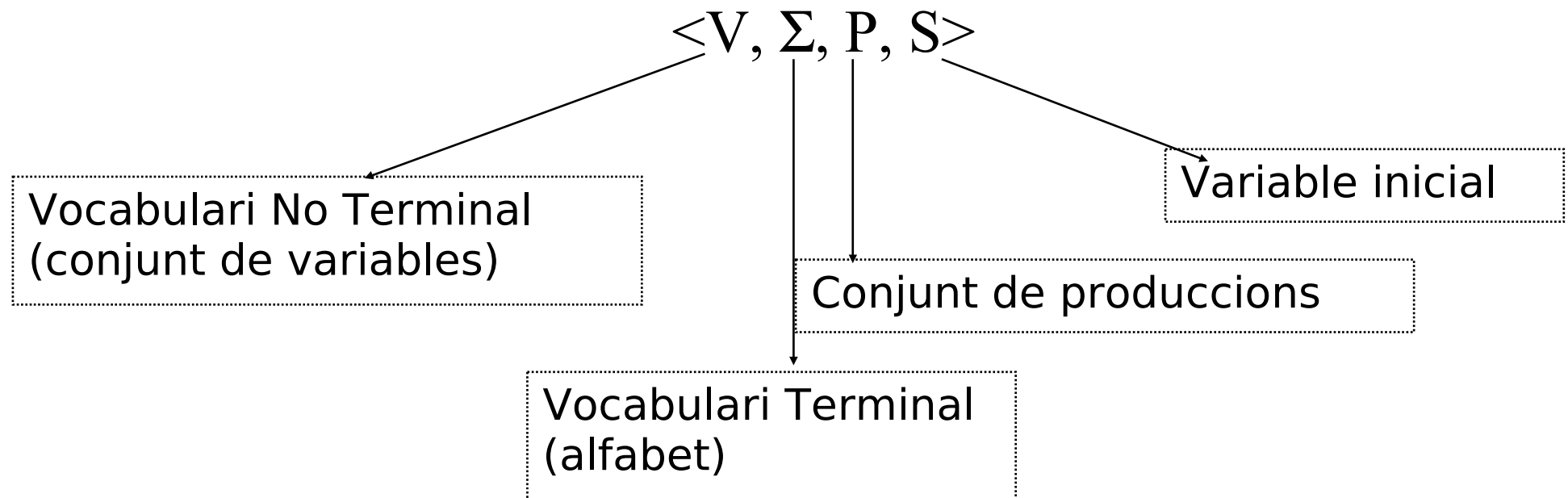
Formes de definir la pertanyença

- 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

- Gramàtiques de constituents
 - Arbres de derivació
- Gramàtiques de dependències
 - Esquemes de dependències
- Gramàtiques de casos
 - Models d'actants → Xarxes semàntiques

Gramàtiques d'Estructura Sintagmàtica



$$\Sigma \cap V = \emptyset$$

$$\Sigma \cup V = \text{Vocabulari}$$

$$S \in V$$

Tipus de Gramàtiques - Jerarquia de Chomsky (1)

Tipus 0 Gramàtiques sense restriccions

- Els elements de P són regles de reescriptura del tipus

$$u \rightarrow w, \quad w, u \in (V \cup \Sigma)^*$$

- Corresponen als llenguatges enumerables recursivament
- Son reconeguts per Maquines de Turing

Tipus 1 Gramàtiques sensibles (Context-sensitive Grammars)

- Es defineix com a restricció la llargària de les regles

$$u \rightarrow w, \quad w, u \in (V \cup \Sigma)^* \text{ i } |u| \leq |v|$$

- Corresponen als llenguatges Contextuals
- Son reconeguts per Linear Bounded Automata

Tipus de Gramàtiques - Jerarquia de Chomsky (2)

Tipus 2 Gramàtiques incontextuals (Context-free Grammars, CFG)

- Els elements de P son regles de reescriptura restringides a les del tipus:

$$A \rightarrow w, \quad A \in V, w \in (V \cup \Sigma)^*$$

- Corresponen als llenguatges incontextuals
- Son reconeguts per automates de pila no deterministes

Tipus 3 Gramàtiques regulars (Regular Grammars, RG)

- Els elements de P son regles de reescriptura dels tipus:

$$A \rightarrow a$$

$$A \rightarrow aB, \quad A, B \in V, a \in \Sigma$$

- Corresponen als llenguatges regulars
- Son reconeguts per automates finits

Condicció de gramaticalitat

- Una frase w (un mot de Σ^*) pertany al llenguatge generat per la gramàtica:

$$w \in L(G) \Leftrightarrow S \xrightarrow[G]{*} w$$

- Podem dir que la gramàtica G pot derivar el mot w utilitzant les produccions a partir de S .

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
- Es el LN un llenguatge incontextual?
- Suficient? NO (normalment)
- Solució
 - CFG + {adició procedimental del contexte}
 - Gramàtiques Lògiques i d'unificació
 - Gramàtiques enriquides amb informació estadística
 - SCFG
 - Gramàtiques lexicalitzades

Exemple de gramàtica incontextual

- | | |
|------------|--------------------|
| (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 ... |

CFG + {addició procedimental del contexte}

| | |
|-------------|--|
| 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" ... |

Factors que incideixen en el procés d'anàlisi sintàctica

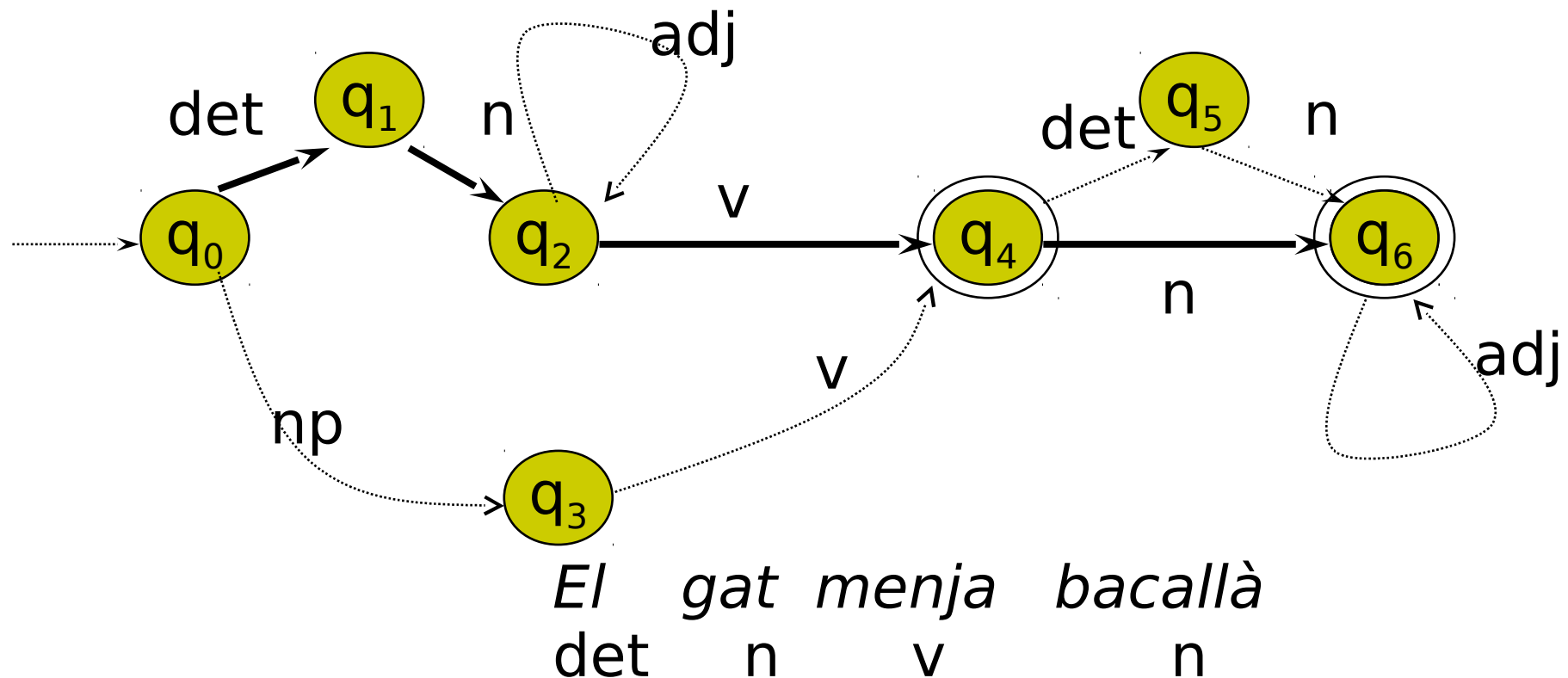
- Expressivitat de la gramàtica
- Àmbit (Coverage)
- Fonts de coneixement implicades
- Estrategia de l'anàlisi
- Direcció de l'anàlisi
- Ordre d'aplicació de les regles
- Gestió de l'ambigüitat
- (in)determinisme
- Enginyeria dels analitzadors

Analitzadors per CFG i extensions

- Simplificacions de CFG
 - CFG \Rightarrow RG
 - Tècniques d'estats finits: **FSA**
 - CFG \Rightarrow DCFG
 - Analitzadors deterministes: **LL, LR**
- Extensions dels FSA
 - TN \Rightarrow RTN \Rightarrow ATN (Woods, 1970)
- WFST, Charts (M. Kay, 1980)
- Mètodes tabulars: **CKY**, Earley (1970)
- Gramàtiques d'estructura de frase:
 - **LSP** (N. Sager, 1981)
 - **Diagram** (A. Robinson, 1981)
- **Parsifal** (M. Marcus, 1980)

Xarxes de Transició (TN)

- Autòmat finit
 - Estats associats a parts de la frase
 - Transicions: Etiquetes que fan referència a categories morfosintàctiques
 - No determinisme



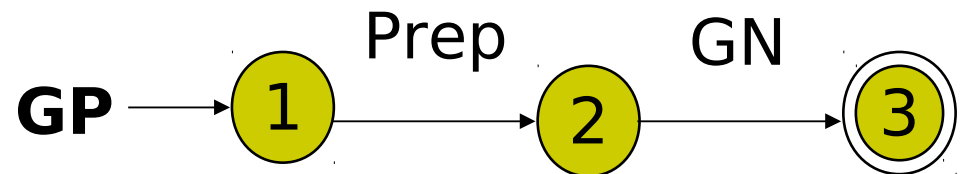
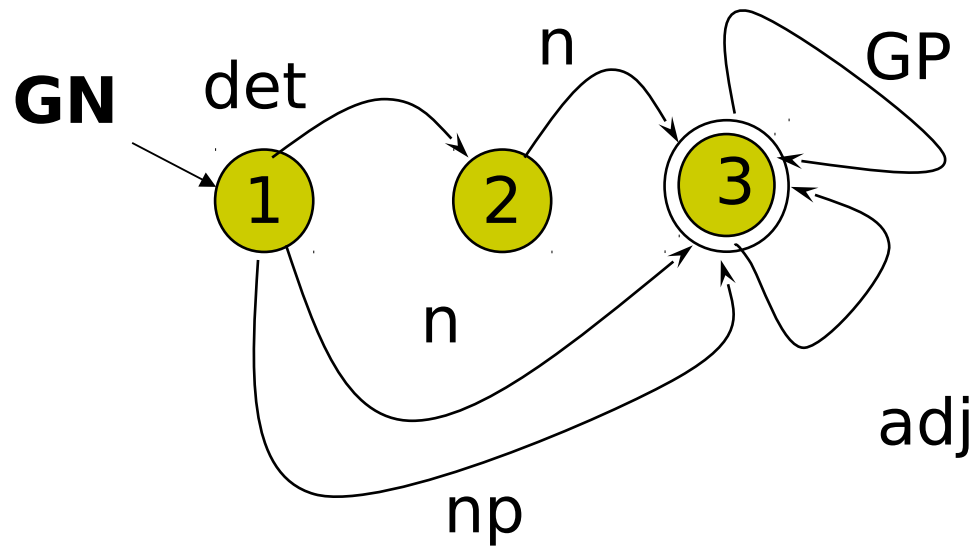
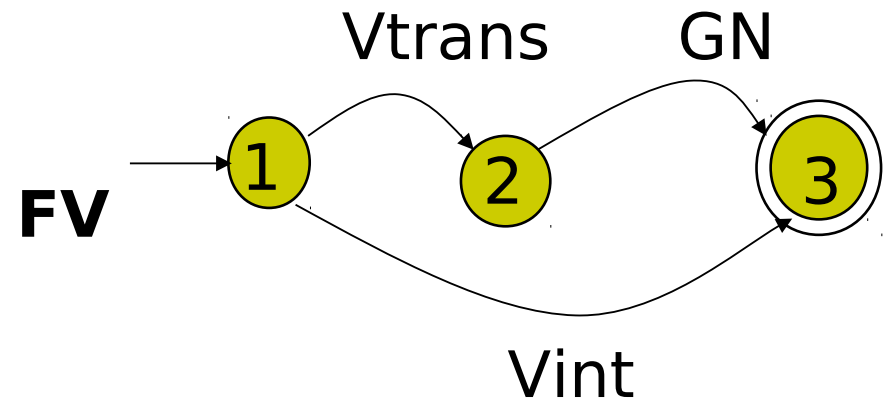
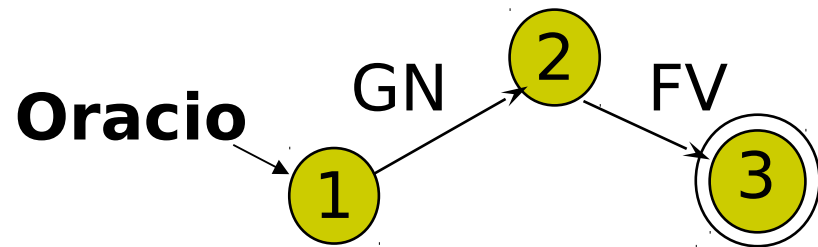
TN: limitacions

- 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

Xarxes de Transició Recurrents (RTN)

- Colecció de xarxes de transició (TN) etiquetades amb un nom
 - Arcs etiquetats amb categories → com xarxes normals
 - etiquetes terminals
 - Arcs etiquetats amb identificadors de xarxes de transició (TN)
 - 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 són dèbilment equivalents a les CFG

RTN: exemple per a frases verbals

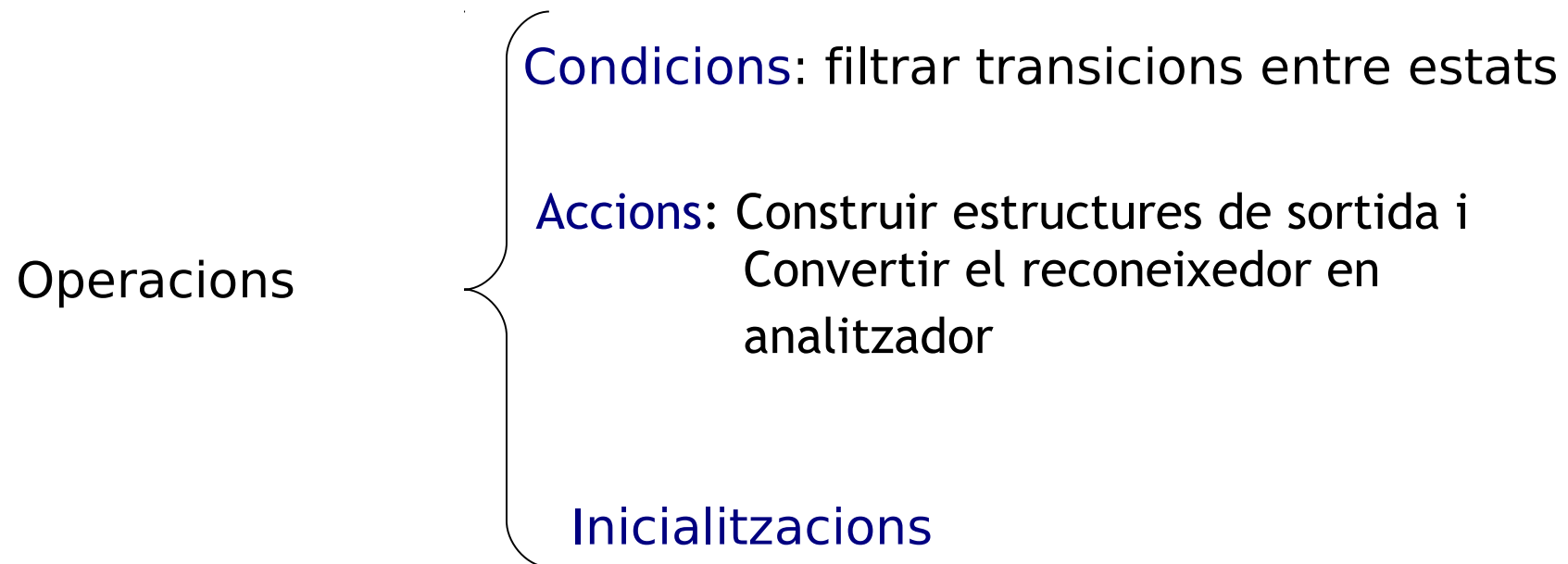


RTN: limitacions

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

Xarxes de transició augmentades (ATN) (Woods, 1970)

- ATN = RTN amb *operacions* afegides als arcs i ús de *registres*.

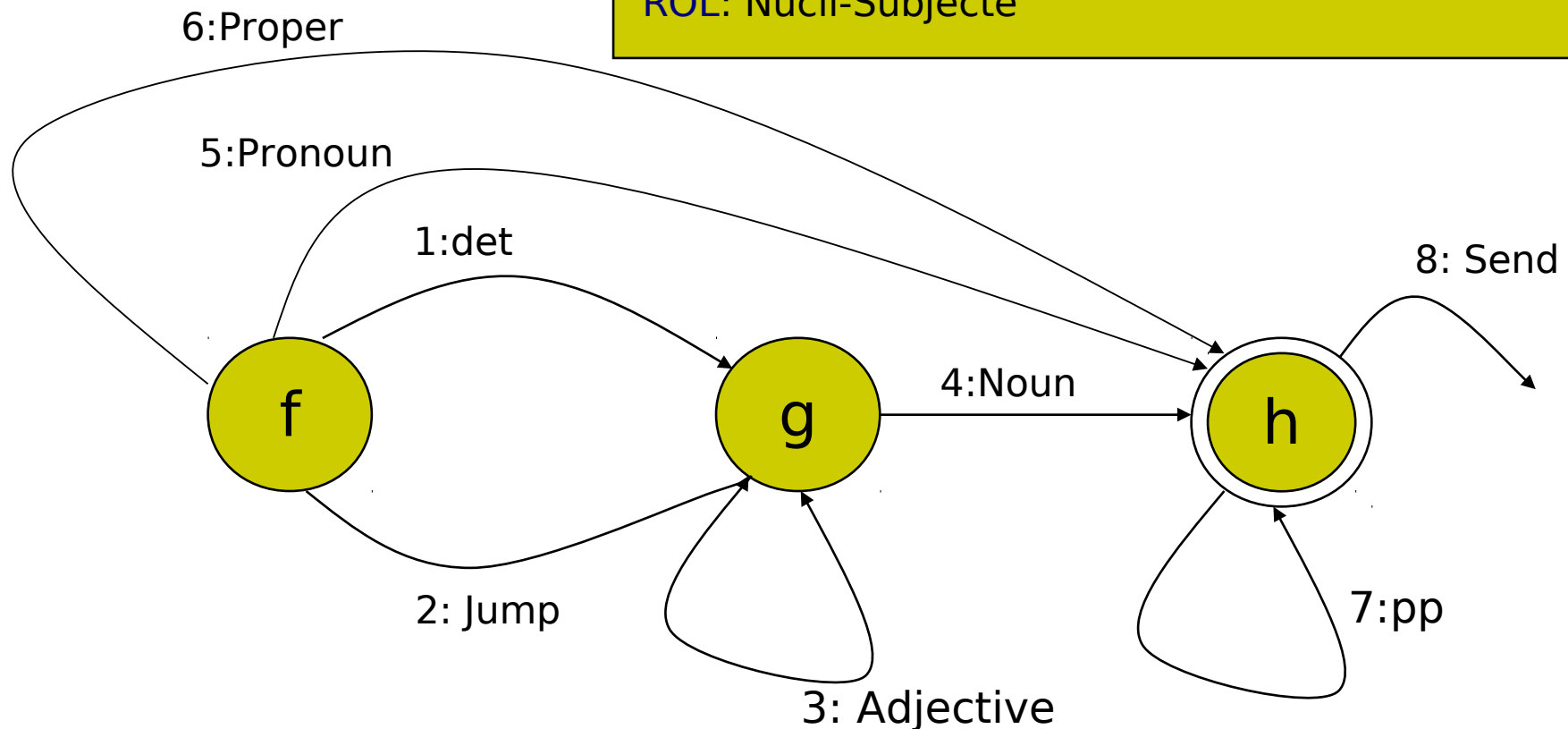


- Permeten expressar les restriccions contextuais

ATN: exemple (Winograd, 1983)

TRETS: Number: Singular, Plural Default: empty
Person: 1st, 2nd, 3rd Default: 3rd

ROL: Nucli-Subjecte



Xarxa per a reconèixer grups nominals (NP)

ATN: exemple (2)

- Inicialitzacions, Condicions i Accions:

NP-1: _fDeterminer_g

A: Set Number to the number of *

NP-4: _gNoun_h

C: Number is empty or number is the number of *

A: Set Number to the number of *

Set Nucli-Subjecte to *

NP-5: _fPronoun_h

A: Set Number to the number of *

Set Person to the Person of *

Set Nucli-Subjecte to *

NP-6: _fProper_h

A: Set Number to the number of *

Set Nucli-Subjecte to *

ATN: limitacions

- Són adequades per l'anàlisi descendent, però no resulta fàcil implementar una anàlisi ascendent o híbrida
- Redundància de les operacions de backtracking
 - ineficiència
- Problemes d'expressivitat notacional:
 - la gramàtica es barreja amb les accions

Charts (Kay, 1973,1980)

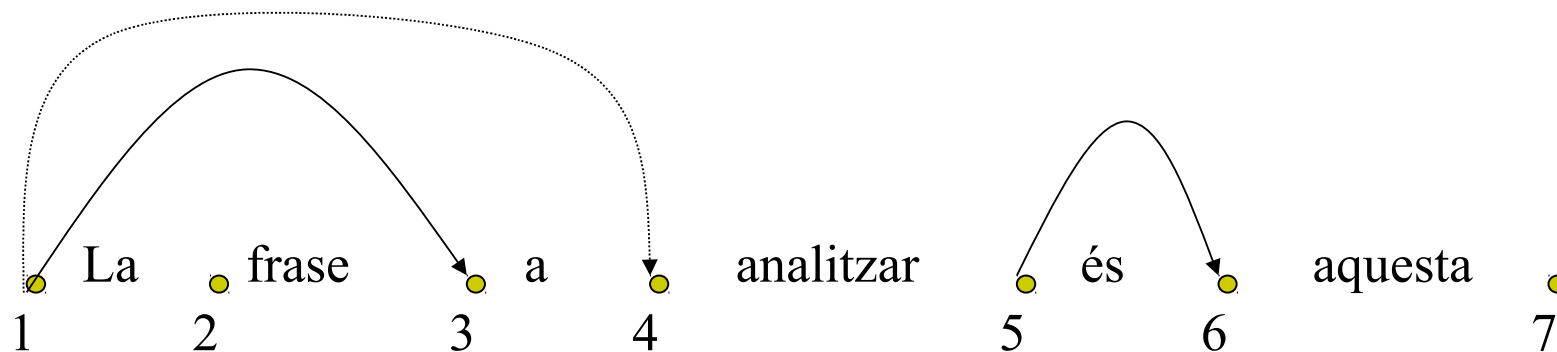
- **Chart** = graf dirigit que es construeix de manera dinàmica i incremental a mesura que es realitza l'anàlisi.
- Els **nodes** corresponen al principi i final de la frase i a les separacions entre paraules (N+1 nodes)

1 2 3 4 5 6 7
● La ● frase ● a ● analitzar ● és ● aquesta ●

- Intenten eliminar redundàncies en l'anàlisi (alleujament del cost del Backtracking) memoritzant estructures parcials ja construïdes.
- Inconvenients: espai, temps de construcció, només guarda components ben formats

Charts (Elements)

- Els arcs es creen dinàmicament.
- Un arc de la posició i a la j ($j \geq i$) 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 $A \rightarrow BCD$ es poden derivar les següents regles puntejades:

$A \rightarrow . B C D$ (corresponent a un arc actiu)

$A \rightarrow B . C D$ ”

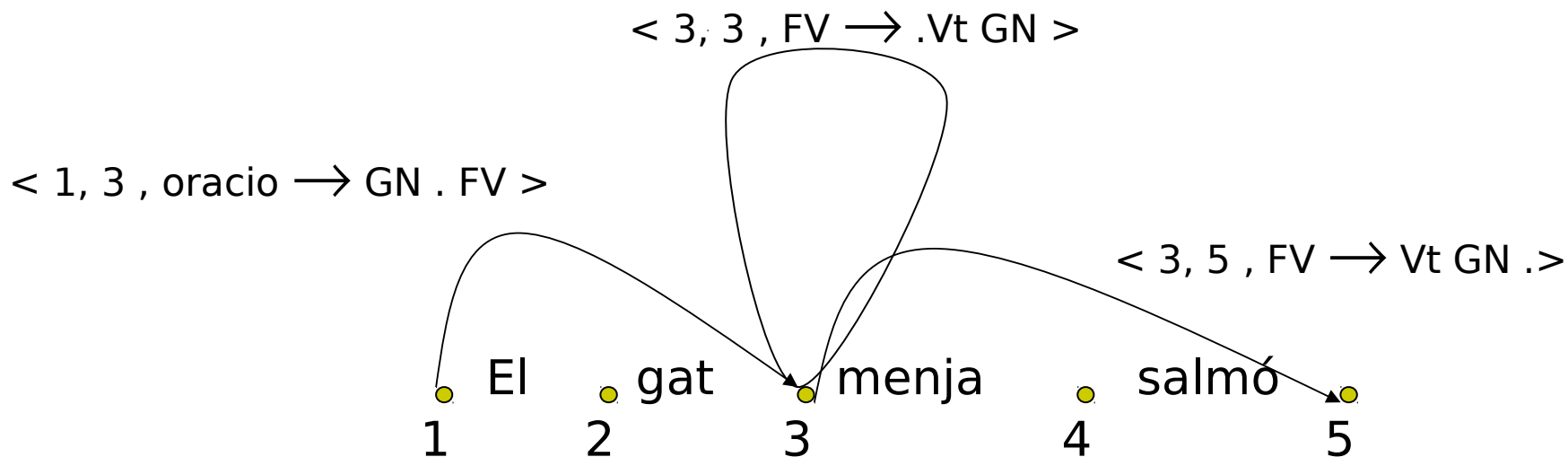
$A \rightarrow B C . D$ ”

$A \rightarrow B C D .$ (corresponent a un arc inactiu)

Charts: notació (2)

- Arc d'un chart: $\langle i, j, X \rightarrow a.b \rangle$

| | |
|---------------------|---------------------------|
| $i, j :$ | 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 $\langle i, j, A \rightarrow a. \rangle$, aleshores s'ha d'afegir al seu extrem esquerre un arc actiu $\langle i, i, B \rightarrow .Ab \rangle$ per cada regla $B \rightarrow A b$ de la gramàtica

Inicialització: afegir els arcs inactius que corresponen a les categories lèxiques (terminals).

Ex: $\langle 1, 2, \text{Det} \rightarrow \text{el}. \rangle$

Estratègia descendent

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

Inicialització: Igual que abans però a més cal afegir l'arc actiu que correspon a l'objectiu d'obtenir una frase.

Ex: $\langle 1, 1, \text{oracio} \rightarrow .SN\ SV \rangle$

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

Formalismes d'Unificació. Gramàtiques Lògiques

- Formalismes basats en unificació \subset Gramàtiques lògiques
- Llenguatge habitual d'implementació: Prolog
- Característiques
 - **Unificació** com a mecanisme bàsic de composició entre constituents
 - **Aproximació sintagmàtica** com a forma bàsica de descripció gramatical

Notació

- Afirmacions (fets)
home (X) \leftarrow
- Condicions (regles) (Consequent \leftarrow antecedent)
mortal (X) \leftarrow home (X)
- Negacions (consultes Existencials)
 \leftarrow immortal(X)
- Contradicció
 \leftarrow

Anàlisi gramatical com a demostració de Teoremes

- L'expressió de la gramàtica i el lexicó com clàusules de Horn permet l'aplicació de la resolució i raonament per refutació com a procediment d'anàlisi.

| | | | |
|------------------|------------------------|---|-----------|
| (1) oracio (X,Y) | ← gnom(X,Z), gver(Z,Y) | } | Gramàtica |
| (2) gnom(X,Y) | ← art(X,Z), nom(Z,Y) | | |
| (3) gver(X,Y) | ← ver(X,Y) | | |

| | | | |
|--------------|--------------|---|--------|
| (4) art(X,Y) | ← el (X,Y) | } | Lexicó |
| (5) nom(X,Y) | ← gos(X,Y) | | |
| (6) ver(X,Y) | ← borda(X,Y) | | |

Exemple (1)

1 el 2 gos 3 borda 4

(7) el(1,2) ←

(8) gos(2, 3) ←

(9) borda (3,4) ←

- TEOREMA a demostrar **oracio (1,4)** ←
- Raonant per refutació hauríem de negar...
 ← oracio (1,4)
- ...i per derivació descendent demostrar
 (una contradicció)

Exemple (2)

1 **El**
 2 **gos**
 3 **borda**
 4

Frase (1,4) ←

(R1) ($X \sqsubseteq 1$, $Y \sqsubseteq 4$) per unificació

← $\text{gnom}(1, Z)$, $\text{gver}(Z, 4)$

(R2) i (R4) aplicades a $\text{gnom}(1, Z)$ i $\text{art}(1, U)$

← $\text{art}(1, U)$, $\text{nom}(U, Z)$, $\text{gver}(Z, 4)$

← $\text{el}(1, 2)$, $\text{nom}(U, Z)$, $\text{gver}(Z, 4)$

(R7) ($U \sqsubseteq 2$)

← $\text{nom}(2, Z)$, $\text{gver}(Z, 4)$

(R5) i (R8) ($Z \sqsubseteq 3$)

← $\text{gos}(2, 3)$, $\text{gver}(3, 4)$

← $\text{gver}(3, 4)$

(R3) i (R6) i (R9)

← $\text{ver}(3, 4)$

← $\text{borda}(3, 4)$

Exemple (3)

1 El 2 gos 3 borda 4

Frase (1,4) ←

← gnom (1,Z), gver(Z,4)
← art(1,U), nom(U,Z), gver(Z,4)
← el, nom(2,Z), gver(Z,4)
← nom(2,Z), gver(Z,4)
← gos(2,Z), gver(Z,4)
← gver(3,4)
← ver(3,4)
← borda(3,4)

Interpretació directa a Prolog !!

Analitzadors d'unificació

- Formalisms lògics
 - Expresivitat i Tractament
 - Les gramàtiques de clàusules definides (DCG)
- El Prolog com analitzador
- El tractament de l'unificació
 - Representació dels termes
 - Algoritme d'unificació

Gramàtiques de Clàusules Definides (DCG)

- Les gramàtiques de clàusules definides permeten escriure gramàtiques lògiques com programes PROLOG
- PROLOG es un llenguatge de regles que fa servir raonament cap enrere com mètode de resolució
- Es defineix una sintaxi especial que permet ocultar el tractament de la frase i diferenciar els elements gramaticals dels procediments que es fan servir per augmentar la gramàtica incontextual
- Les regles fan servir variables per comunicar-se informació i fer les comprovacions necessàries que exigeixi la gramàtica

Gramàtiques de Clàusules Definides (Sintaxi)

- Una regla gramatical té la següent sintaxi
 - `izq --> der1, der2, der3, ..., derN`
- Cadascun dels símbols de la gramàtica pot tenir variables per diferents usos (passar o rebre informació d'altre producció, construir un resultat)
- Es consumeixen elements de l'entrada fent servir corxets i indicant una llista de variables y/o constants que es volen unificar amb l'entrada
 - `aaa --> [W], bbb`
- Es pot introduir codi PROLOG posant-lo entre claus
 - `aaa(W) --> [W], bbb(W), {number(W)}`
- Per executar una DCG es crida al símbol principal de la gramàtica amb dos paràmetres, una llista amb les paraules de la frase i una llista buida
 - `frase([el,gat,menja,bacalla],[])`

Gramàtiques de Clàusules Definides (Exemple)

```
analisi(X,Y):- asercion(X,Y).
```

```
asercion --> sn, verb, compl.
```

```
compl --> [].
```

```
compl --> prep, sn.
```

```
compl --> sn.
```

```
sn --> npr.
```

```
sn --> det, n.
```

```
verb --> [W], {verbo(W)}.
```

```
npr --> [W], {npropio(W)}.
```

```
n --> [W], {nombre(W)}.
```

```
det --> [W], {determ(W)}.
```

```
prep --> [W], {prepo(W)}.
```

```
npropio(clara).
```

```
npropio(maria).
```

```
npropio(barcelona).
```

```
nombre(hombre).
```

```
nombre(profesor).
```

```
nombre(libro).
```

```
determ(un).
```

```
determ(el).
```

```
verbo(esta).
```

```
verbo(rie).
```

```
verbo(piensa).
```

```
verbo(habla).
```

```
verbo(lee).
```

```
prepo(en).
```

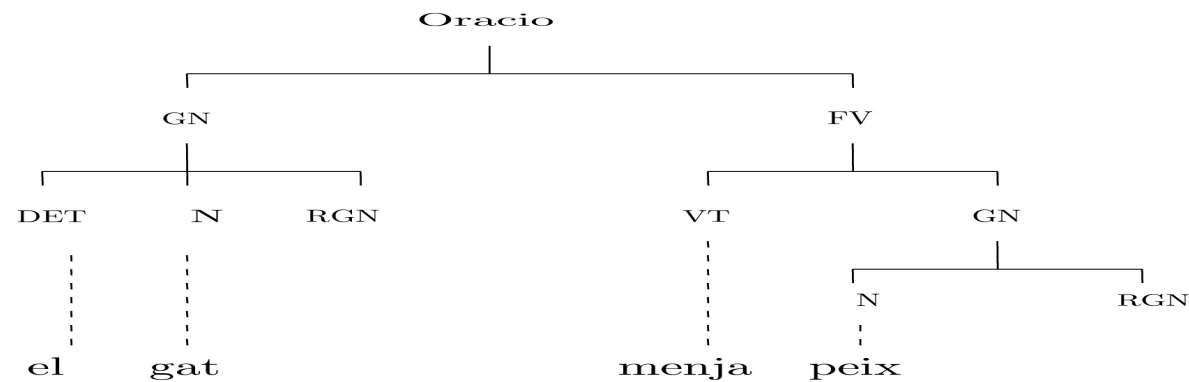
```
prepo(con).
```

```
prepo(de).
```

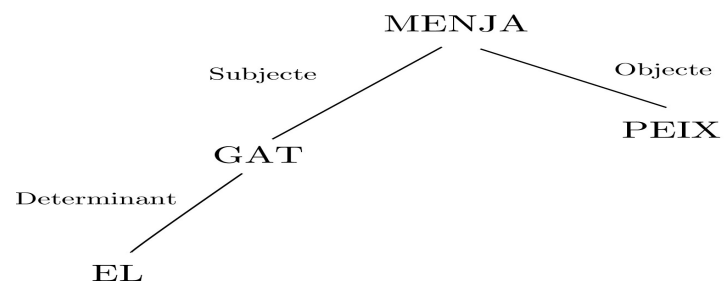
Resultat del anàlisi sintàctic

- Arbre d'anàlisi
 - Estructura de components
- Estructura de dependències
- Model d'actants
 - Xarxa semàntica
- Forma lògica

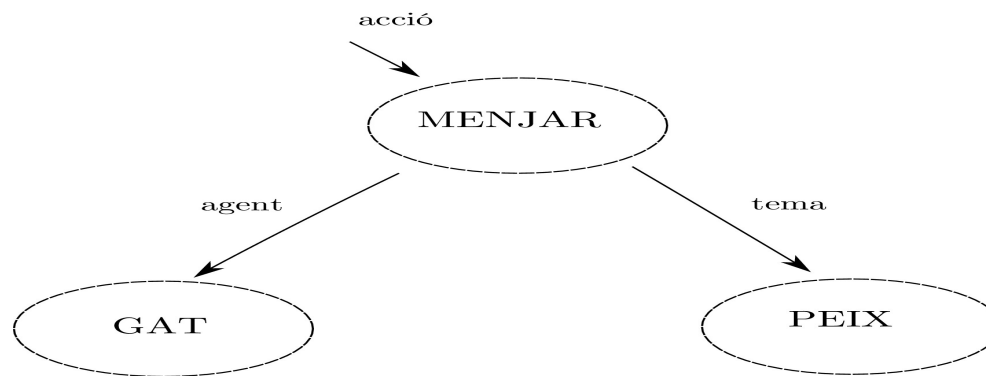
Estructura de components



Estructura de dependències



Model d'actants

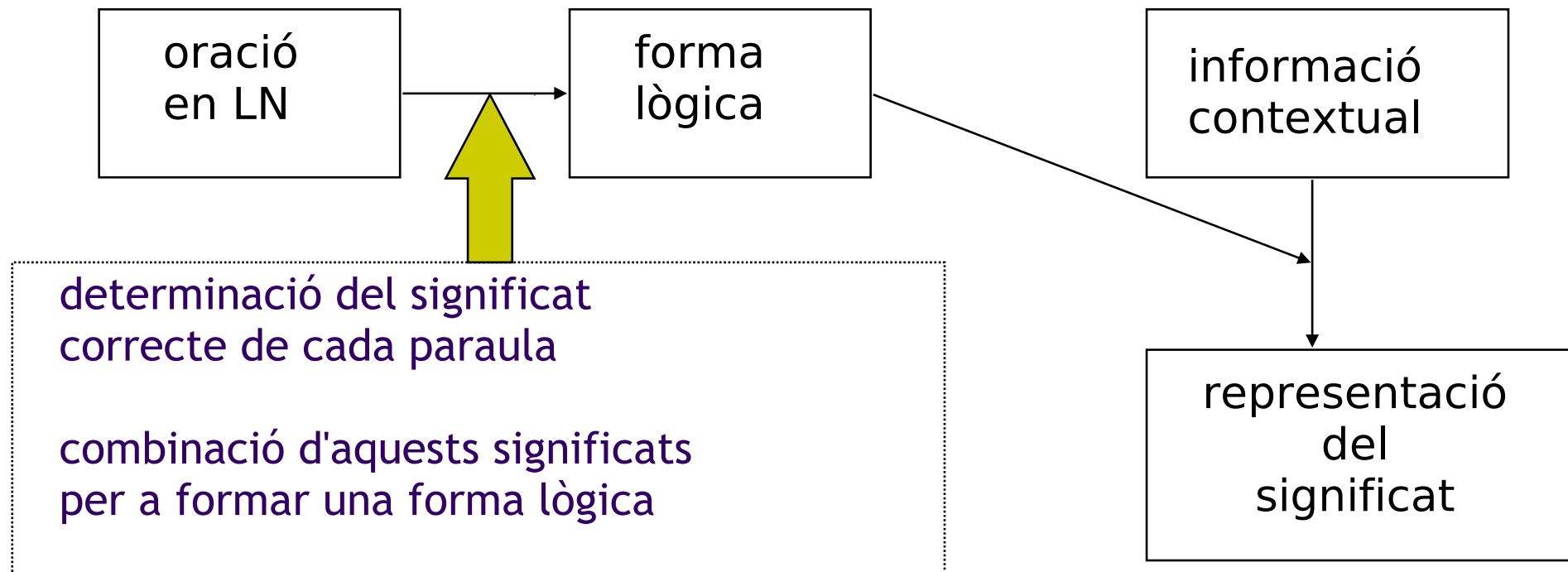


Forma lògica

```
existeix(X,  
  i(gat(X),  
    existeix(Y,  
      i(peix(Y),  
        menja(X,Y))))))
```

La Semàntica

- La Semàntica tracta el significat de les oracions.
- La Interpretació Semàntica (IS) es el procés d'extracció d'aquest significat.



Propietats del Model Semàntic

- **Semàntica compositiva:** la representació semàntica d'un objecte s'ha de poder realitzar a partir de les representacions semàntiques dels seus components:

$$IS = f(IS(\text{Components Sintàctiques}))$$

- Basat en una teoria
- Definició d'un Sistema de Representació Semàntica.
- Interfície Sintaxi-Semàntica
- La IS ha de ser robusta en front a les ambigüitats
- La IS ha de poder tractar fenòmens complexos com: *la quantificació, les modalitats, la negació.*

Dos problemes

- La representació del significat
- La interpretació semàntica

La Representació del significat (1)

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))))))
```

La Representació del significat (2)

- En aquesta fórmula apareixen quatre tipus diferents d'informació:
 - Estructura lògica
 - Contingut conceptual (semàntic)
 - Indicació dels actes de parla
 - Anotacions pragmàtiques
- El formalisme hauria de ser capaç de proporcionar una capacitat expressiva suficient per a garantir la descripció d'aquests quatre tipus d'informació.

La Representació del significat (3)

- Formes de representació:
 - CP1
 - Freqüentment en forma clausal
 - Altres formalismes lògics
 - Xarxes semàntiques
 - Frames

Representacions basades en lògica.

- Un vocabulari de **predicats** en el qual haurem d'indicar la aritat (el nombre d'arguments i, de vegades, el seu tipus).
- Un vocabulari de **constants i variables**.
- Un conjunt de **conectors lògics**.
- Un vocabulari de **funcions** de les quals indicarem també la aritat.
- Un conjunt de **quantificadors** que actuaràn sobre els predicats que hagin de ser (o puguin ser) quantificats.

La interpretació semàntica

- Interacció amb l'anàlisi sintàctica:
 - procés en cascada (1)
 - sintaxi | semàntica
 - procés en cascada (2)
 - {sintaxi + filtre semàntic} | semàntica
 - procés en paral·lel
 - {sintaxi, semàntica}
- Forma de realitzar la composició
 - funció de composició
 - activació de la IS

Exemple de Funció de Composició

- Interpretació mitjançant lambda avaluacions:

$$(\text{lambda } (x) (\dots)) = (\lambda (x) (\dots))$$

Gramàtica

Oracio \rightarrow GN FV (2 1)

GN \rightarrow np (1)

FV \rightarrow vi (1)

Fv \rightarrow vt GN (1 2)

Lexicó

Pere \rightarrow np, pere

Maria \rightarrow np, maria

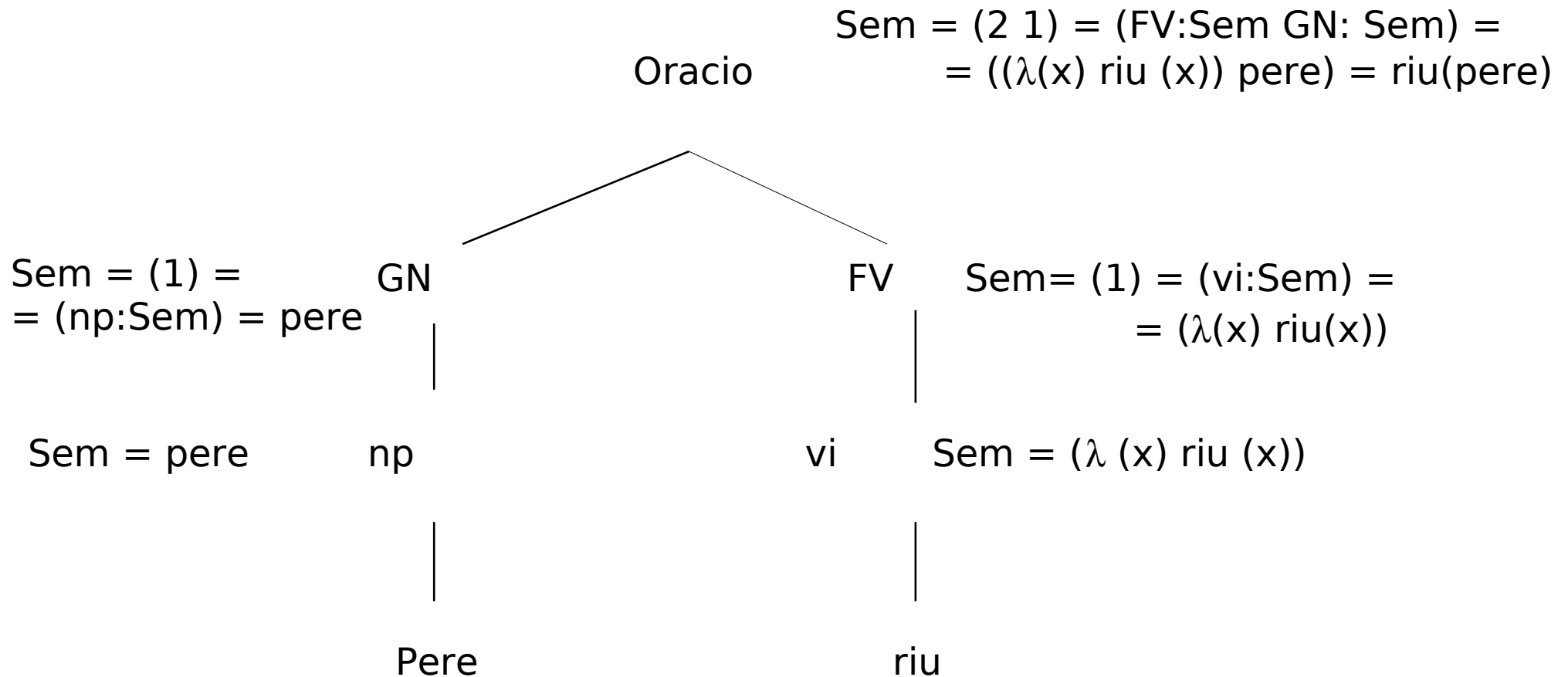
riu \rightarrow vi, $(\lambda(x) (\text{riu}(x)))$

estima \rightarrow vt,

$((\lambda(x) (\lambda(y), \text{estima}(y, x))))$

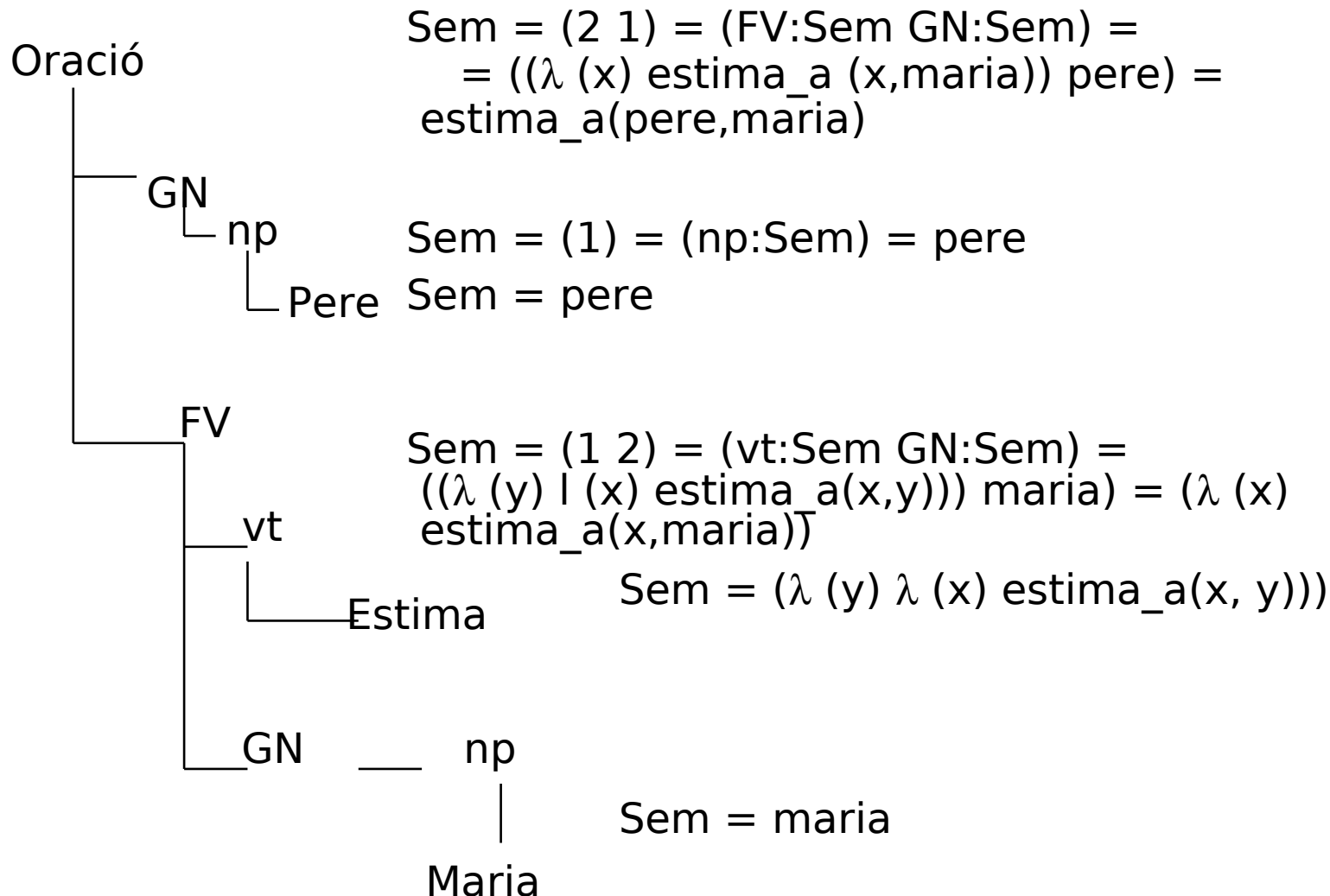
Funció de Composició: exemple 1

“Pere riu”

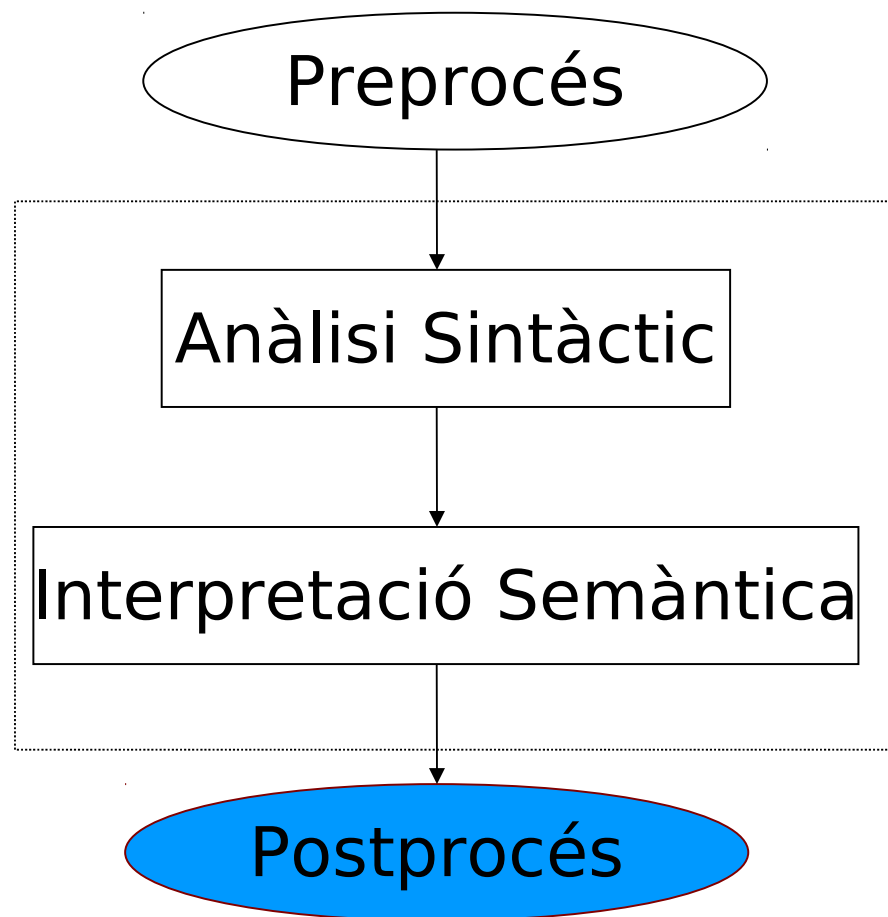


Funció de Composició: exemple 2

“Pere estima Maria”



Processament del LN



“Quina es la capital de França ?”

...

```
(oracio
  (oracio_interrogativa
    (pron_interrogatiu (quina))
    (verb (es)
      (sn (... la capital de França)))
    )
  )
)
```

...

pregunta(X), capital(X,frança)

Postprocés

- Resolució de la referència
- Anàlisi semàntica-pragmàtica
- Anàlisi il·locutiva
 - Reconeixement d'intencions
- ...

PLN a l'actualitat (1990-)

- **PLN empíric**
- Basat en **corpus** textuais
- Processament **Estadístic** vs. **Lingüístic**
- **Combinació** de models lingüístics i estadístics
- Aplicació de mètodes d'**Aprenentatge Automàtic** per a modelar el llenguatge

Aplicacions “clàssiques”

- Basades en **textos**
 - Traducció automàtica
 - Extracció de informació a partir de textos
- Basades en **diàlegs**
 - Interfícies en llenguatge natural
 - Accés a BD's
 - Sistemes de consulta telefònica
 - Sistemes tutors intel.ligents

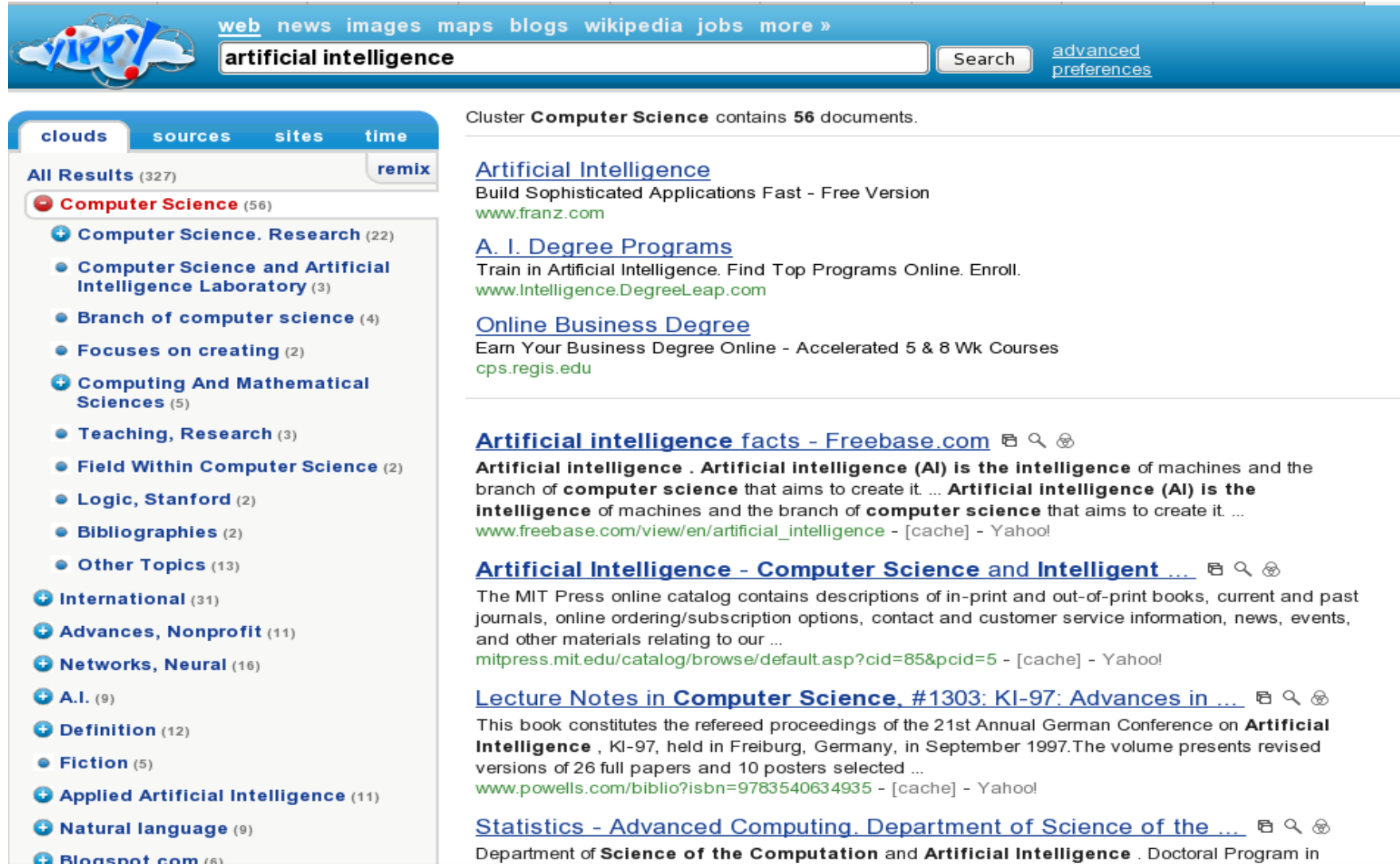
Aplicacions basades en grans coleccions de documents

- **Fonts**
 - Grans corpora de documents
 - Internet
- **Recuperació d'Informació (IR)**
 - Valor afegit de la component lingüística
 - Accés i recuperació independent de la llengua
- **Categorització de documents**
 - Document routing:
 - agències de premsa, classificació de e-mails, etc.
 - Document filtering:
 - e-mails comercials, netnews, etc.
 - (Re)organització automàtica de col·leccions de documents:
 - BD's textuais, Internet, etc.
 - Topic Detection and Tracking

Aplicacions basades en grans coleccions de documents(2)

- Extracció d'Informació sobre el Web
 - Integració d'informació
 - Clustering i detecció de relacions inter/intra documents
- Resum automàtic
- Question Answering

Extracció d'informació de la Web



The screenshot shows the Vippy search engine interface. At the top, there's a navigation bar with links for 'web', 'news', 'images', 'maps', 'blogs', 'wikipedia', 'jobs', and 'more'. A search bar contains the text 'artificial intelligence', and a 'Search' button is next to it. To the right of the search bar are links for 'advanced' and 'preferences'. Below the search bar, there's a sidebar on the left with tabs for 'clouds', 'sources', 'sites', and 'time'. Under the 'clouds' tab, there's a section for 'All Results (327)' and a 'remix' button. A list of categories is shown, including 'Computer Science' (56), 'Computer Science. Research' (22), 'Computer Science and Artificial Intelligence Laboratory' (3), 'Branch of computer science' (4), 'Focuses on creating' (2), 'Computing And Mathematical Sciences' (5), 'Teaching, Research' (3), 'Field Within Computer Science' (2), 'Logic, Stanford' (2), 'Bibliographies' (2), 'Other Topics' (13), 'International' (31), 'Advances, Nonprofit' (11), 'Networks, Neural' (16), 'A.I.' (9), 'Definition' (12), 'Fiction' (5), 'Applied Artificial Intelligence' (11), 'Natural language' (9), and 'Blogspot.com' (63). The main content area shows a cluster of results for 'Computer Science' containing 56 documents. The first result is 'Artificial Intelligence' with the description 'Build Sophisticated Applications Fast - Free Version' and the URL 'www.franz.com'. The second result is 'A. I. Degree Programs' with the description 'Train in Artificial Intelligence. Find Top Programs Online. Enroll.' and the URL 'www.Intelligence.DegreeLeap.com'. The third result is 'Online Business Degree' with the description 'Earn Your Business Degree Online - Accelerated 5 & 8 Wk Courses' and the URL 'cps.regis.edu'. Below these, there are three more results: 'Artificial intelligence facts - Freebase.com', 'Artificial Intelligence - Computer Science and Intelligent ...', and 'Lecture Notes in Computer Science, #1303: KI-97: Advances in ...'. Each result includes a brief description and a URL. The last result is 'Statistics - Advanced Computing. Department of Science of the ...' with the description 'Department of Science of the Computation and Artificial Intelligence . Doctoral Program in'.

web news images maps blogs wikipedia jobs more »

artificial intelligence Search advanced preferences

clouds sources sites time remix

All Results (327)

Computer Science (56)




- Computer Science. Research (22)
- Computer Science and Artificial Intelligence Laboratory (3)
- Branch of computer science (4)
- Focuses on creating (2)
- Computing And Mathematical Sciences (5)
- Teaching, Research (3)
- Field Within Computer Science (2)
- Logic, Stanford (2)
- Bibliographies (2)
- Other Topics (13)
- International (31)
- Advances, Nonprofit (11)
- Networks, Neural (16)
- A.I. (9)
- Definition (12)
- Fiction (5)
- Applied Artificial Intelligence (11)
- Natural language (9)
- Blogspot.com (63)




Cluster **Computer Science** contains **56** documents.




[Artificial Intelligence](#)
Build Sophisticated Applications Fast - Free Version
www.franz.com




[A. I. Degree Programs](#)
Train in Artificial Intelligence. Find Top Programs Online. Enroll.
www.Intelligence.DegreeLeap.com

[Online Business Degree](#)
Earn Your Business Degree Online - Accelerated 5 & 8 Wk Courses
cps.regis.edu

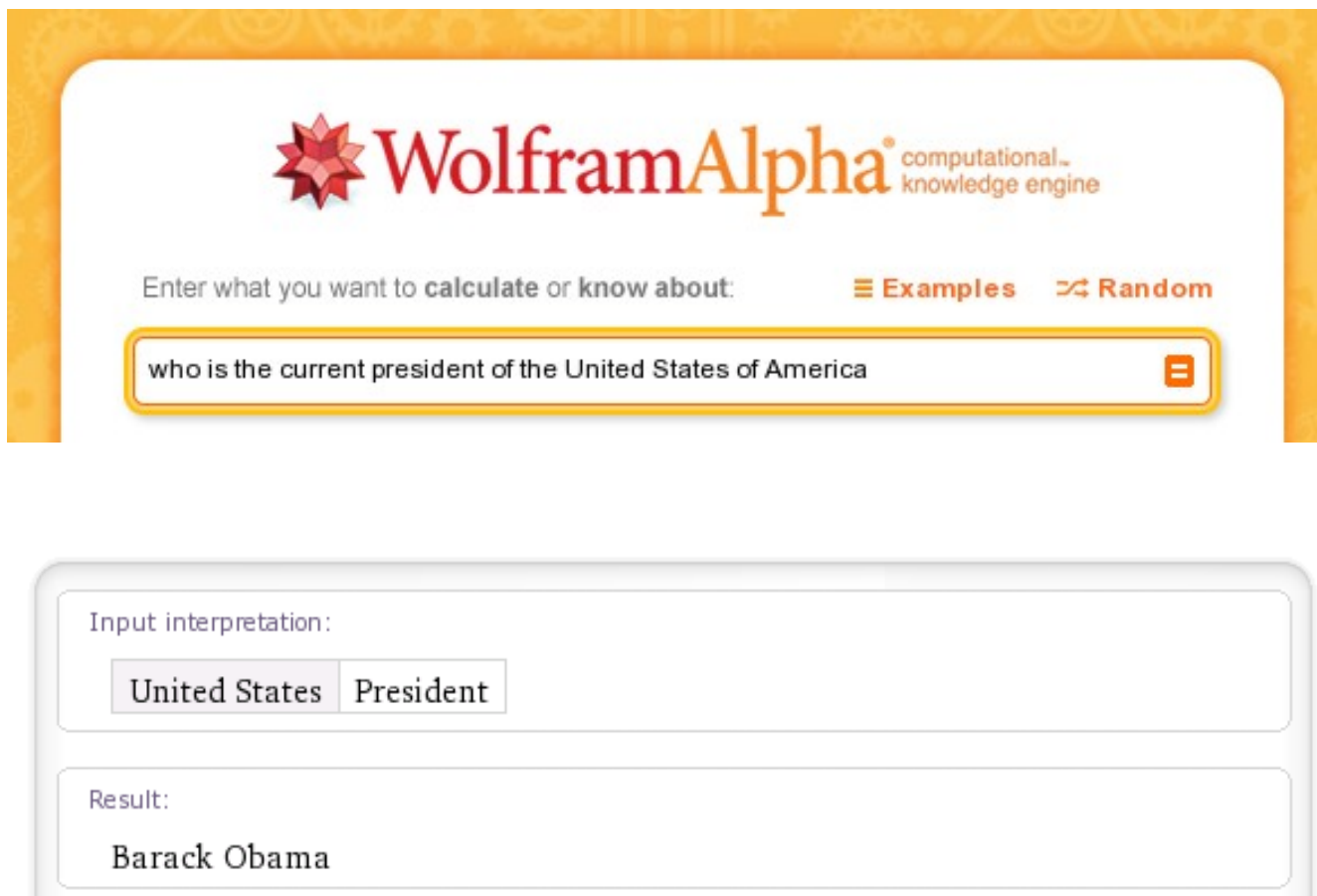
[Artificial intelligence facts - Freebase.com](#)   
Artificial intelligence . **Artificial intelligence (AI)** is the **intelligence** of machines and the branch of **computer science** that aims to create it ... **Artificial intelligence (AI)** is the **intelligence** of machines and the branch of **computer science** that aims to create it ...
www.freebase.com/view/en/artificial_intelligence - [cache] - Yahoo!

[Artificial Intelligence - Computer Science and Intelligent ...](#)   
The MIT Press online catalog contains descriptions of in-print and out-of-print books, current and past journals, online ordering/subscription options, contact and customer service information, news, events, and other materials relating to our ...
mitpress.mit.edu/catalog/browse/default.asp?cid=85&pcid=5 - [cache] - Yahoo!

[Lecture Notes in Computer Science, #1303: KI-97: Advances in ...](#)   
This book constitutes the refereed proceedings of the 21st Annual German Conference on **Artificial Intelligence** , KI-97, held in Freiburg, Germany, in September 1997. The volume presents revised versions of 26 full papers and 10 posters selected ...
www.powells.com/biblio?isbn=9783540634935 - [cache] - Yahoo!

[Statistics - Advanced Computing. Department of Science of the ...](#)   
Department of **Science of the Computation and Artificial Intelligence** . Doctoral Program in

Question Answering



The image shows a screenshot of the WolframAlpha website. At the top, the WolframAlpha logo is displayed with the tagline "computational knowledge engine". Below the logo, there is a text input field with the prompt "Enter what you want to calculate or know about:". To the right of the input field are links for "Examples" and "Random". The input field contains the text "who is the current president of the United States of America". Below the input field, the "Input interpretation:" section shows the text "United States" and "President" as separate tokens. The "Result:" section displays "Barack Obama" as the answer.

WolframAlpha[®] computational knowledge engine

Enter what you want to calculate or know about: [Examples](#) [Random](#)

who is the current president of the United States of America

Input interpretation:

United States President

Result:

Barack Obama

Question Answering

- Wolfram Alpha <http://www.wolframalpha.com/>
- Answers.com <http://www.answers.com/>
- Start <http://start.csail.mit.edu/>
- Qualim <http://demos.inf.ed.ac.uk:8080/qualim/>
- TextMap <http://brahms.isi.edu:8080/textmap/>