

**Vincenzo Lombardo**

**Informatica e studio del linguaggio**

L'accesso elettronico alla conoscenza

*vincenzo@mfn.unipmn.it*

# Il linguaggio “naturale”

- E' il principale veicolo della comunicazione
- Esistono grandi moli di dati codificati in linguaggio naturale (corpus)
- Accedere a tali risorse è uno dei compiti più importanti dell'informatica attuale
- La presenza del WWW ha garantito l'accesso immediato a una marea di informazioni in linguaggio naturale

# Comprendere il linguaggio

- Associare a ogni frase del linguaggio una rappresentazione non ambigua del suo significato
- Tale rappresentazione deve essere interpretabile da programmi ...
  - che ragionano
  - che eseguono dei comandi
  - che rispondono a domande
  - che risolvono problemi
  - ...

# Informatica e linguaggio

- Studi di tipo linguistico (computazionale)
  - Definire la competenza sintattica di un parlante nativo
  - Esistenza di una grammatica universale
- Interazione uomo-macchina
  - Interfacce vocali: comprensione/produzione di linguaggio parlato
  - Interfacce di tipo testuale
- Applicazioni pratiche che operano sui testi

# Lo studio gerarchico del linguaggio

- livello dei suoni (convenzioni sull'alfabeto)
  - suoni accettabili in una lingua: ca VS ça VS tcha
- livello delle parole (convenzioni sull'ortografia)
  - parole accettabili in una lingua: portosa VS rtosapo
- livello delle combinazioni di parole (sintassi)
  - “Giorgio volere di pane bianco” VS
  - “pane di volere Giorgio bianco”
- livello del significato (semantica)
  - “La macchia invisibile crede nel cielo” VS
  - “Idee verdi senza colore sognavano furiosamente”

# **Processi della comprensione del linguaggio**

- Livello lessicale
- Livello sintattico
- Livello semantico
- Livello pragmatico

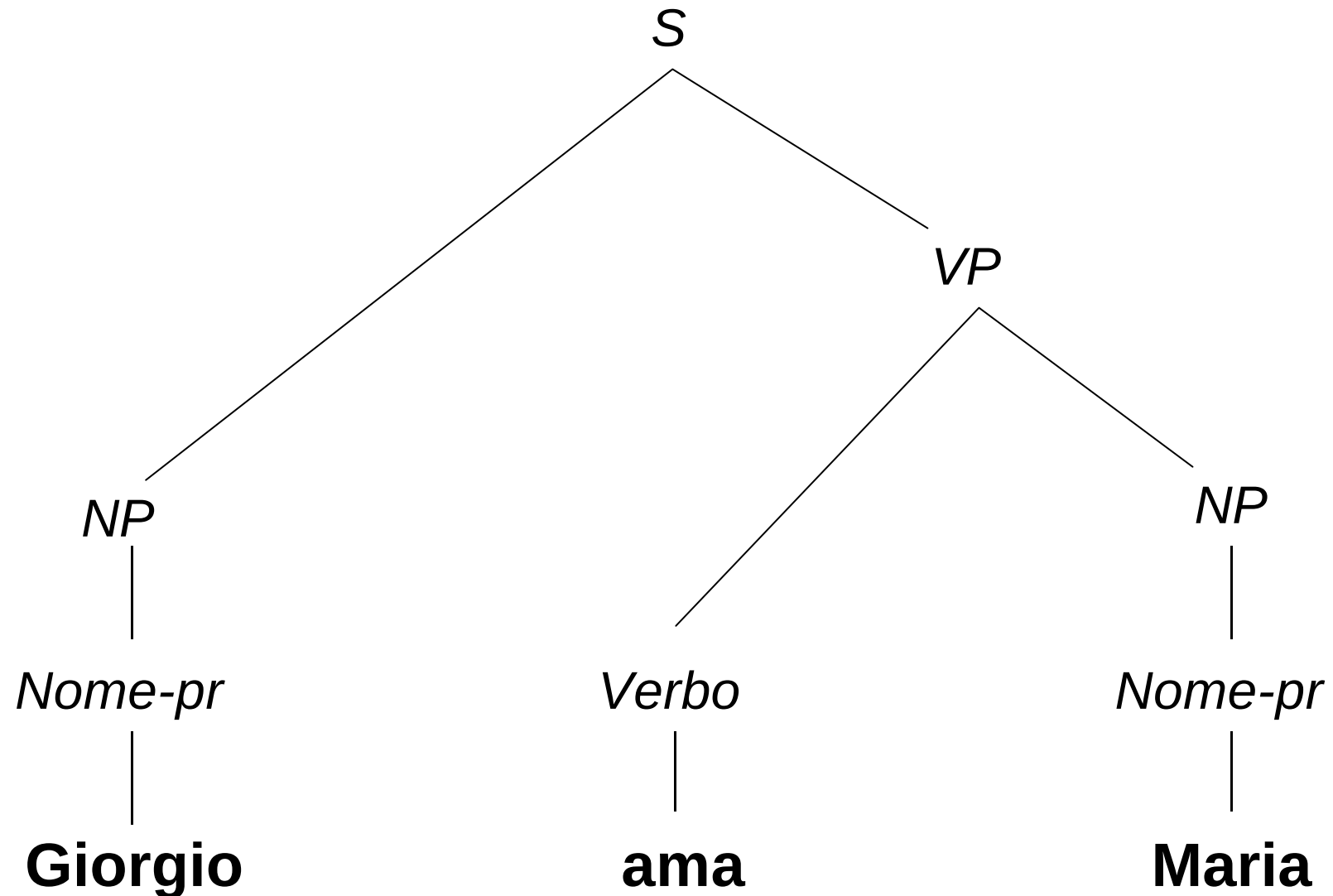
# Livello lessicale

- Obiettivo: elaborazione a livello della parola
- Fasi
  - riconoscimento dei suoni
  - segmentazione delle parole (tokenizing)
  - identificazione delle parole: analisi morfologica e ricerca delle radici
- Risultato: un testo suddiviso in parole a cui sono associate tutte le info sintattiche e semantiche

# Livello sintattico

- Obiettivo: come le parole si combinano per formare le frasi
- Fasi
  - tagging
  - parsing
- Risultato: un albero di struttura sintattica della frase

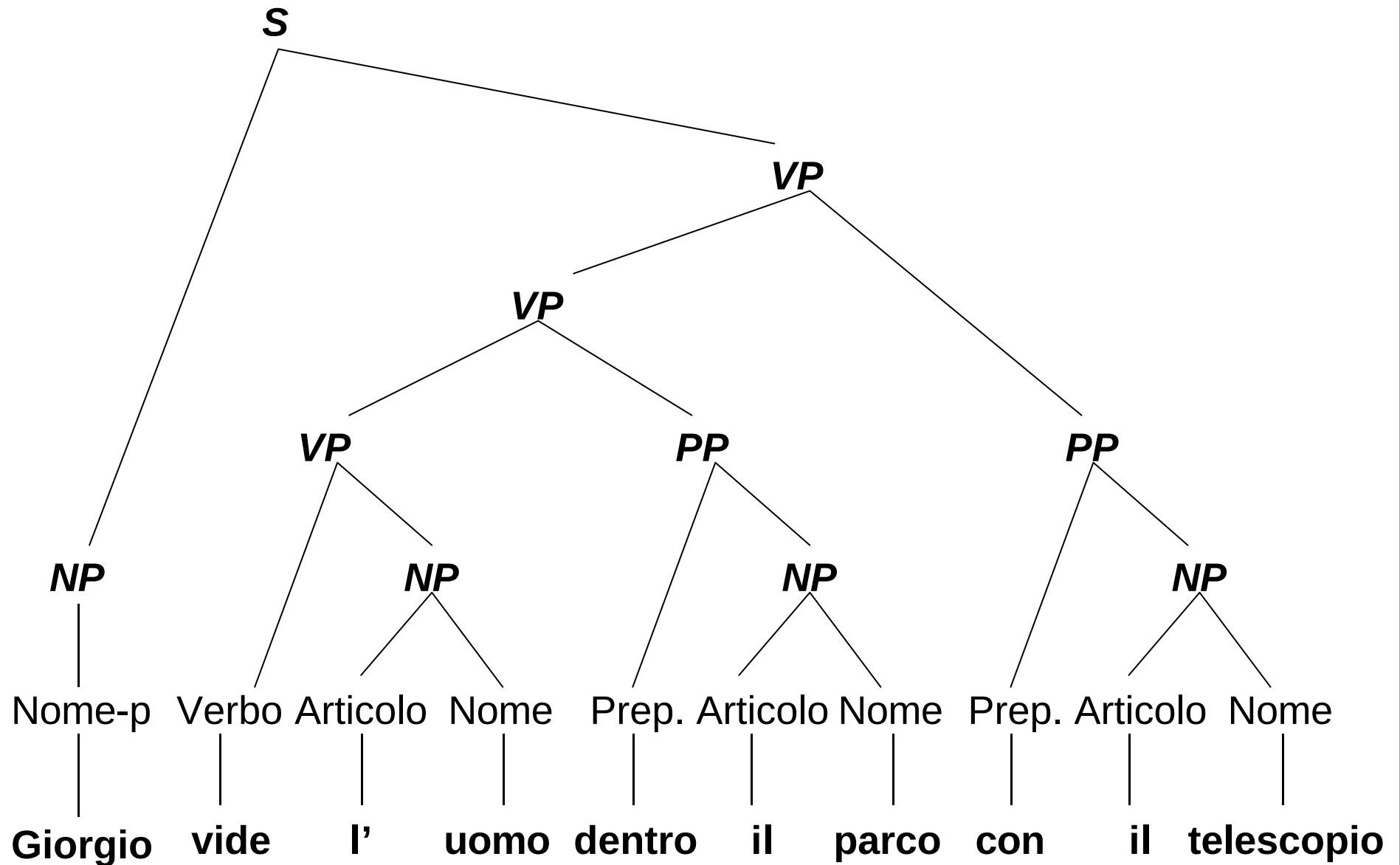
# Parsing di “Giorgio ama Maria”



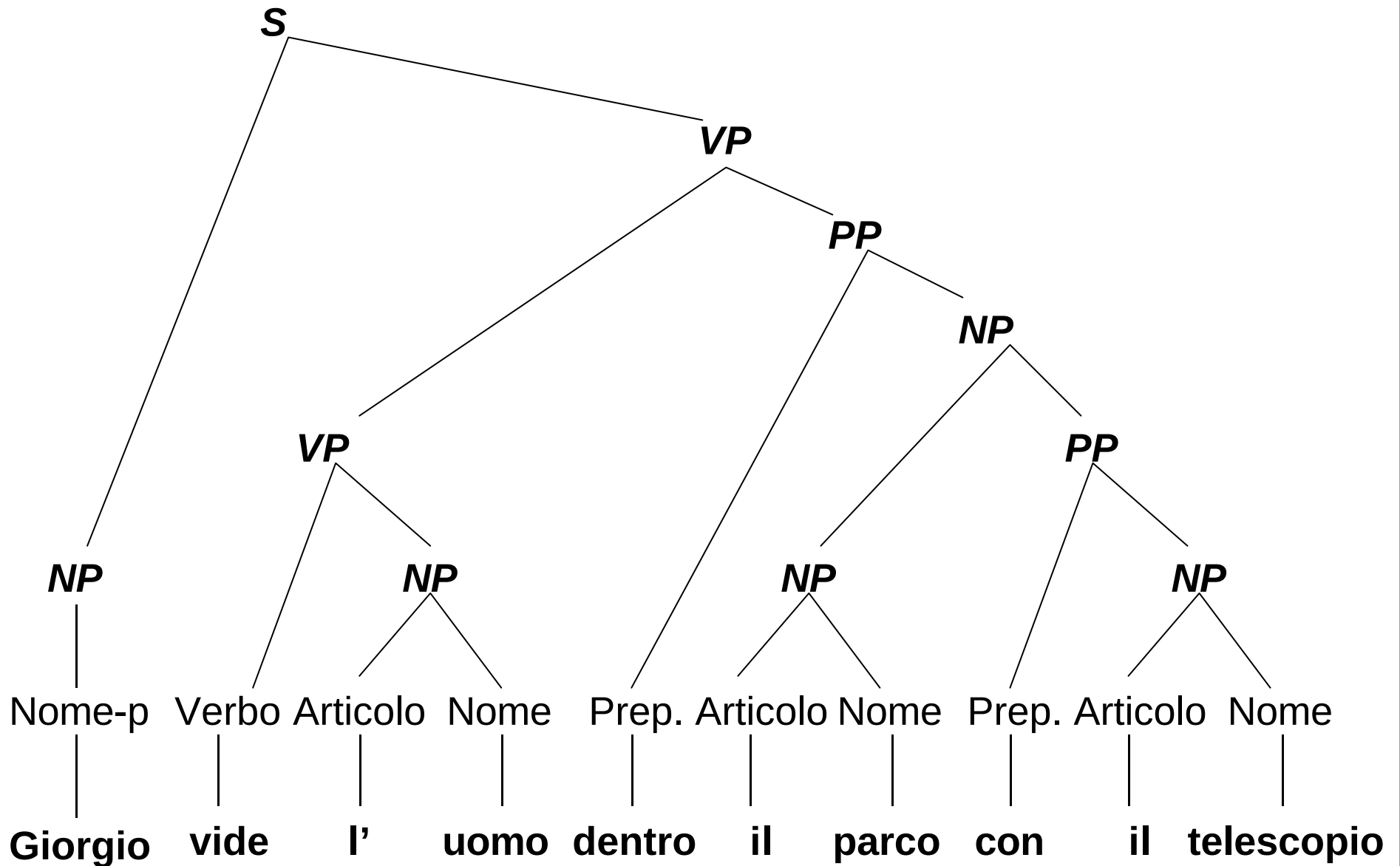
# **Una frase ambigua**

Giorgio vide l'uomo dentro il parco con il telescopio

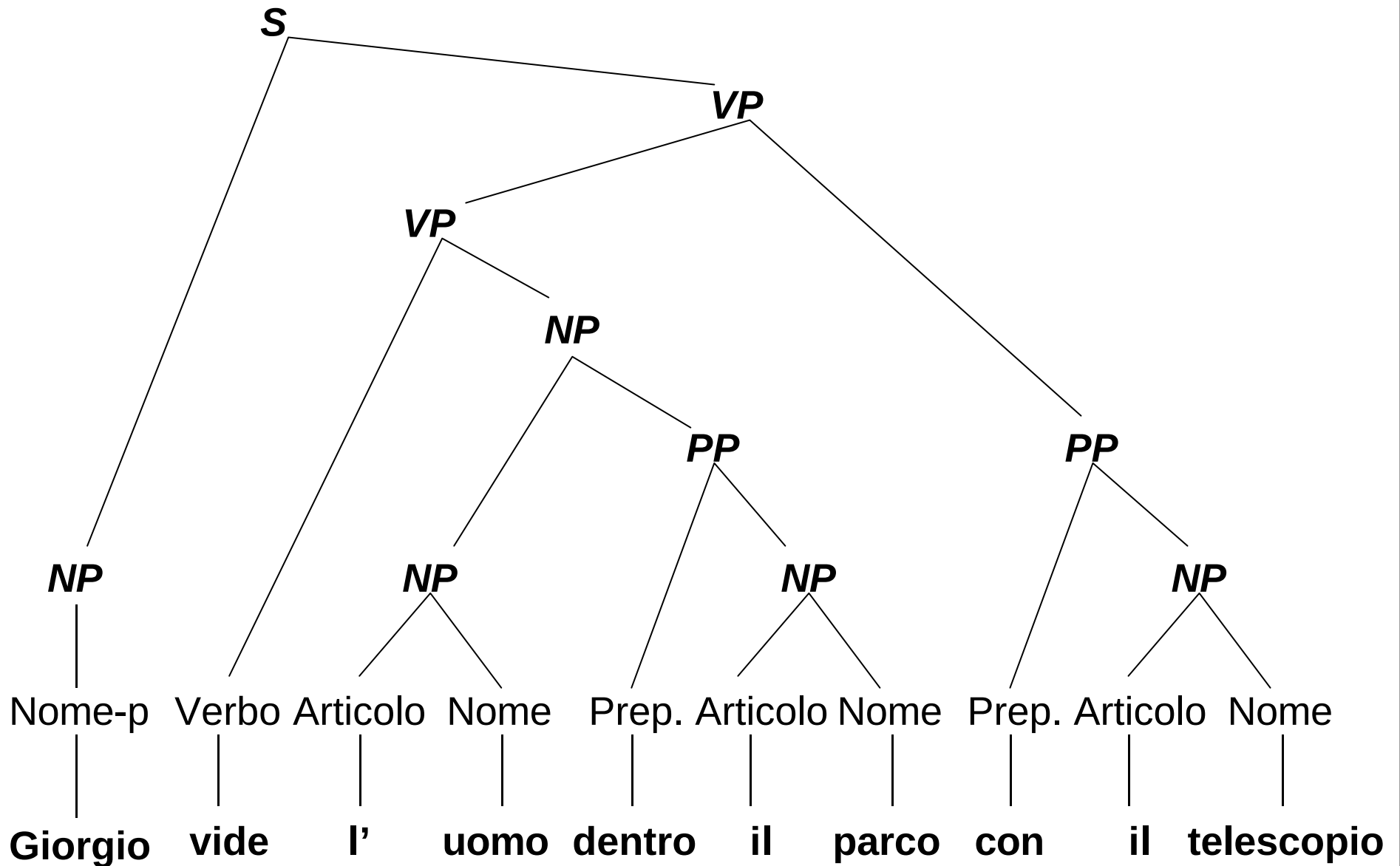
# Albero di derivazione (1)



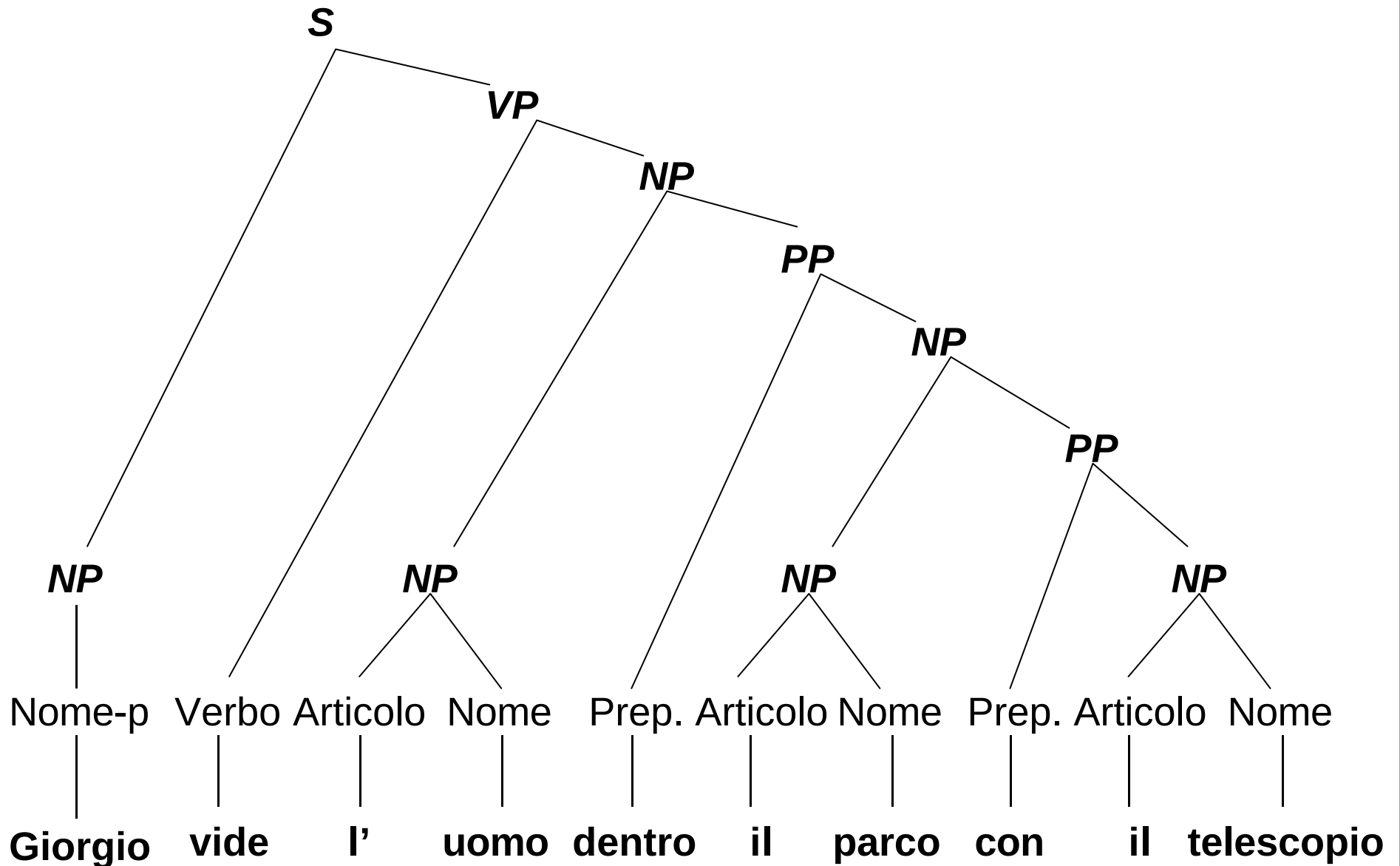
## Albero di derivazione (2)



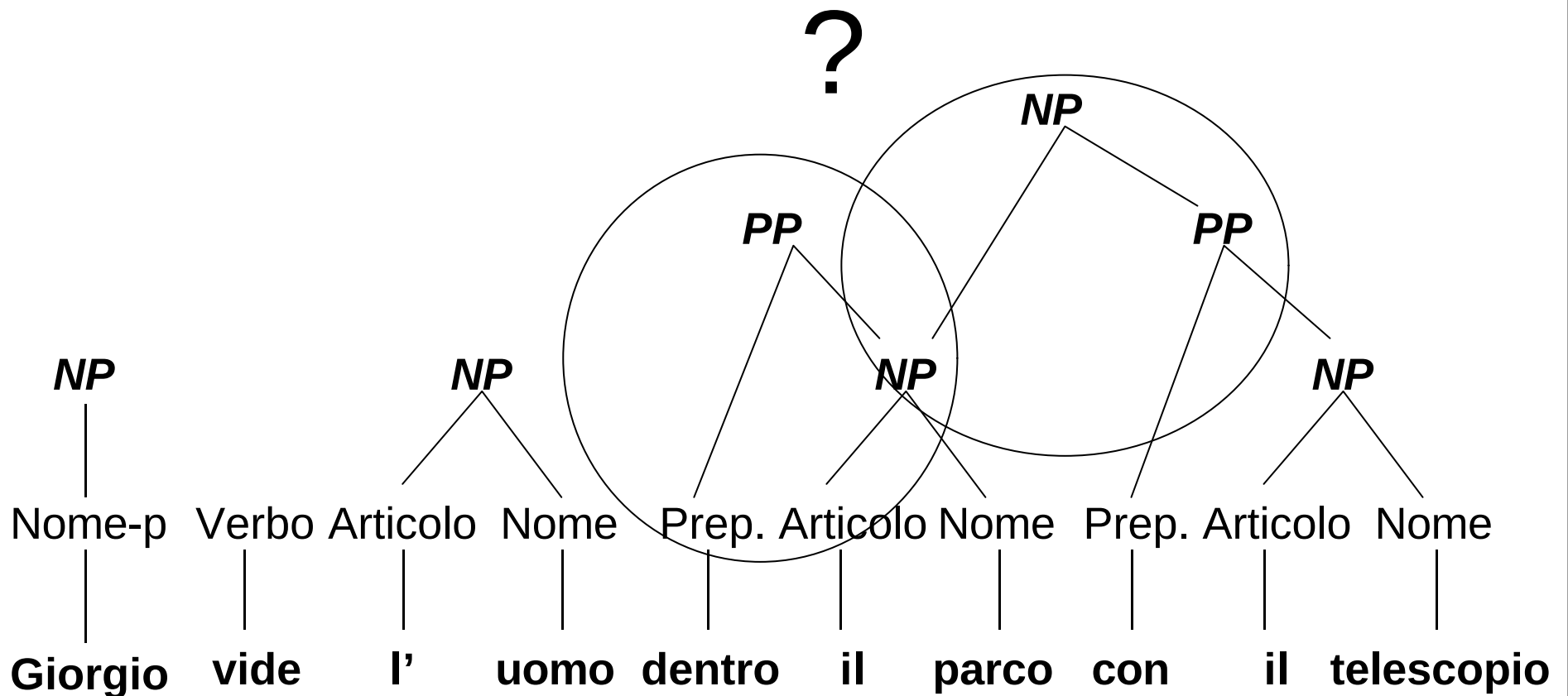
# Albero di derivazione (3)



# Albero di derivazione (4)



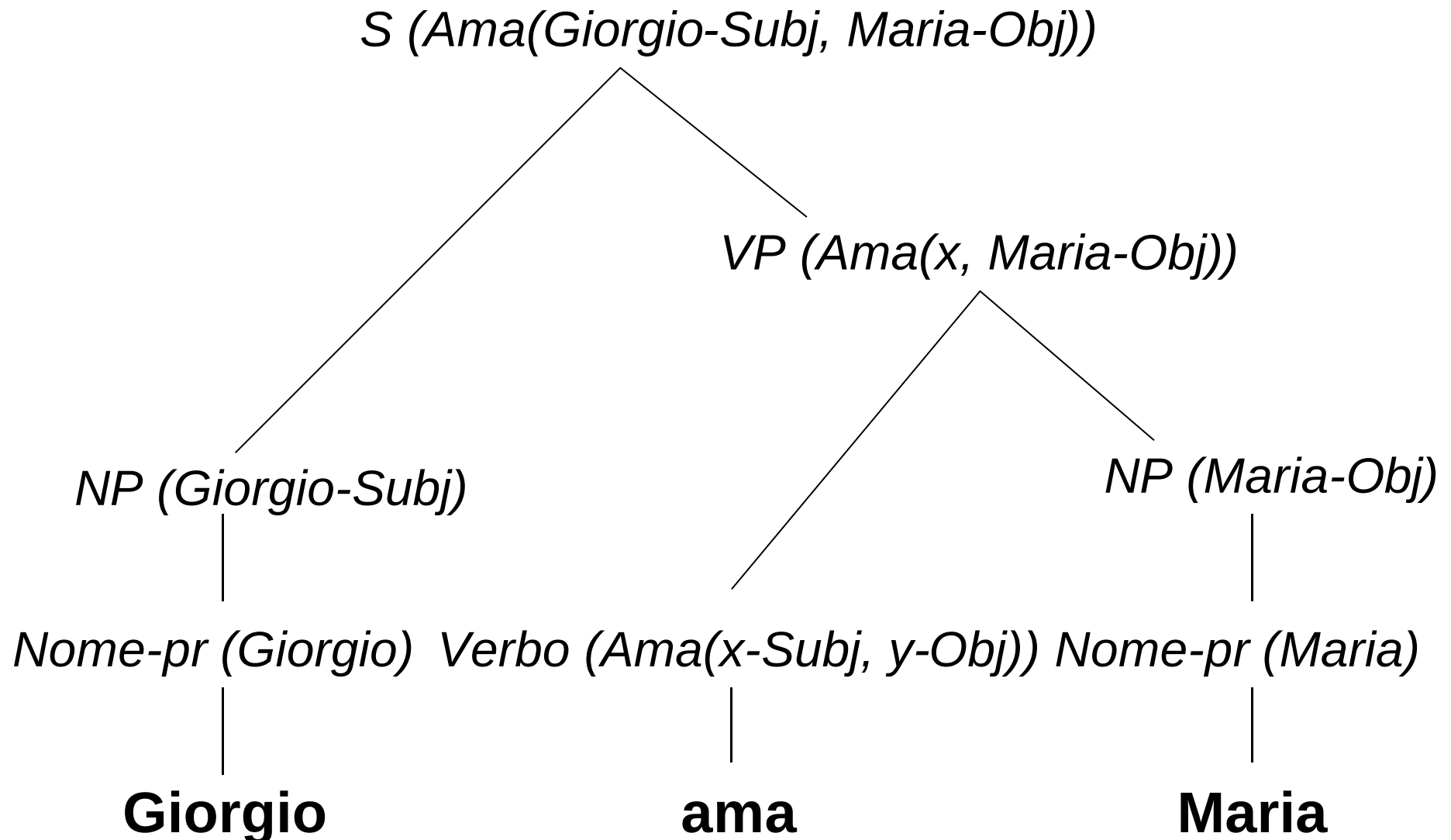
# Ambiguità nel parsing



# Livello semantico

- Obiettivo: assegnare alla frase una rappresentazione non ambigua del suo significato
- Fasi
  - recupero dei significati delle singole parole, un esempio, sinonimi
  - combinazione dei significati per un'intera frase
- Risultato: una formula logica che rappresenta il significato

# Interpretazione semantica di “Giorgio ama Maria”



# Esempio

“Ogni uomo ama una donna”

- 2 possibili forme logiche (interpretazioni):
- Per ogni uomo, esiste una donna che egli ama
- Esiste una donna che ogni uomo ama

# **Livello pragmatico**

- Obiettivo: interpretare la frase nel contesto del discorso e della situazione in cui viene enunciata
- Fasi
  - ricerca dei referenti
  - soluzione delle anafore e delle ellissi
  - modello dell'utente
- Risultato: un'interpretazione della frase in contesto

# Ambiguità lessicale

- Una parola ha più di un significato.
- brucia (bruciare, 3a persona, indicativo presente, intransitivo)
  - La carta brucia (“è un combustibile”)
  - La casa di Mario brucia (“ha preso fuoco”)
  - Il peperoncino brucia (“è piccante”)
  - La minestra brucia (“è troppo calda”)
  - La gola brucia (“causa dolore fisico”)
  - La condanna brucia (“causa dolore mentale”)

# Ambiguità lessicale

- NB. Non si è considerato tutto l'universo di “bruciare” (transitivo, riflessivo, ...)
  - La contessa brucia le sostanze (“consuma”)
  - L'acido solforico brucia i tessuti (“provoca ustioni”)
  - Mi sono bruciato ... al sole, la carriera, ...
- NB. Ambiguità lessicale trans-categoriale
  - “Tutti hanno un telefonino e a chi telefonino non si capisce” (da un quotidiano)
  - PESCA
    - » nome (il frutto, lo sport)
    - » verbo (lo sport, l'estrazione, ...)
    - » aggettivo (il colore)

# Ambiguità sintattica (o strutturale)

- occorre anche senza l'ambiguità lessicale
  - “Giorgio vide un uomo nel parco con il telescopio”
- l'ambiguità sintattica può causare un'ambiguità semantica
  - “Giorgio vide un uomo con un telescopio”
- anche l'ambiguità lessicale può causare una ambiguità semantica
  - “L'astronomo sposò una stella”
- ambiguità nell'assegnazione delle relazioni grammaticali
  - “Chi uccise il poliziotto?”

# Ambiguità semantica e pragmatica

- l'ambiguità semantica può sorgere anche senza ambiguità lessicale o sintattica
  - “la casa verde”
- ambiguità referenziale
  - soluzione delle anafore (“la mela”, “egli”, ...)
  - “la-mela-che-ho-mangiato-oggi-a-pranzo”
- ambiguità pragmatica
  - quando il parlante e l'ascoltatore sono in disaccordo sulla descrizione della situazione corrente
  - “Qui fa caldo”, “Il teatro è sulla destra”

# Ambiguità locali globali

- Una frase è ambigua localmente quando una sua parte può avere più analisi, ma solo una è valida data l'intera frase.
  - “I soldati, avvertiti del pericolo ...
  - ..., condussero il raid di mezzanotte.
  - ... i cittadini, condussero il raid di mezzanotte.
- Una frase è ambigua globalmente quando può avere più analisi tutte valide.
  - “La vecchia porta la sbarra”

# **Applicazioni pratiche**

# Applicazioni pratiche

- focalizzate su domini particolari piuttosto che testi senza restrizioni
- focalizzate su compiti particolari piuttosto che sulla comprensione completa del testo
- Traduzione automatica
- Accesso a data base
- Elaborazione di testi

# Traduzione automatica

Nessun risultato teorico fondamentale, ma sistemi di ausilio alla traduzione

Costi di start-up elevati (lessici da 20.000 a 100.000, grammatiche da 100 a 10.000 regole)

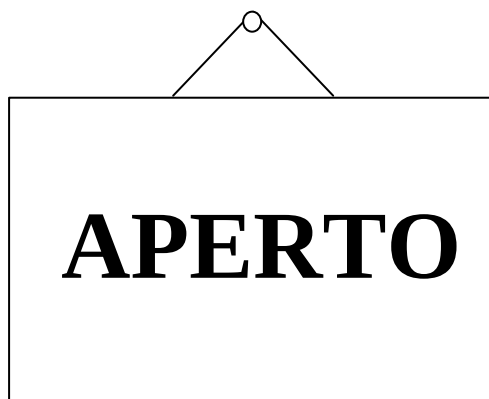
- Sistema TAUM-METEO (Università di Montreal)
  - rapporti meteorologici da inglese a francese
  - dominio ristretto + costrutti linguistici specifici
- Sistema SPANAM
  - traduzioni comprensibili, ma raramente fluenti
  - utile per post-editing
- Approccio pre-editing o sublanguage
  - “Caterpillar English”

# **Traduzione automatica**

La traduzione è difficile perché, nel caso generale, richiede una comprensione profonda del testo, e ciò a sua volta richiede una comprensione profonda della situazione che si sta comunicando.

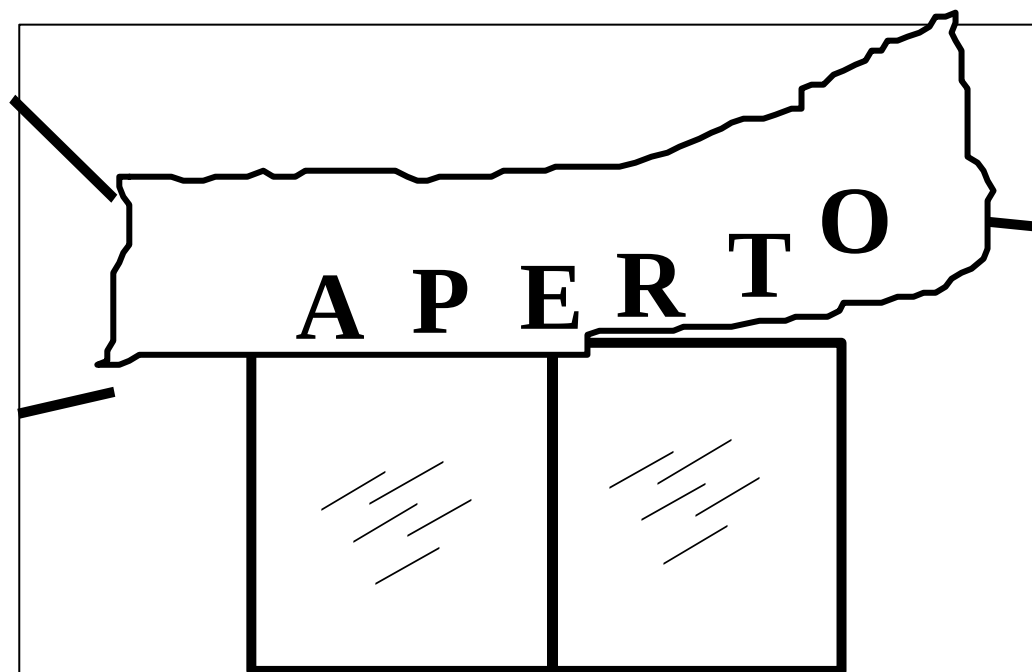
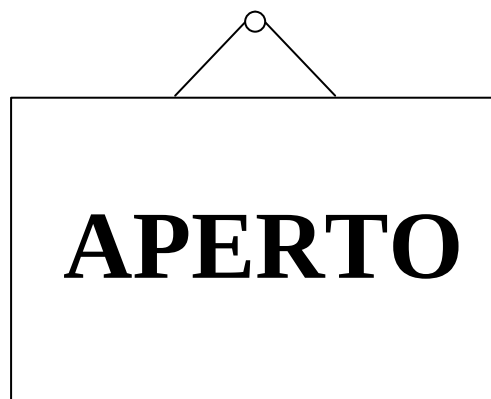
# Traduzione automatica

La traduzione è difficile perché, nel caso generale, richiede una comprensione profonda del testo, e ciò a sua volta richiede una comprensione profonda della situazione che si sta comunicando.



# Traduzione automatica

La traduzione è difficile perché, nel caso generale, richiede una comprensione profonda del testo, e ciò a sua volta richiede una comprensione profonda della situazione che si sta comunicando.



# Traduzione automatica

- linguaggi differenti categorizzano il mondo in modo differente
  - open                      aperto (nella maggioranza dei casi)
  - open questions      domande aperte
  - open market          mercato (aperto) (?)
  - running tap          rubinetto aperto
- il traduttore (uomo o macchina) deve ...
  - comprendere esattamente la situazione di riferimento
  - trovare un testo corrispondente nel linguaggio di destinazione che descriva al meglio la stessa situazione.
  - you                      lei? o tu?

# Accesso a data base

- Applicazioni comuni negli anni '70
  - interfacce complicate in linguaggi non amichevoli
  - interesse ridotto oggi: interfacce grafiche, fogli eletr., ...
- LUNAR (NASA - Woods 1973)
  - domande di geologi sulla composizione chimica delle rocce lunari portate a terra dalla missione Apollo
  - solo prototipo (78% di query risposte corrette)
- CHAT (Pereira 1983)
  - domande su un database geografico
  - qual è' la copertura del sistema?
  - problemi di gestione del discorso

# Elaborazione di testi

la maggior parte delle informazioni on-line è memorizzata in forma testuale

- e-mail, notizie, articoli, libri, enciclopedie
- troppa informazione da cui selezionare (in modo efficace)

- information retrieval
- text categorization
- information extraction

# Information retrieval

- selezionare da un insieme di documenti quelli rilevanti per una query (come combinazione di parole chiave)
- il contributo del LN a livello della parola
  - analisi morfologica e tagging (linguistica computazionale - computazione linguistica) e uso di content words
  - pesi di discriminazione: se un termine appare in pochi documenti ha un peso discriminativo maggiore
- poco uso della sintassi

# Classificazione di testi

- assegnare al testo una categoria tra un numero fissato a priori di categorie
  - servizi commerciali di informazioni
  - sistemi automatici classificano con correttezza del 90%
- può sfruttare le tecniche di LN perché l'insieme di categorie è fissato a priori
  - si può fare un tuning fine sul problema
  - Ex. parola “crude” (volgare), nel WSJ (grezzo 100%)

# Estrazione di informazioni

- elaborazioni affermazioni strutturate da testo on-line
- sistema SCISOR (General Electric 1990)
- ESEMPIO: N.Y. La Pillsbury ha guadagnato 3-4 punti sui 3.1 milioni di azioni dell'affare proposto, dopo che la britannica Grand Metropolitan ha alzato la sua offerta ostile di 3\$ per azione fino a 63\$. La compagnia ha immediatamente rifiutato lo zuccherino, che era venuto dopo che le due parti non si erano messe d'accordo su un'offerta più alta fatta in termini amichevoli nel weekend.
- TEMPLATE
  - Subevent: offerta aumentata, offerta rifiutata
  - Type: ostilità
  - Target: Pillsbury
  - Suitor: Grand Metropolitan
  - Share-Price: 63
  - Stock Exchange: Wall Street
  - Volume: 3.1 milioni
  - Effect-on-stock: {Rialzo: 3-4 punti}

**FINE**