Introduzione alla Linguistica Computazionale

Elena CABRIO

Université de Nice Côte d'Azur

elena.cabrio@unice.fr

UNIVERSITÉ CÔTE D'AZUR

Che cos'è la linguistica computazionale?

Introduzione

Che cos'è la linguistica computazionale?

Area di ricerca **interdisciplinare** a cavallo fra l'informatica, la linguistica, la psicologia e l'ingegneria che si occupa di:

1. mettere in grado i calcolatori di «comprendere» il linguaggio umano (NLP = Natural Language Processing, anche noto come TAL = Trattamento Automatico del Linguaggio)

1. studiare le proprietà formali e matematiche del linguaggio, interpretato come sistema di stringhe, quindi come un sistema matematico.

Finalità applicative della LC

- *Interrogare* una base di dati in linguaggio naturale, per iscritto o oralmente
- *Tradurre* automaticamente il linguaggio, sia parlato che scritto
- *Indicizzare*, *riassumere* automaticamente e poi effettuare *ricerche* su base semantica partendo da testo non strutturato
- Sviluppare filtri che riconoscono messaggi con contenuto inappropriato (es. *anti-spam*)
- Individuare automaticamente i casi di plagio
- Creare tecnologie di *supporto* alle disabilità (es. analisi della leggibilità dei testi)
- Analizzare opinioni, predire tendenze raccogliendo informazioni disponibili online

Finalità conoscitive della LC

- Creare un modello computazionale del modo in cui gli uomini usano il linguaggio:

1) Studiare come formalizzare le nostre conoscenze su un linguaggio = renderle *senza ambiguità* usando un linguaggio definito artificialmente

Es. "Adoro la pesca"

"La vecchia porta la sbarra"

"Giorgio vide un uomo nel parco con il telescopio"

2) Capire meglio il funzionamento delle strutture del linguaggio, dei meccanismi con cui gli esseri umano lo apprendono, lo producono e lo comprendono

I livelli di comprensione nella LC

Trattare il linguaggio naturale richiede l'analisi di vari *livelli* di comprensione/competenza:

- Livello lessicale: riguarda le convenzioni sulle singole parole
  - saporito, poritosa, *rtoisapo

- Livello sintattico: riguarda l'ordine corretto delle parole e i suoi effetti sul significato
  - il cane ha morso il bambino, il bambino ha morso il cane
  - idee verdi incolori dormono furiosamente
  - * morso ha bambino cane il il

I livelli di comprensione nella LC

Trattare il linguaggio naturale richiede l'analisi di vari *livelli* di comprensione/competenza:

- Livello semantico: riguarda il significato delle parole e delle frasi
  - la gola brucia, la casa brucia, la condanna brucia, la minestra brucia
  - *idee verdi incolori dormono furiosamente
- Livello pragmatico: riguarda il contesto comunicativo e sociale generale e i suoi effetti sull'interpretazione
  - questo è bello, il panino chiede un'altra birra

L’ambiguità è pervasiva

- Riconoscimento del parlato
  - “Lo scontro ha causato 10 **contusi**” - “Lo scontro ha causato 10 **confusi**”
- Analisi lessicale
  - “Tutti hanno un **telefonino** e a chi **telefonino** non si capisce”
- Analisi sintattica
  - “Ho mangiato gli spaghetti **con** la forchetta” - “Ho mangiato gli spaghetti **con** la pancetta”
- Analisi semantica
  - “Mi piace la **pesca** noce” - “Mi piace la **pesca** d’altura”
- Interpretazione semantica
  - “Ogni uomo ama **una** donna”

L’ambiguità è pervasiva

- Analisi del discorso
  - “Ha messo il carciofo nel piatto e **l’**ha mangiato”
  - “I saggi parlano perché hanno qualcosa da dire; gli sciocchi perché devono dire qualcosa”

- Analisi pragmatica
  - “Se smetti di fumare ti pago da bere” (Promessa)
  - “Se salti le lezioni ti metto in punizione” (Minaccia)

Un po’ di storia…

Nascita

- fondazione dell’Association of Computational Linguistics (ACL) nel 1962

Fisionomia

- pluralità di programmi di ricerca e metodologie
- interdisciplinarità e multidisciplinarità

Obiettivi

- applicazioni destinate a specialisti del linguaggio
- applicazioni informatiche di uso comune

Automi, algoritmi et modelli

Automa

- dal greco *autòmatos*, “che agisce da sé”
- macchine che, sulla base di istruzioni, eseguono un’azione o compiono atti di tipo linguistico

Calcolo

- alcune caratteristiche accomunano il pensiero e il linguaggio a calcoli e lingue algebriche
  - Thomas Hobbes, Gottfried Leibniz

Intelligenza artificiale

- costruire macchine che possano svolgere compiti linguistici

Il comportamento della macchina

Input

- L'*input* è pensabile come lo stimolo (sensoriale, linguistico, ecc.), o il dato, che viene fornito alla macchina per essere trattato

Output

- L'*output* è il comportamento che la macchina esibisce dopo aver ricevuto l'input: produzione di una risposta, un suono, una azione, un movimento, ecc.

Modello

- Il modello filtra l'input, lo analizza e vi associa, a seconda delle sue caratteristiche, mediante una serie di *algoritmi*, un output

Quali dati di input?

- **Dati strutturati:**
  - **Basi di dati**: le informazioni sono codificate in tabelle e sono accessibili tramite un apposito linguaggio di interrogazione. Esiste uno “schema” che permette di interpretare in modo non ambiguo i dati.
  - **Basi di conoscenza**: permettono anche di eseguire inferenze (ragionamenti).
- **Dati semi-strutturati**
  - Tabelle inserite in documenti o su Web
  - Directories di portali su Web (ad esempio Google e Yahoo!)
  - Documenti XML (Extensible Markup Language)
  - I dati sono parzialmente interpretabili

Quali dati di input?

- **Dati non strutturati:**
  - **Testi scritti** in vari formati
  - Documenti Word, pdf, Power Point
  - Giornali on-line
  - Pagine web in HTML
  - SMS
  - **Campi testuali** nelle basi di dati
  - **Messaggi di posta elettronica**
  - Messaggi sulle news group
  - Frequently Asked Questions (**FAQ**)
  - **News** di agenzie
  - **Trascrizioni** automatiche di tele o radio giornali

Quali dati di input?

- **Dati multilingui e multimediali:**
  - **Dati multilingui** in vari formati
    - **Siti multilingui** con lo stesso testo disponibile in pagine diverse dedicate.
    - Testi al cui interno compaiono sezioni in lingue diverse.
    - **Traduzioni**, per esempio manuali d'uso di prodotti.

- **Informazioni multimediali**
  - **Immagini** inserite all'interno di un testo, eventualmente con una didascalia
  - **Filmati**
  - **File audio,** con messaggi parlati

Processamento di dati: Quali esigenze?

Alcuni esempi …

1. **Trovare** informazione contenuta in fonti di tipo testuale.
2. **Estrarre** informazione contenuta in formato testuale.
3. **Organizzare** documenti in formato testuale.
4. **Costruire reti di utenti** basate sul loro interesse verso alcuni documenti (vedi recenti studi sui social media e le reti sociali)

Scoprire l'informazione

- **Recuperare** informazione (*information retrieval*): l'utente sottomette una richiesta (*query*) e ottiene documenti rilevanti per quella richiesta.
- **Cross-language retrieval**: la *query* è in una lingua diversa da quella dei documenti.
- **Question Answering**: la *query* è una domanda in linguaggio naturale, la risposta è una porzione di testo.
- **Tradurre** documenti da una lingua ad un'altra.

Estrarre informazioni

- **Riassumere** il contenuto di un documento utilizzando poche frasi significative.
- **Riempire degli schemi (*template*) prefissati**, con informazioni del tipo chi, dove, quando, …
- **Selezionare i termini rilevanti** da un insieme di documenti, ad esempio per costruire l'indice tematico di un libro.

Organizzare le informazioni

- **Categorizzazione** dei testi: assegnare una certa categoria ad ogni documento di una collezione.
- **Raggruppare** (*clustering*) i documenti in gruppi omogenei per contenuto. Ad esempio per estrarre opinioni e giudizi relativamente ad un certo prodotto.
- **Individuare l'argomento** (*topic*) di un documento, ad esempio di un messaggio di posta elettronica, per poterlo inviare ad un destinatario appropriato.
- **Classificare** documenti in una gerarchia di concetti.

Costruire reti di utenti

- Gli utenti vengono classificati rispetto al loro interesse per certi documenti.
- **Modellizzazione dell'utente**: viene costruito un profilo personalizzato, che viene poi usato per proporre nuovi documenti.
- **Sistemi di raccomandazione** di documenti.

Estrazione di informazioni

La Rolo banca 1473 ha reso noto che al 31 agosto sono state collocate obbligazioni del prestito "Rolo Banca 1473 Spa a tasso fisso convertibile 1996/1999 - prima emissione" per 522.15 miliardi. Il prestito, di durata triennale - spiega una nota della banca - ha cedole semestrali predeterminate (prima cedola 4.15%, cedole successive 3.90%) con facoltà per l'emittente di conversione a tasso variabile indicizzato al Libor Lira 6 mesi.

SCHEMA OBBIGAZIONE

TIPO-OBBIGAZIONE	*Prestito Obbligazionario*
EMITTENTE	*Rolo Banca 1473 Spa*
QUANTITA'	*522.15 miliardi*
DATA-EMISSIONE	*31 agosto*

Question answering (1)

- Trovare la risposta ad una domanda all'interno di una collezione di testi

"Qual è la **stella** più **luminosa** visibile dalla Terra?"

1. **Sirio** è la più **brillante stella** visibile dalla Terra pur essendo una ….
2. **Stefania Sandrelli**, la **stella** più **brillante** del panorama cinematografico italiano, ….

Question answering (2)

- **Scoprire relazioni implicite tra domanda e risposta**

*Chi è l'**autore** de "I promessi sposi"?*

…Alessandro Manzoni **scrisse** "I promessi sposi" nel 1840.

…la regista Roseanne Barr ha **messo in scena** la rappresentazione dei "*Promessi sposi*" nel 1978 …

Question answering (3)

- **Scoprire relazioni implicite tra domanda e risposta**

Quale è la data di nascita di Mozart?

…. Mozart (**1751** – 1791) ….

Question answering (4)

- **Scoprire relazioni implicite tra domanda e risposta**

Quale è la distanza tra Napoli e Ravello?

"Dall'aereoporto di Napoli seguire le indicazioni Autostrade (segnali verdi). Proseguire in direzione Salerno (A3). Guidare per circa 6 Km. Pagare il pedaggio (1.20 Euro). Guidare ancora per circa 25 Km. Lasciare l'autostrada a Angri (uscita Angri). Girare a sinistra, seguire le indicazioni per Ravello. Guidare per circa due Km. Girare a destra, seguire le indicazioni per "Costiera Amalfitana". Dopo 100 metri si arriva ad un semaforo prima di un ponte molto stretto. State attenti a non perdere il prossimo cartello "Ravello" a circa 1 Km. dal semaforo. Ora potete rilassarvi e godervi il panorama (seguite questa strada per 22 Km.). Arrivati a Ravello …."

IBM Watson Jeopardy

- https://www.youtube.com/watch?v=WFR3lOm_xhE

Parametri delle applicazioni computazionali

Parametri di valutazione

Robustezza

- È la capacità dell'applicazione di gestire materiale linguistico in input contenente *rumore*
- e di accettare e analizzare input parziali o incompleti

Potenza

- Descrive la capacità di copertura della lingua dell'applicazione, il suo raggio di azione
- Considera «quanto» della lingua viene trattato accuratamente dall'applicazione

Portabilità

- È la possibilità di applicazione a nuovi domini (altre lingue, altri linguaggi settoriali, altre tipologie testuali), modificandone al minimo la struttura

Generalizzabilità

- È la capacità del modello computazionale di dare conto di fenomeni linguistici nuovi, applicando modelli desunti da materiale linguistico relativamente ridotto

Parametri delle applicazioni computazionali

Programmazione

Economia di programmazione

- la semplicità/complessità del modello, il tempo che ci vuole per perfezionarlo

Complessità della computazione

- quante risorse vengono adoperate per giungere a un output e quanto tempo ci vuole per il processamento delle informazioni

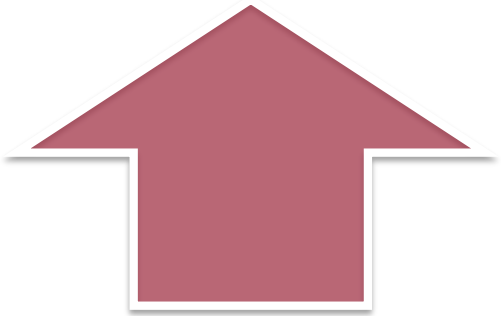
Principali approcci all'analisi del testo

Approcci di tipo statistico

- dati estratti da testi reali
- metodo induttivo
- riproduzione di comportamenti che simulano le tendenze proprie della produzione linguistica concreta

Approcci basati su regole

- di tipo «grammaticale»
- serie di condizioni necessarie e sufficienti a specificare una data produzione
- metodo deduttivo



Principali approcci all'analisi del testo

1) Approcci basati su regole, knowledge-rich

- La rappresentazione del dominio è esplicita, espressa sotto forma di regole.
- Tale rappresentazione corrisponde alla conoscenza di un esperto della materia

Es. Estrazione di relazioni in campo biomedico
"*L'alcol interagisce con le proteine cerebrali…*"

Principali approcci all'analisi del testo

1) **Approcci basati su regole: Problemi**

- E' molto difficile concepire tutto il sistema di regole necessario a fornire a un calcolatore le conoscenze linguistiche per l'elaborazione del linguaggio.
- E' anche molto difficile gestire la complessità e le interazioni del sistema di regole.

Principali approcci all'analisi del testo

1) **Approcci basati su regole: Problemi**

- E' molto difficile concepire tutto il sistema di regole necessario a fornire a un calcolatore le conoscenze linguistiche per l'elaborazione del linguaggio.
- E' anche molto difficile gestire la complessità e le interazioni del sistema di regole.

Soluzione: Invece di un esperto che fornisce al calcolatore le informazioni linguistiche sotto forma di regole, l'esperto annota un testo con informazione linguistica e il programma impara da solo le regole e il loro uso.

Principali approcci all'analisi del testo

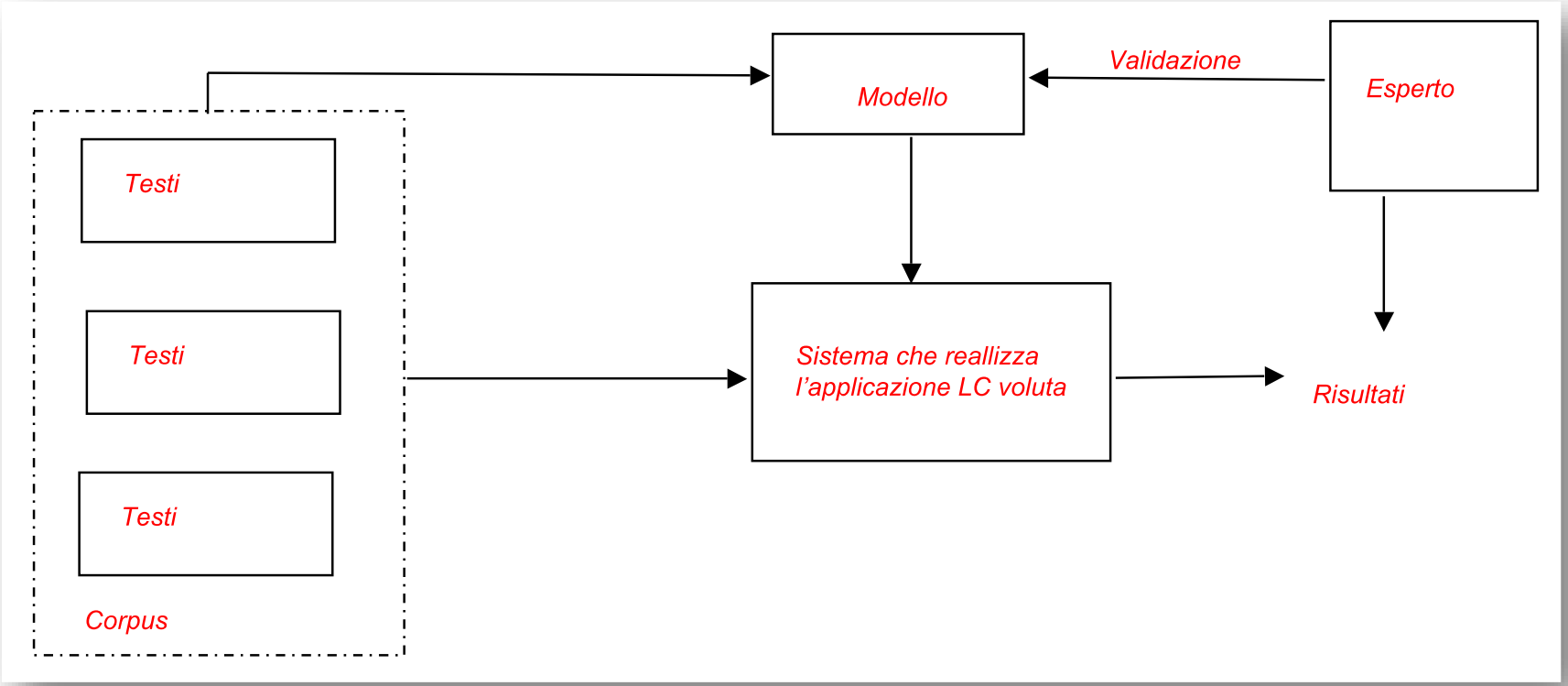
2. Approcci statistici (data-driven, knowledge-poor)

- La rappresentazione della conoscenza del dominio è implicita, espressa sotto forma di annotazione di un testo o di un corpus.
- Un programma impara automaticamente le regole e la loro frequenza d'uso nel testo
- Sono usati **modelli probabilistici**, che descrivono il comportamento delle parole nei testi (*GoogleTranslate)*

Es. Estrazione di relazioni in campo biomedico
"*L'alcol interagisce con le proteine cerebrali…*"

Principali approcci all'analisi del testo

2. Approcci statistici (data-driven, knowledge-poor)



Principali approcci all'analisi del testo

2. Approcci statistici: Problemi

- E' molto difficile e anche costoso costruire risorse linguistiche rappresentative in quantità sufficiente.
- Non si cerca più di riprodurre la competenza linguistica con modelli che formalizzano le nostre facoltà di comprensione linguistica, ma si cerca di riprodurre, per una classe di applicazioni data, la *performance linguistica associata.*
- Questo lo si fa con modelli automaticamente estratti dai dati, che devono essere in *grandi quantità* e *caratteristici* della applicazione voluta.

Principali approcci all'analisi del testo

2. **Approcci statistici: Vantaggi**
   - Permettono di cogliere regolarità presenti in collezioni di testi di grandi dimensioni.
   - Vengono osservati fenomeni "oggettivi" riguardo ad una lingua, che possono sfuggire all'analisi "soggettiva" praticata da linguisti.

Principali approcci all'analisi del testo

2. Approcci statistici: Vantaggi

- *Acquisizione*: identificazione e codifica automatica delle conoscenze necessarie
- *Copertura*: si coprono automaticamente tutti i fenomeni linguistici nel dominio di applicazione.
- *Robustezza*: adattamento più facile al "rumore" e ai dati imprevisti
- *Portabilità*: adattamento più facile ad una nuova lingua.
- *Valutazione*: si arriva a una valutazione sperimentale dei sistemi e delle ipotesi scientifiche

Algoritmi di apprendimento

- Definizione: Un programma *apprende* a partire da un esperimento di addestramento A per eseguire il compito C valutato da una misura di performance P, se la performance P al compito C migliora in seguito all'esposizione ad A.

- **Esempio**

Compito C: classificare i verbi in classi predefinite

Esperimento di addestramento A: base dati di coppie di verbi con i loro attributi e le risposte corrette

Misura di performance P : % di nuovi verbi classificati correttamente (rispetto a una classificazione stabilita da un esperto)

Apprendimento per classificazione

- Il compito più studiato in apprendimento automatico (*machine learning*) consiste nell' inferire una funzione che assegna gli esempi rappresentati come vettori di tratti distintivi ad una classe fra un insieme finito di categorie date.

Esempio

- Sia dato un insieme di verbi.
- Compito: classificazione *binaria*: verbi di movimento (es. *correre, saltare, passeggiare*) e verbi di cambiamento di stato (es. *fondere, cuocere*).
- Proprietà: per ogni volta che troviamo il verbo nel corpus è transitivo? è passivo? Il suo soggetto è animato?

Apprendimento per classificazione: un esempio

Esempio	Trans?	Pass?	Anim?	Classe
correre	5%	3%	90%	MoM
saltare	55%	5%	77%	MoM
fondere	10%	9%	20%	CoS
cuocere	80%	69%	88%	CoS

- Se Pass < 9% e Anim > 20% allora il verbo è MoM, altrimenti CoS

- Come classificare un nuovo verbo?
- "passeggiare": Trans 2%, Pass 1%, Anim 90% ⟶ MoM

Una questione aperta: L'incalcolabilità delle lingue

Calcoli

- un *calcolo*, per essere definito tale, prevede una serie di condizioni tra cui la presenza di un inventario di simboli finito e di un insieme finito di regole di combinazione dei simboli in stringhe/segni del linguaggio

…ma le lingue non sono *completamente* calcolabili

La potenziale infinitezza dei segni

- Le lingue naturali, così come i calcoli, possono produrre un numero potenzialmente infinito di segni
- Posso creare sempre nuove frasi, e posso creare nuovi lessemi, nuove parole, che esprimano nuovi significati
- Non vi è limite di lunghezza nella produzione dei segni
- L'inventario delle unità di prima articolazione (i morfi, dotati di significante e significato) è aperto

L’importanza della valutazione

- È importante poter valutare in modo sperimentale (**replicabile**) i risultati ottenuti su un certo compito.
- Le prestazioni di un algoritmo vengono verificate rispetto al comportamento degli umani (*gold standard*).
- Gli algoritmi vengono migliorati fino a che non si avvicinano ai giudizi degli umani.
- Lo studio della valutazione e delle metriche di valutazione di sistemi di trattamento automatico del linguaggio è una parte fondamentale della Linguistica Computazionale
- Ogni sistema viene valutato confrontandolo con lo “stato dell’arte” = peformance del sistema che ha ottenuto fino a quel momento risultati migliori su un dato *gold standard*

Metriche di valutazione

- Calcolare la performance di un sistema relativamente ad un task specifico
- La valutazione quantitativa considera:
  - Un task da svolgere in modo automatico
  - Un insieme di dati per
    - ``Allenare'' (parametrare) i modelli (training set)
    - Migliorare i modelli (development set)
    - Valutare i modelli (test set)

  ➡ Effort per preparare i dati ➡ Evitare di influenzare i dati
- Correlazione tra la modellizzazione e la sua valutazione
  - Per un task, le performances possono variare di molto a seconda di:
    - Tipo di dati (topic,qualita', etc.)
    - Metrica di valutazione usata

➡ Tenere in considerazione le condizioni in cui viene effettuata la valutazione

➡ Ricerca dell'oggettivita'

Valutazione di un task di classificazione

- (*altro esempio*): a partire da un insieme di testi, trovare solo quelli **pertinenti** rispetto ad una **classe** definita
- Valutare l'abilita' di un sistema di trovare i testi **pertinenti**, e **solamente** quelli

- Quanto un sistema fornisce una risposta relativamente ad un documento e ad una classe, ha due scelte:
  - Il documento **appartiene** secondo lui alla classe
  - Il documento **non appartiene** secondo lui alla classe

- Rispetto a queste due possibilita' di risposta, esistono due casi:
  - Il documento **appartiene** alla classe
  - Il documento **non appartiene** alla classe

Valutazione di un task di classificazione

Nome del caso	Abbreviazione	Descrizione
Vero positivo	TP	Il sistema identifica **correttamente** il documento come **appartenente** alla classe
Falso positivo	FP	Il sistema identifica **erroneamente** il documento come **appartenente** alla classe
Vero negativo	TN	Il sistema identifica **correttamente** il documento come **non appartenente** alla classe
Falso negativo	FN	Il sistema identifica **erroneamente** il documento come **non appartenente** alla classe

Valutazione di un task di classificazione

- Metriche per il calcolo delle performances: le **recall** et la **precisione**
- **Recall:** numero di documenti pertinenti trovati dal sistema, rispetto al numero di documenti pertinenti nell'insieme dei testi

$$recall = \frac{numero\ di\ documenti\ correttamente\ attribuiti\ \ alla\ classe\ i}{numero\ di\ documenti\ appartenenti\ \ alla\ classe\ i}$$

- **Precisione**: numero di documenti pertinenti identificati, rispetto al numero totale di documenti proposti dal classificatore come appartenenti ad una certa classe

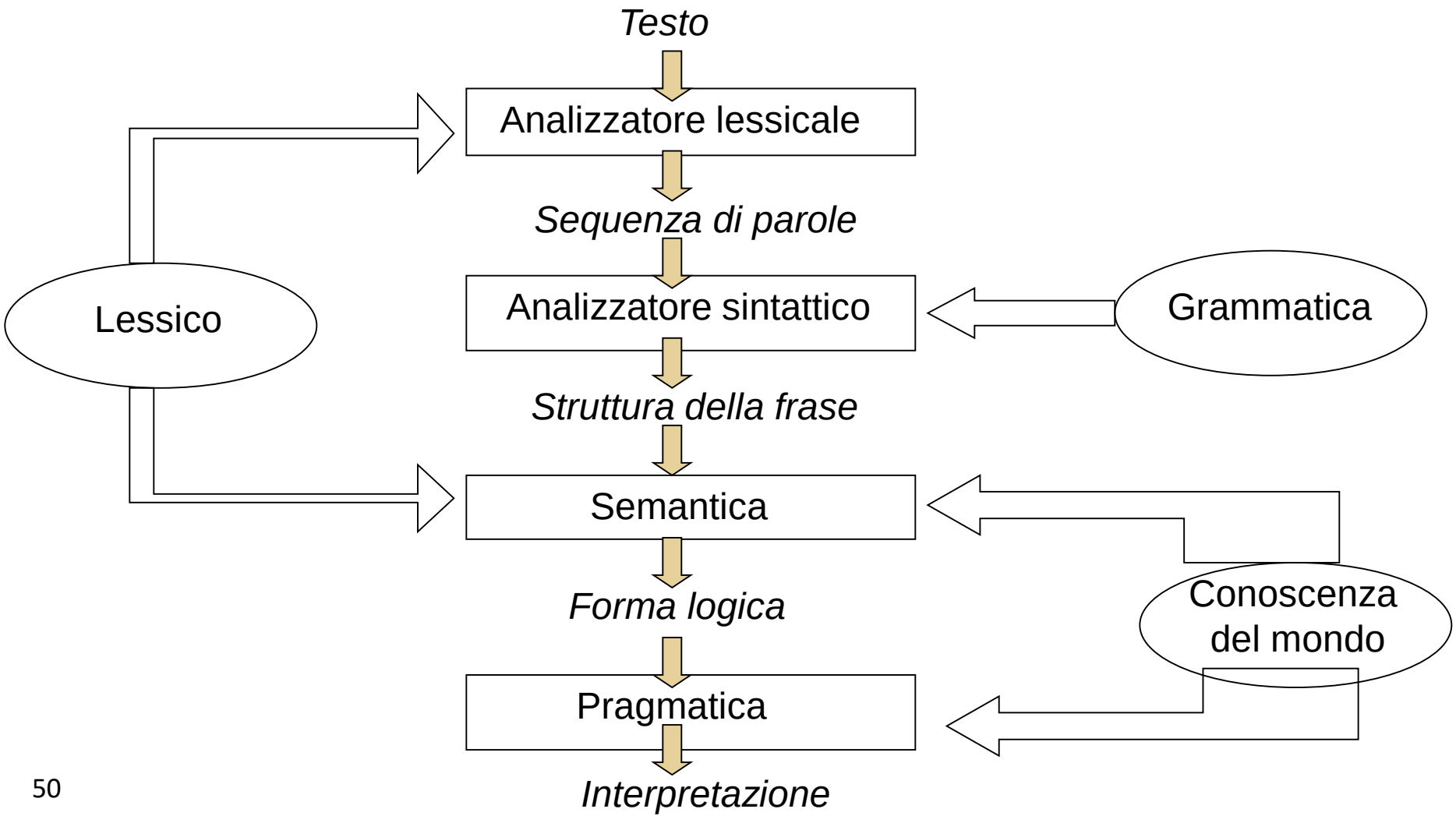
$$precisione = \frac{numero\ di\ documenti\ correttamente\ attribuiti\ \ alla\ classe\ i}{numero\ di\ documenti\ attribuiti\ alla\ classe\ i}$$

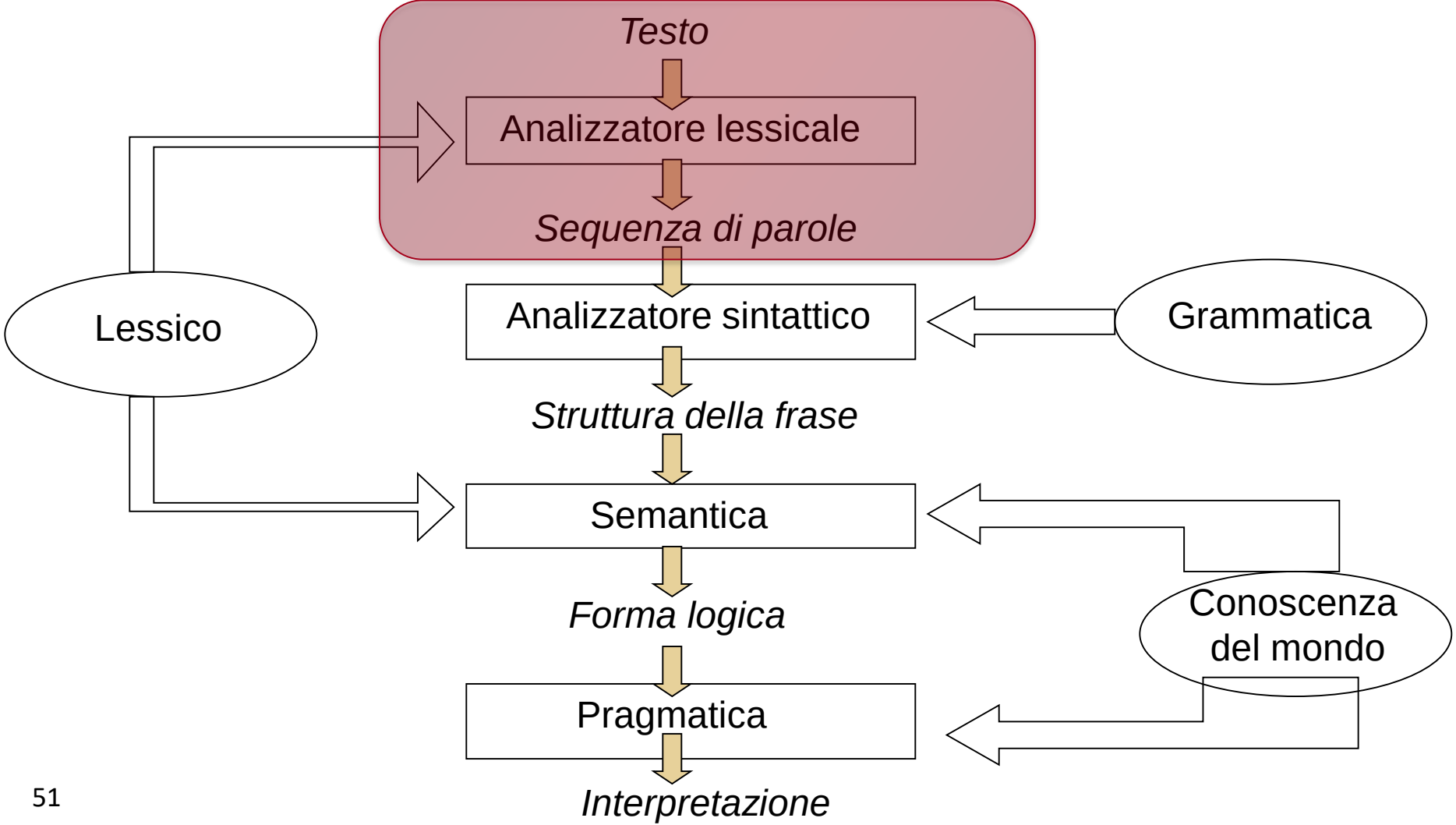
Valutazione di un task di classificazione

- Metrica standard che combina precisione e recall e la loro media armonica, chiamata F-mesure o F-score

$$\text{F-measure} = 2 \cdot \frac{precisione \cdot recall}{precisione + recall}$$

Analisi completa di un testo


Testo
Analizzatore lessicale
Sequenza di parole
Analizzatore sintattico
Grammatica
Lessico
Struttura della frase
Semantica
Forma logica
Conoscenza del mondo
Pragmatica
Interpretazione


Testo
Analizzatore lessicale
Sequenza di parole
Lessico
Analizzatore sintattico
Grammatica
Struttura della frase
Semantica
Forma logica
Conoscenza del mondo
Pragmatica
Interpretazione

Analisi lessicale

Parole, tokens, forme, lemmi…

- Unità ``logiche'' per il trattamento dei testi:
  Documento ⊃ paragrafo ⊃ frase ⊃ ``parola'' ⊃ ``carattere''

- Ma una « parola » non e' un'unità ben definita :
  - Esempi : aereo, mangiato, molto, Roberto, AVIS, 42...

Analisi lessicale

- **Forma :** nozione grafica di parola (Igor Mel'čuk)
- **Lemma :** intersezione tra una forma (grafica) e un senso, a volte per composizione di morfemi
- **Morfema :** la più piccola unità portatrice di senso (per es. « ri »)
- **Token** (gettone) : unità minima di informazione identificata attraverso l'« analisi lessicale » o « tokenizzazione » (lessema)

Tokenizzazione

- **Segmentare un testo in « unità minime » per poterle processare**
- Insieme di automi in grado di riconoscere i token definendo delle stringhe di caratteri
  - Lessemi : -?[A-Z] ?[a-z]*
  - Punteggiatura : .|...|,|!|?
  - Numeri : -?[0-9]*(,|.)[0-9]*
  - ...

Example:

Gli studenti, quelli di Bologna, hanno tutti 29 di media ?

Gli **|** studenti **|** , **|** quelli **|** di **|** Bologna **|** , **|**hanno **|** tutti **|** 29 **|**di **|** media **|** ?

Lemmatizzazione

- **Lemma :** unità autonoma (composta da morfemi) che costituisce il lessico di una lingua
  - **Morfemi**: le « parti » di un lemma
  - **Autonomo** : può essere utilizzato così com'è in una frase
- **Lemmatizzazione, trovare i lemmi per ogni token di una frase**

Example:

Parto in luna di miele.

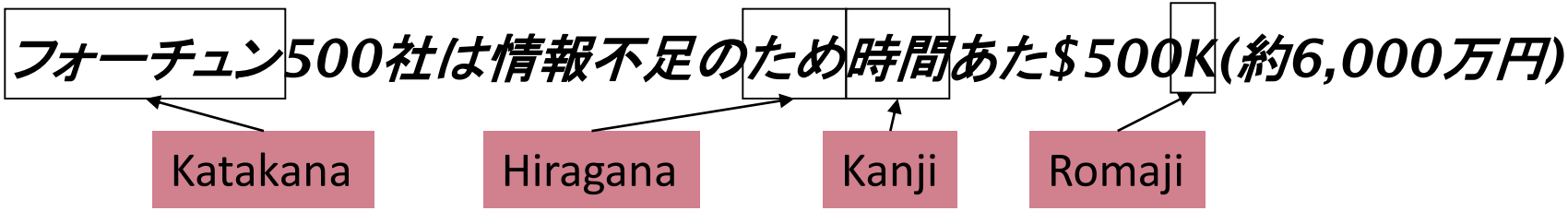
partire **|** in **|** luna di miele

Tokenizzazione: problemi

- Italiano
  - ***L'insieme*** $\rightarrow$ un token o due?
    - ***L*** ? ***L'*** ? ***Il*** ?
    - Vogliamo che ***l'insieme*** possa corrispondere a ***un insieme***
- Le espressioni in tedesco non sono segmentate
- ***Lebensversicherungsgesellschaftsangestellter***
  - 'impiegato di una compagnia di assicurazioni-vita'

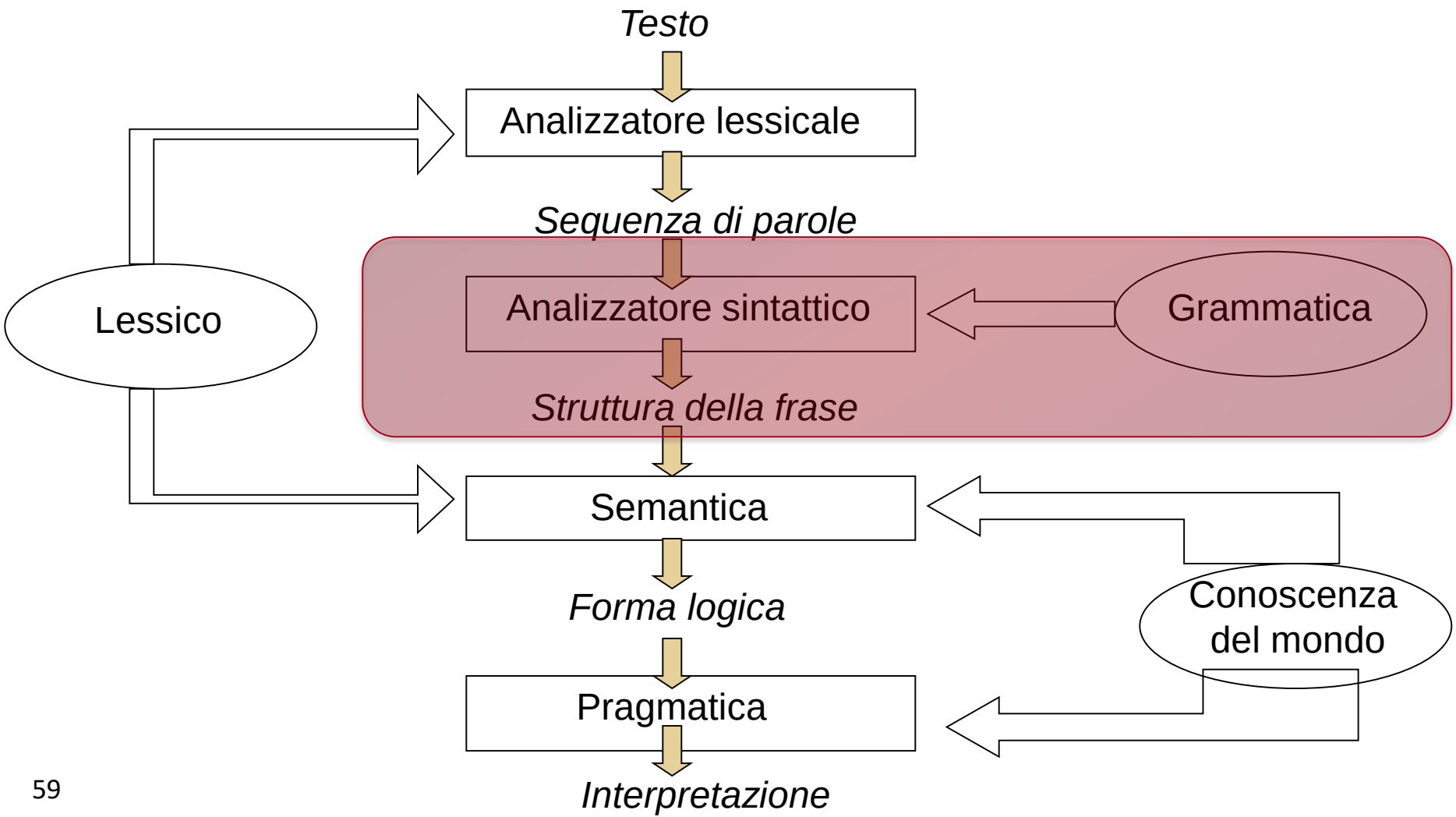
Tokenizzazione: problemi

- Cinese e Giapponese:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - 莎拉波娃 现在 居住 在 美国 东南部 的 佛罗里达
  - Sharapova vive adesso nel sud-est degli Stati Uniti, in Florida
- Più complicato in Giapponese, più alfabeti
- Date/quantità in formati diversi



Normalizzazione

- Bisogno di «normalizzare» i termini
  - Ricerca di informazioni: il testo indicizzato e i termini usati nella richiesta devono avere la stessa forma (es. Spa e S.P.A)
- Definiamo implicitamente le classi di equivalenza dei termini
  - Per esempio, sopprimendo i punti in una parola (es. M.)
- Alternativa: espansione asimmetrica:
  - Termine indicizzato: ***window*** -> Ricerca: ***window***, ***windows***
  - Termine indicizzato : ***windows*** -> Ricerca: ***Windows, windows, window***
  - Termine indicizzato : **Windows** -> Ricerca: **Windows**
- Potenzialmente più potente, ma meno efficace

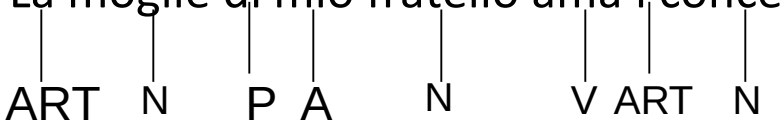

Testo
Analizzatore lessicale
Sequenza di parole
Analizzatore sintattico
Grammatica
Lessico
Struttura della frase
Semantica
Forma logica
Conoscenza del mondo
Pragmatica
Interpretazione

Dalla stringa all'albero di parsing

- Il *Parsing* è l'analisi sintattica di una frase ottenuta automaticamente da un sistema, il *parser*.

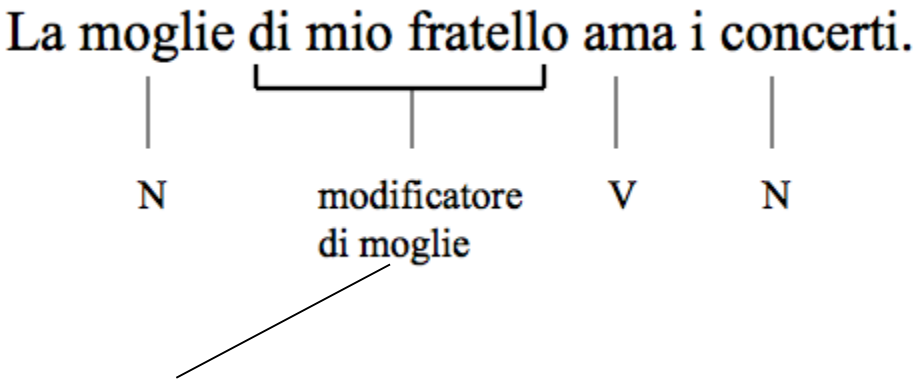
- Partiamo da un esempio: Voglio analizzare la frase

"La moglie di mio fratello ama i concerti"



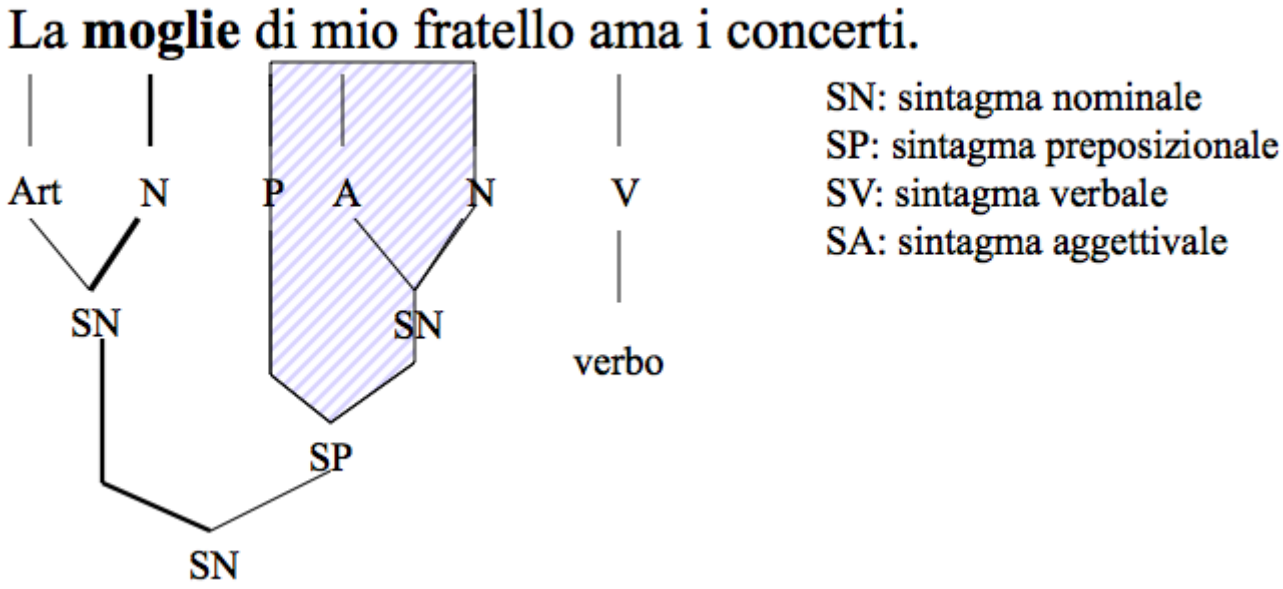
Per capire veramente qual è la struttura della frase, però, devo creare dei raggruppamenti tra parole.

Dalla stringa all'albero di parsing



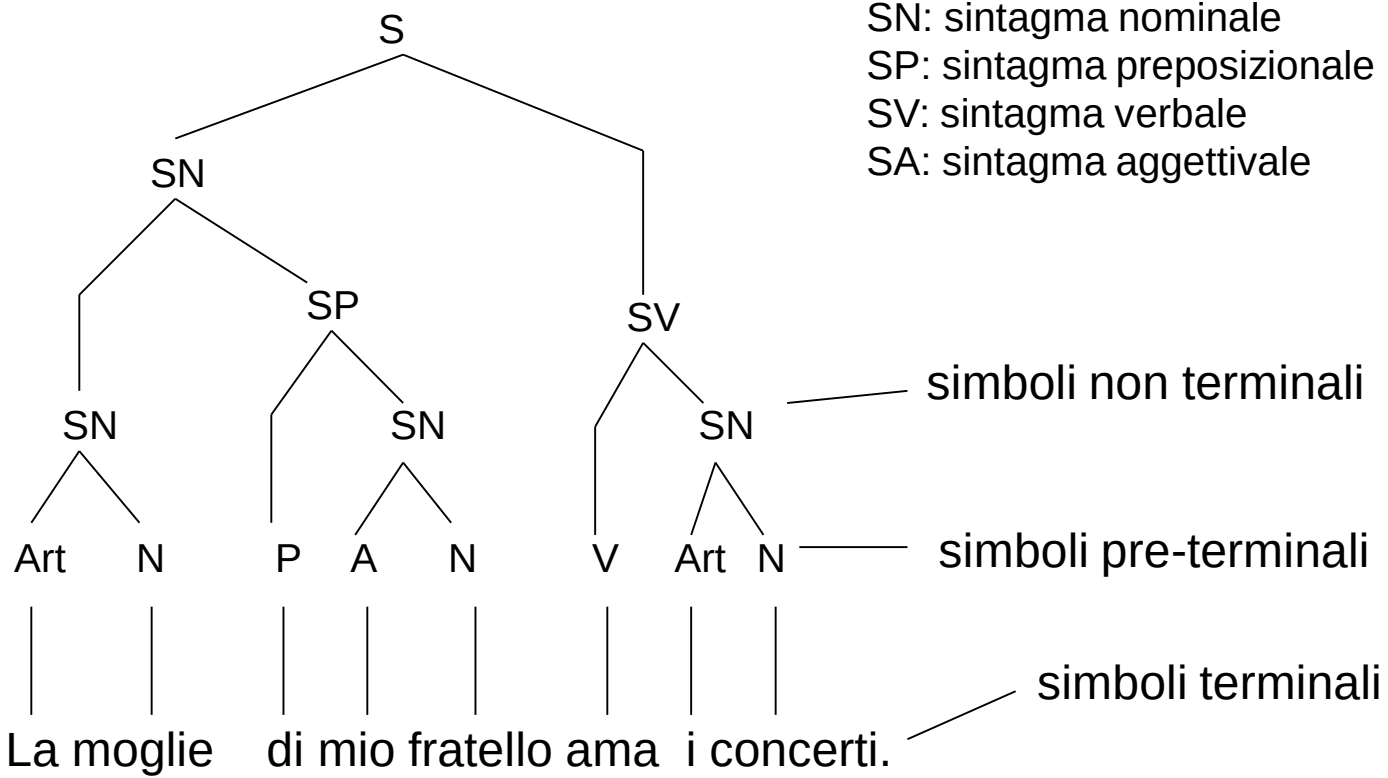
Come faccio a capire che questa sequenza di parole va insieme e rappresenta un **modificatore**?
Devo introdurre raggruppamenti di parole: i **sintagmi**

Dalla stringa all'albero di parsing

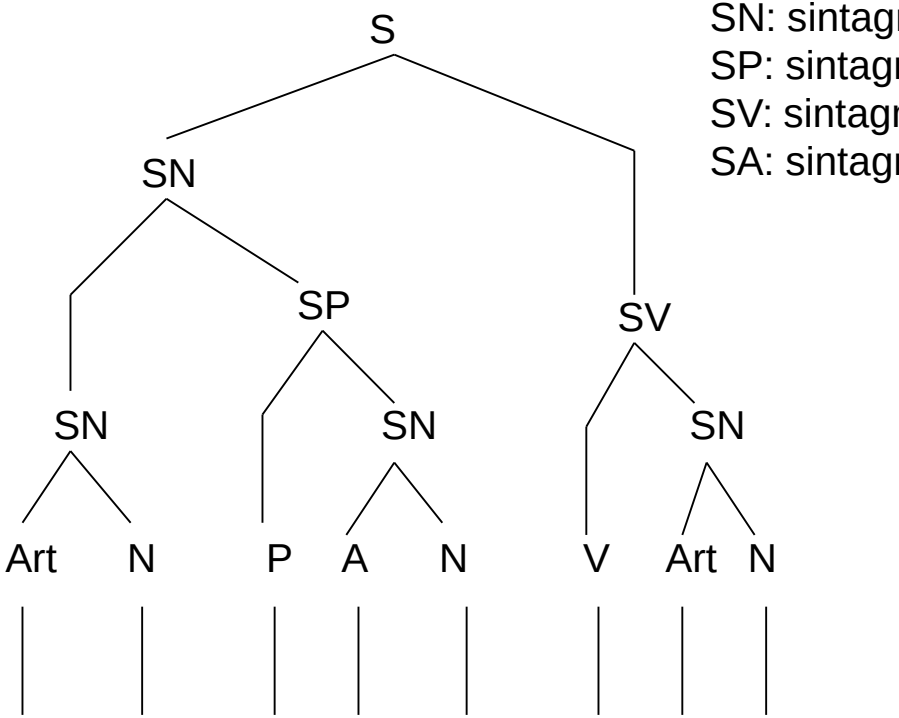


Alberi sintattici (1)

SN: sintagma nominale
SP: sintagma preposizionale
SV: sintagma verbale
SA: sintagma aggettivale



Relazioni tra simboli: regole di riscrittura



SN: sintagma nominale
SP: sintagma preposizionale
SV: sintagma verbale
SA: sintagma aggettivale

S → SN SV
SN → SN SP
SP → P SN
SV → V SN
SN → Art N
SN → A N
Art → la
Art → i
P → di
N → moglie
N → fratello
N → concerti
V → ama
A → mio

Regole di riscrittura (o produzioni)

S → SN SV
SN → SN SP
SP → P SN
SV → V SN
SN → Art N
SN → A N
Art → la
Art → i
P → di
N → moglie
N → fratello
N → concerti
V → ama
A → mio

S: sintagma
SN: sintagma nominale
SP: sintagma preposizionale
SV: sintagma verbale
SA: sintagma aggettivale

Simboli non terminali

Simboli terminali: sono parole e compaiono sulla destra della regola.

Una regola di riscrittura accoppia un simbolo non terminale (a sinistra) con simboli non terminali e/o terminali (a destra)

Grammatiche libere da contesto

- La modellazione matematica più comune utilizzata per modellare la struttura in sintagmi (o costituenti) si chiama **Grammatica libera da contesto** (Context-Free Grammar, **CFG**). Sono anche chiamate Grammatiche sintagmatiche o Phrase-structure grammar.

- Formalizzate per la prima volta da Chomsky nel 1956.

- Una CFG include una serie di regole o produzioni che esprimono il modo in cui i simboli nel linguaggio possono essere raggruppati e ordinati per es. il modo in cui un articolo e un nome vanno a comporre un sintagma nominale.

Grammatiche libere da contesto

- Definizione formale di una **grammatica libera da contesto**:
  - un insieme di **simboli terminali (Σ)**
  - un insieme di **simboli non terminali (Ψ)**
  - un insieme di **regole di riscrittura (P)**
  - un **simbolo iniziale (S)**

- L'applicazione delle regole di riscrittura (partendo dal simbolo iniziale S) consente di **generare** un insieme di frasi di una lingua (il **linguaggio generato**).

Grammatiche libere da contesto

$G = \Sigma = \{$ ama, moglie, la, i, concerti, la, pizza, mia $\}$

$\Psi = \{$ N, V, Art, SN, SV, S $\}$

$P = \{$

i. S → SN SV
ii. SN → Agg-poss N
iii. SN → Art N
iv. SV → V SN
v. N → pizza
vi. N → concerti
vii. N → moglie
viii. V → ama
ix. Art → i
x. Art → la
xi. Agg-poss → mia $\}$

Ci possono essere più regole con lo stesso simbolo nella parte sinistra.

Definiamo una grammatica ***G***

Grammatiche libere da contesto

Vogliamo dimostare che la grammatica *G* riesce a riconoscere la frase: *mia moglie ama i concerti*

passo	regola	Stringa prodotta
0		S
1	i.	**SN** SV
2	ii.	**Agg-poss** N SV
3	ix.	mia N SV
4	vii.	mia moglie **SV**
5	iv.	mia moglie **V** SN
6	viii.	mia moglie ama **SN**
7	iii.	mia moglie ama **Art** N
8	ix.	mia moglie ama i **N**
9	vi.	mia moglie ama i concerti

Grammatiche generative (3)

La grammatica *G* non riesce a generare la frase:

mia moglie ama i concerti in piazza

perché non ha a disposizione un simbolo per costruire un sintagma preposizionale (SP) per generare la porzione di frase *in piazza*.

La frase non fa dunque parte del linguaggio generato dalla grammatica *G*.

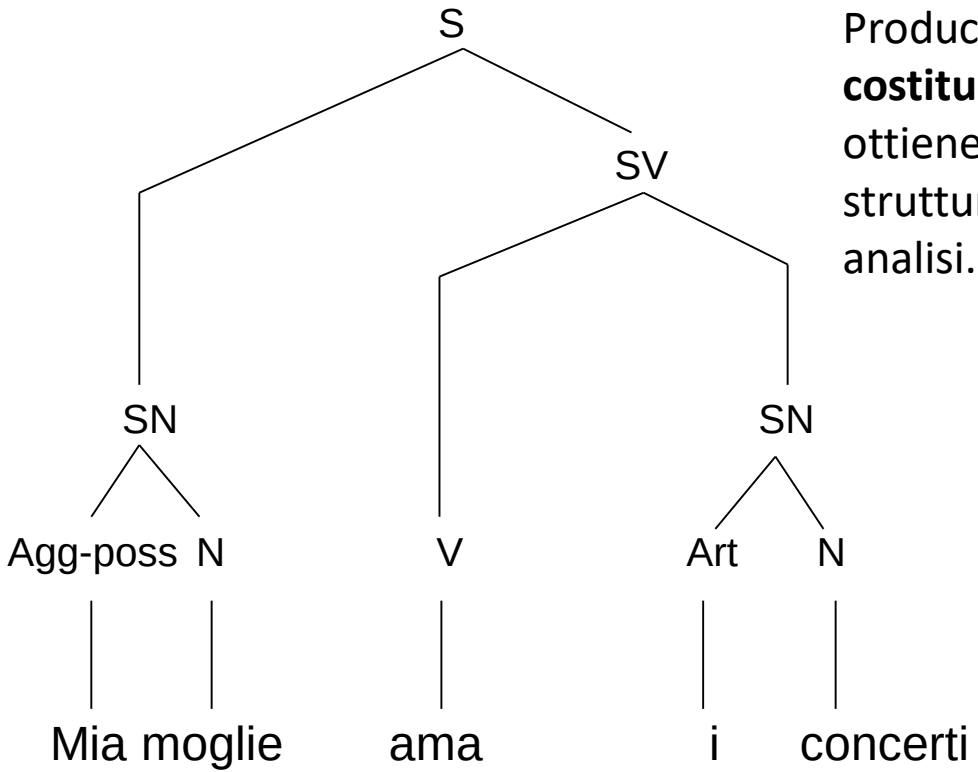
Teniamo traccia delle strutture intermedie

- Non è sufficiente sapere se, data una grammatica *G,* una certa frase appartiene o meno alla lingua descritta da quella grammatica.

- Quello che serve è una rappresentazione completa della struttura della frase.

- L'analisi sintattica della frase deve ritornare i singoli **costituenti** (o **sintagmi**) che sono stati utilizzati per riconoscere la frase.

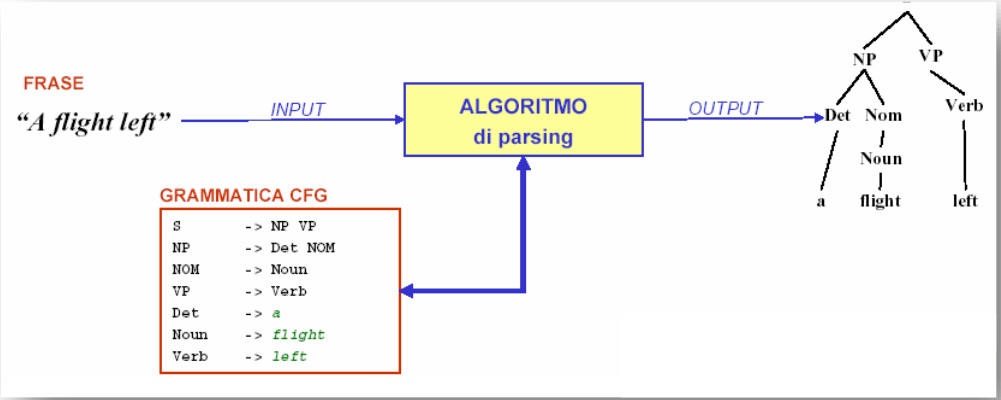
Il Parsing: Analisi sintattica di testo



Produce un albero di **costituenza** (**albero sintattico**) che si ottiene componendo le descrizioni strutturate ottenute ad ogni passo di analisi.

Parsing e Grammatiche

- Il Parsing e la Grammatica sono due cose diverse
- Una **Grammatica** è un modello dichiarativo che definisce un linguaggio
- Il **Parsing** è un processo di assegnazione di una struttura ad una stringa in base ad una grammatica secondo un **algoritmo**



Parsing e Grammatiche

- Il Parser è il programma che effettua l'analisi sintattica (il Parsing) e che si articola in un numero ristretto di funzioni:

  - **Scansione della frase** da analizzare (frase di input). Il procedimento porta ad analizzare una parola alla volta. In alcuni casi, riuscire a esaminare fino a una o due parole più avanti aiuta a ridurre l'ambiguità. In questo caso, il parser è dotato di una finestra o un *look-ahead* di *n* parole.

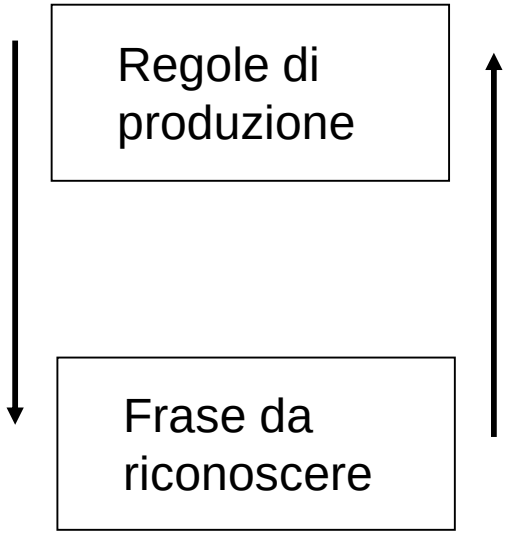
  - **Scansione della grammatica** e reperimento delle regole rilevanti. Le strategie possono essere *bottom-up* o *top-down*

  - **Strutturazione** della frase e memorizzazione progressiva della struttura.

Due tipologie di algoritmi di parsing

Algoritmi Top-down:
Si parte dalle regole e si cerca di applicarle a porzioni di frase.

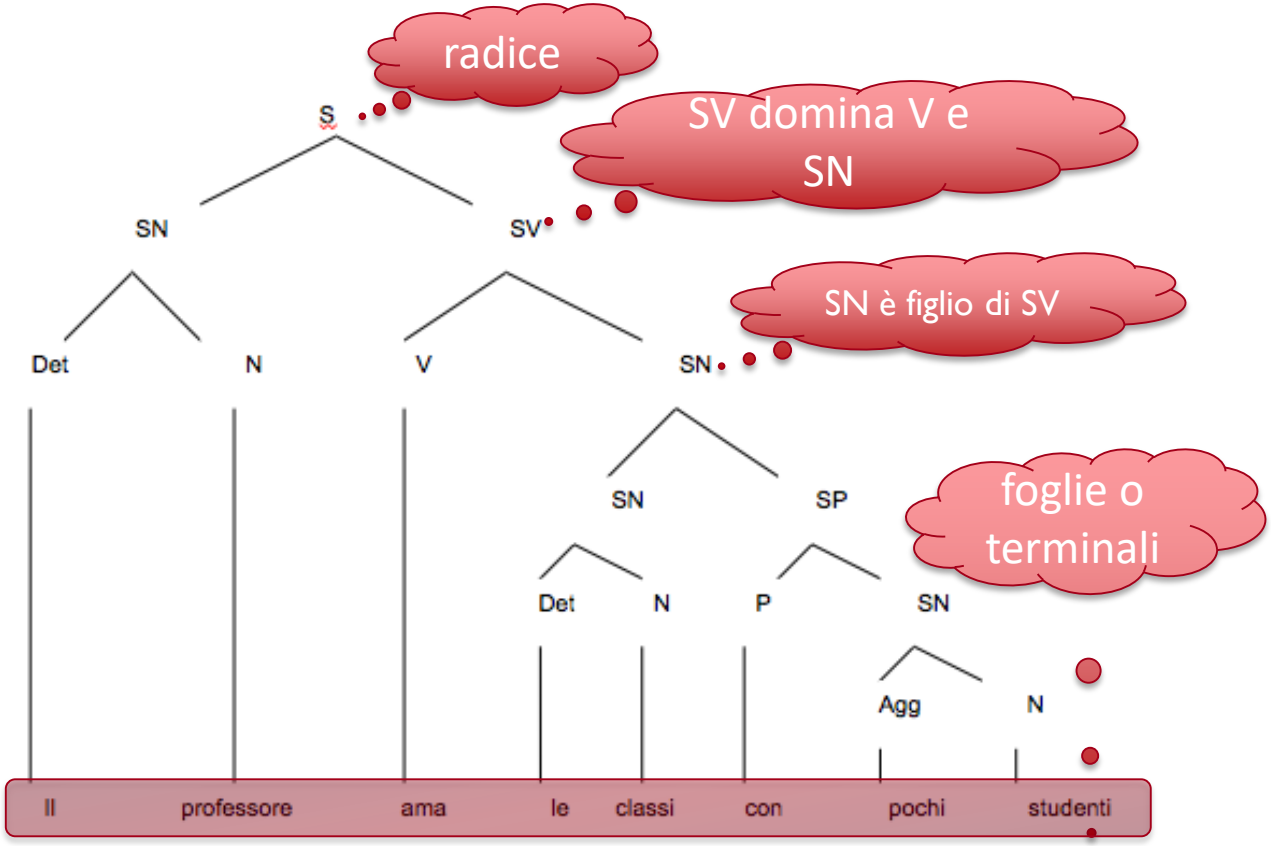
Problemi: se si parte con una ipotesi sbagliata, l'errore viene riconosciuto dopo molti tentativi.



Algoritmi Bottom-up:
Si parte dalle parole e si cerca di accorparle cercando regole appropriate.

Problemi: partendo dalle parole, l'assenza di una regola viene verificata dopo molti tentativi.

Gli alberi di parsing



I Treebank

- Abbiamo visto che è molto importante nello sviluppo di sistemi di trattamento automatico del linguaggio avere a disposizione dati annotati
- Questo permette sia di effettuare analisi quantitative su determinati fenomeni a partire da testi (es. statistiche sulle costruzioni sintattiche più frequenti, sui sintagmi), ma anche di **allenare sistemi supervisionati.**
- Un corpus in cui ogni frase è annotata sintatticamente tramite l'assegnazione di un albero di parsing si chiama **Treebank**

I Treebank

- Esistono Treebank in molte lingue. Vedi la lista al link http://nlp.stanford.edu/links/statnlp.html#Treebanks

- La procedura per creare un Treebank consiste di solito in una fase di annotazione automatica tramite *parsing* e poi di una correzione manuale

- Il Treebank più utilizzato per l'inglese è il **Penn Treebank**

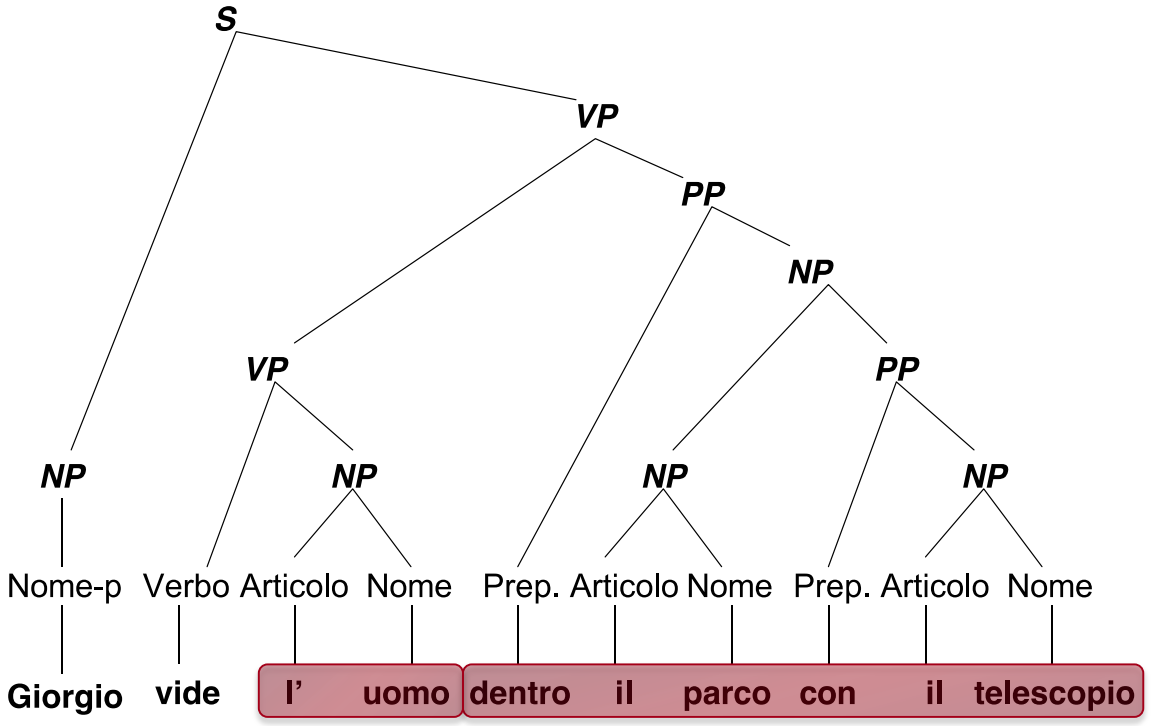
http://www.cis.upenn.edu/~treebank/

Il Problema dell'Ambiguità sintattica

- Spesso per una frase sono ammissibili diverse analisi sintattiche
- Una grammatica libera da contesto non riesce a gestire l'ambiguità perchè non include informazioni su quali forme siano più probabili di altre
- L'analisi corretta secondo una CFG è quella che viene risolta prima secondo l'ordine in cui sono definite le regole

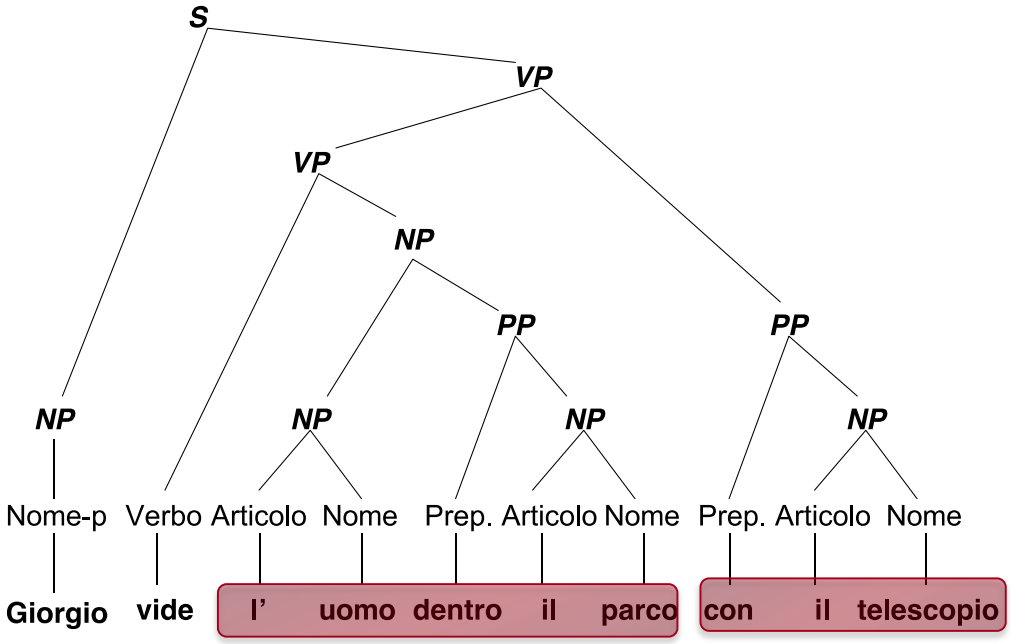
Il Problema dell'Ambiguità sintattica

"Giorgio vide l'uomo dentro il parco con il telescopio"



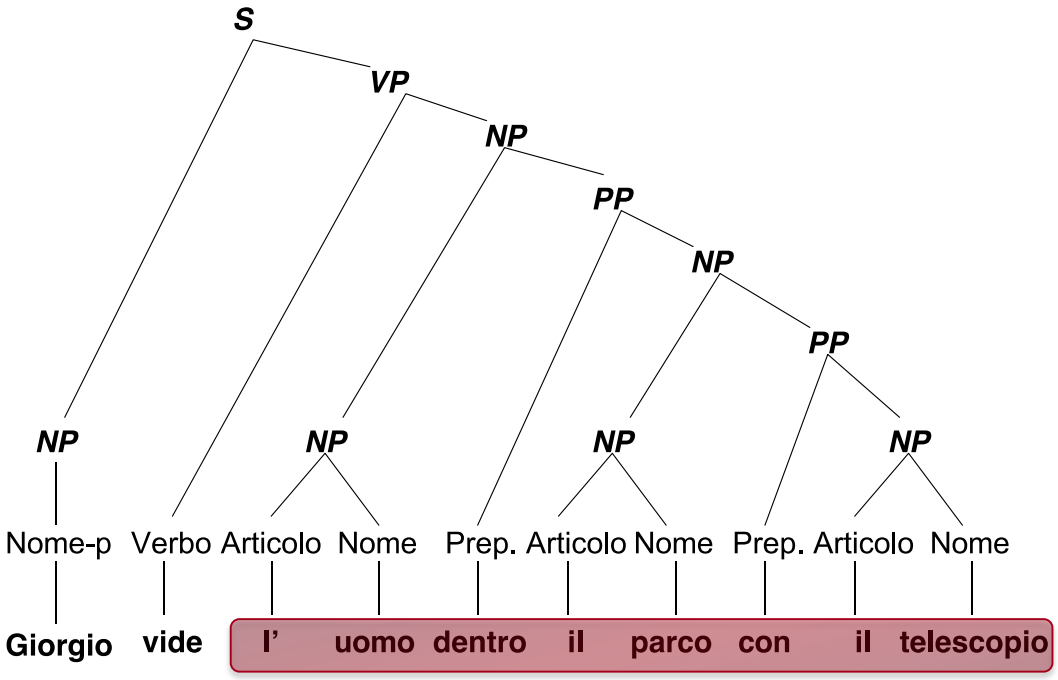
Il Problema dell'Ambiguità sintattica

"Giorgio vide l'uomo dentro il parco con il telescopio"

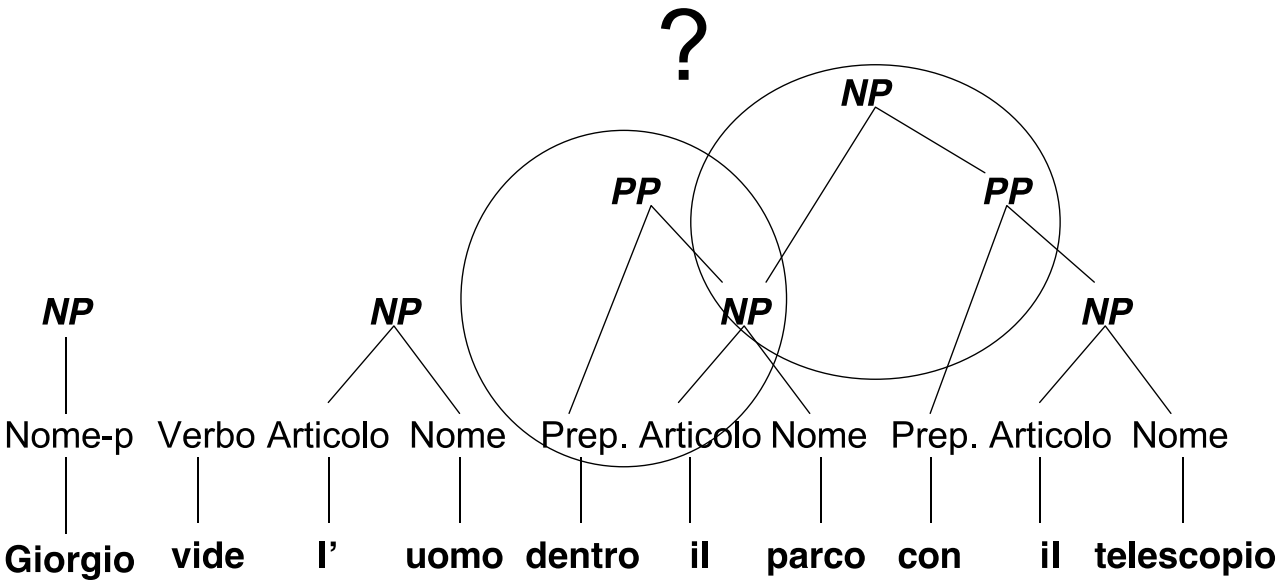


Il Problema dell'Ambiguità sintattica

"Giorgio vide l'uomo dentro il parco con il telescopio"



Il Problema dell'Ambiguità sintattica

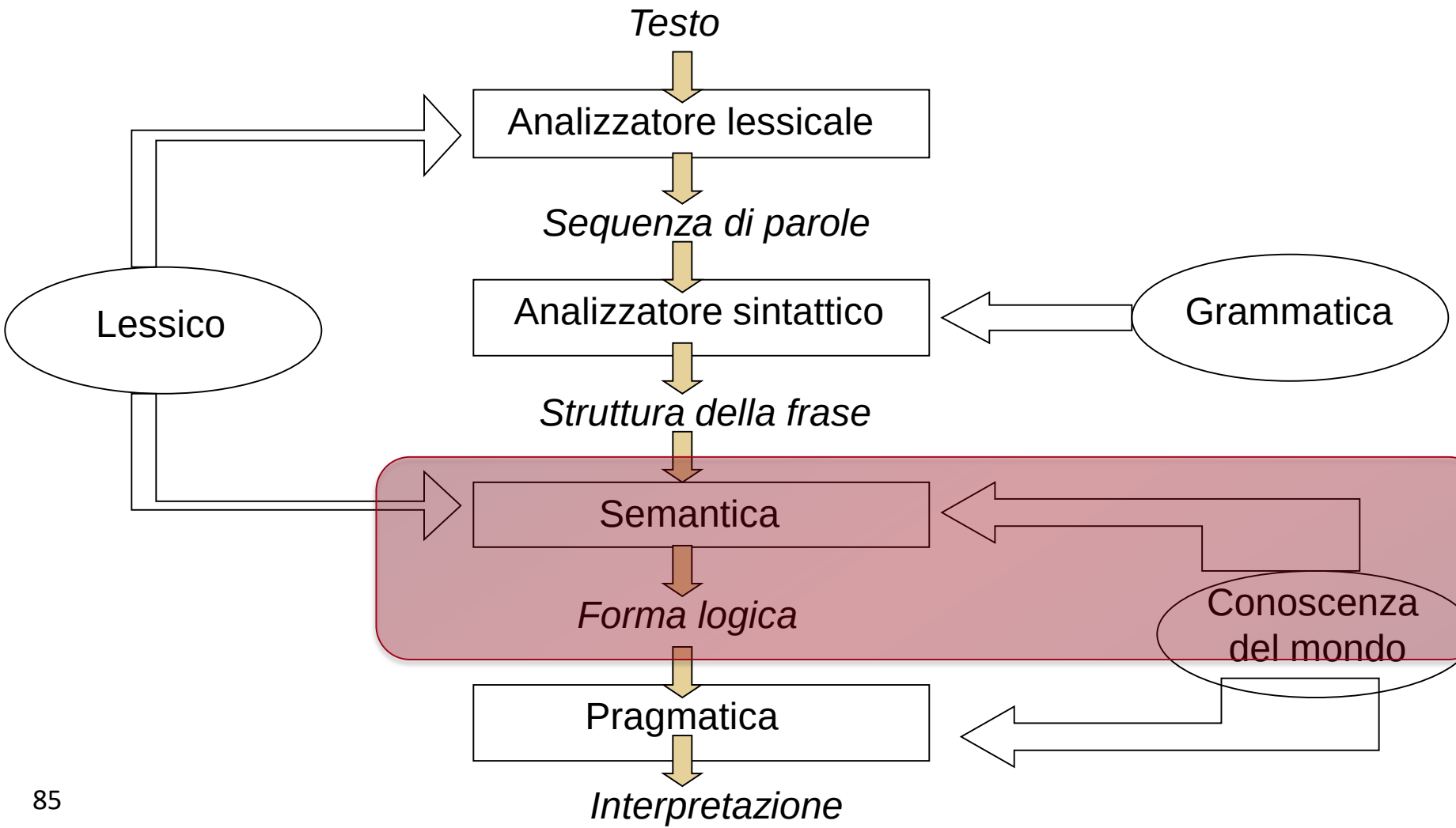


Il Problema dell'Ambiguità sintattica

- L'analisi sintattica di una frase può produrre più di un albero sintattico. Un caso tipico sono le ambiguità che derivano dall' **attaccamento dei sintagmi preposizionali** (PP-attachment)

- L'ambiguità nasce da come è formulata la grammatica.

- Problema molto studiato perchè difficile da trattare automaticamente


Testo
Analizzatore lessicale
Sequenza di parole
Analizzatore sintattico
Grammatica
Lessico
Struttura della frase
Semantica
Forma logica
Conoscenza del mondo
Pragmatica
Interpretazione

Semantica lessicale

- **Semantica :** studio del significato
- **Lessico :** insieme delle parole di una lingua

==> Semantica lessicale**:** studio del significato delle parole di una lingua

A cosa serve la semantica in CL ?

- Per la traduzione automatica
  - *Mangio una pesca/Vado a pesca -> peach/to fish*
  - *Paolo e' caduto per terra/Paolo e' caduto in trappola-> fell down/got trapped*
- Per la ricerca di documenti
  - Per evitare il silenzio es. : *vendere/cedere/dare*
  - Per evitare il ``rumore'' es.: *pesca*
- Per i sistemi di risposta automatica
  - *Paolo ha regalato delle rose a Maria*
    - Paolo ha regalato dei fiori a Maria?
    - Maria ha delle rose ?

A cosa serve la semantica in CL ?

- Per la construzione di dizionari

  apparecchi di misurazione/sistemi di misurazione : varianti

- Per la generazione di testi
  - Scegliere le collocazioni corrette :

  - tendere un tranello *- prendere fiato*

  - realizzare un desiderio *- mantenere una promessa*

- Per la disambiguazione e la comprensione del testo

  Il vagone 1 si trova in testa al treno

  Si è procurato una ferita alla testa

Relazioni semantiche tra unità lessicali

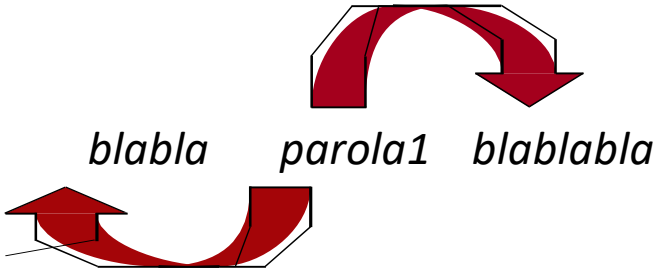
- **Relazioni paradigmatiche** (sostituzione)

blabla *parola1* *blablabla*

parola2

sinonimia, iperonimia, meronimia, antonimia

- **Relazioni sintagmatiche** (concatenazione di unità lessicali nel testo)



argomenti, collocazioni

Disambiguazione del significato delle parole

- La word sense disambiguation (WSD) riguarda il problema di individuare quale senso di una parola è usato in una data frase
- Si applica alle parole con più significati (polisemia)
- Richiede un dizionario che elenca i possibili sensi di ogni parola
- Si può affrontare su singole parole o congiuntamente su tutte le parole della frase (si considerano le combinazioni di significati)

  Ho mangiato un piatto freddo

  Ho lavato un piatto sporco

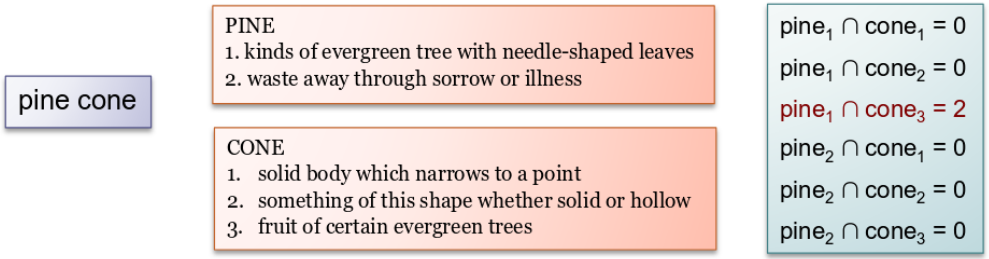
  Hanno servito un piatto freddo

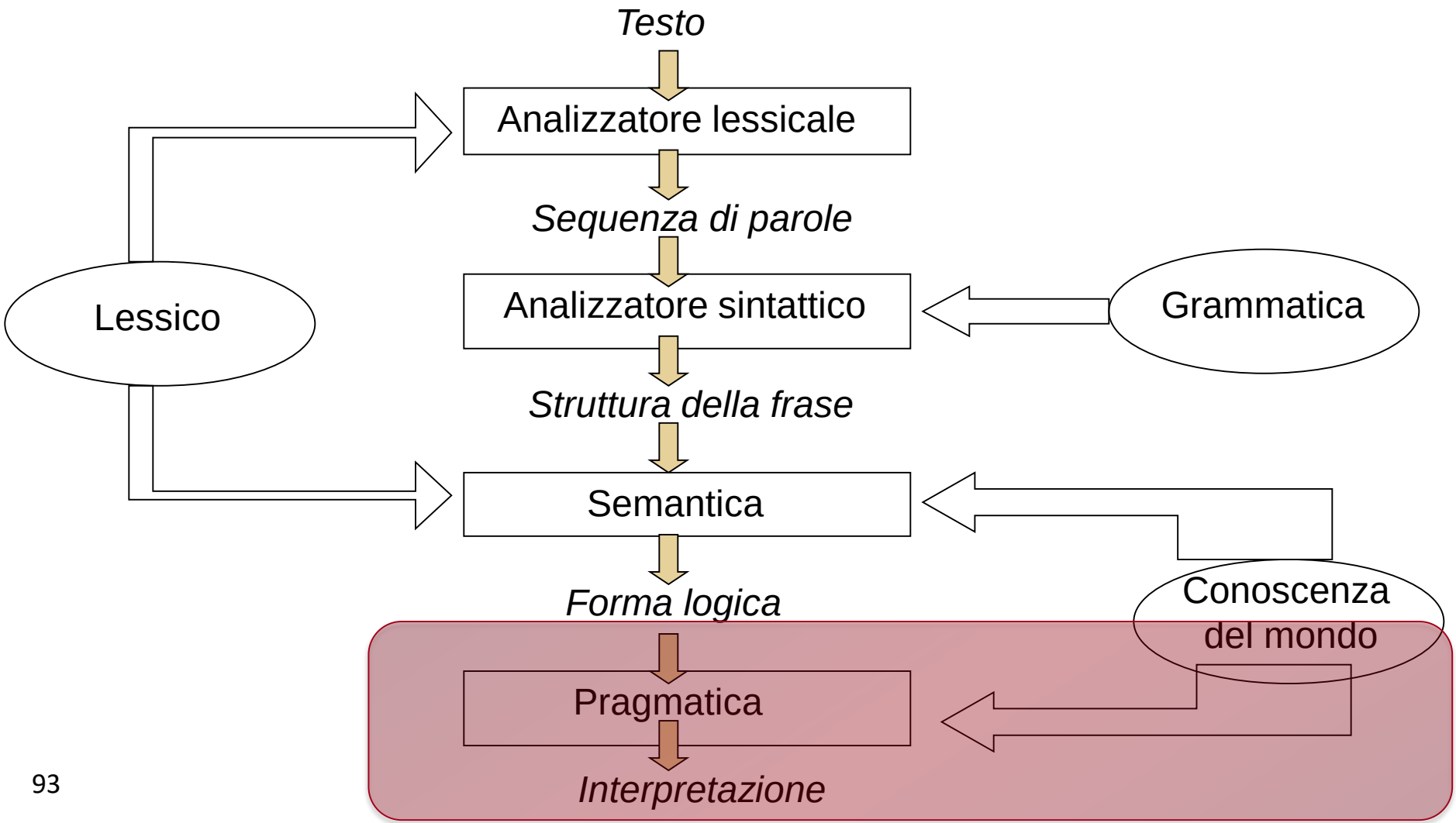
Apprendimento supervisionato

- La WSD può essere affrontata come un problema di classificazione: il senso corretto è la classe da predire
- la parola è rappresentata con un insieme (vettore) di feature in ingresso al classificatore
- Il classificatore può essere stimato con tecniche di apprendimento automatico a partire da un dataset etichettato
- Si possono usare diversi modelli per costruire il classificatore (Naïve Bayes, reti neurali, alberi di decisione…)

Dictionary-based methods

- Un dizionario può fornire informazioni sul contesto legato ai sensi delle parole (le glosse)
- L'algoritmo più semplice è quello di Lesk (1986) :
  - Si calcola la sovrapposizione fra le glosse associate ai vari significati delle parole nella frase
  - Si sceglie la combinazione di significati che fornisce il massimo livello di sovrapposizione complessiva (complessità è combinatoria nel numero di sensi)




Testo
Analizzatore lessicale
Sequenza di parole
Analizzatore sintattico
Lessico
Grammatica
Struttura della frase
Semantica
Forma logica
Conoscenza del mondo
Pragmatica
Interpretazione

Pragmatica

- Più difficile ancora si presenta il livello pragmatico.
- Per dialogare correttamente è necessario rappresentarsi le intenzioni degli interlocutori, che sono solo parzialmente rispecchiate nelle loro parole
- A tutt'oggi, i tentativi di inserire un livello di analisi pragmatica nei sistemi di elaborazione automatica del linguaggio non sono numerosi, e spesso hanno un valore puramente esplorativo.
- Alcuni di questi tentativi si basano su un'interessante teoria sviluppata dai filosofi del linguaggio: la teoria degli atti linguistici (speech acts)

Ricerca @UCA:
Estrazione di strutture argumentative da testi

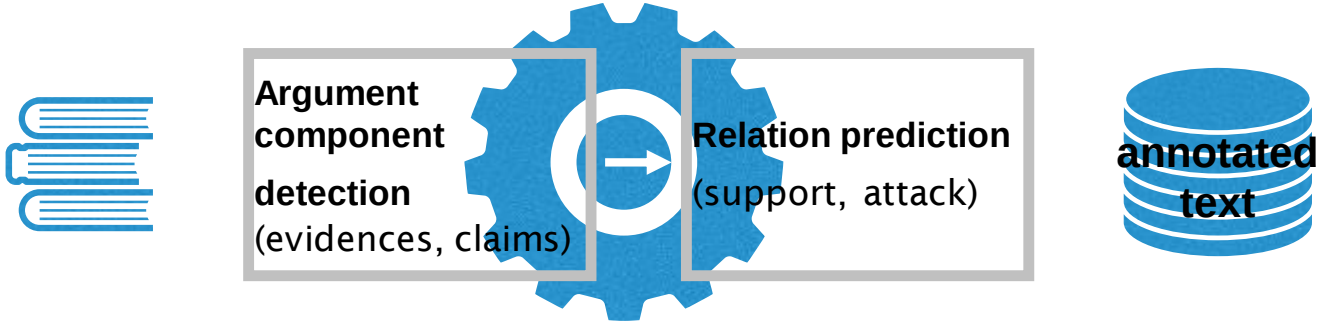
New challenges for AI

- **Huge amount of documents** available on the Web;
- **Machine learning** and natural language processing methods;
- **Heterogeneous data** (e.g., clinical reports, legal documents, user-generated content, political debates);
- Additional elements: **domain knowledge**, **sentiment**, **emotions**, …

Argument Mining: understanding the WHY

Opinion mining (or sentiment analysis): **what** users think about a certain product, event, political party, …

Argument mining: **why** users have this opinion about a certain product, event, political party, …



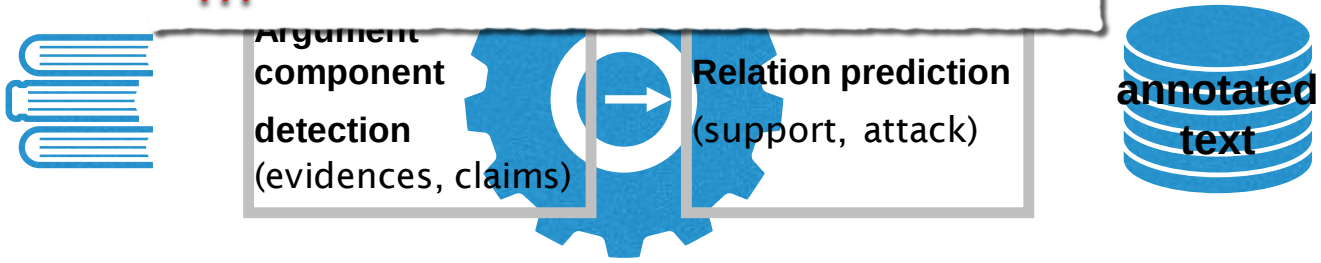
Argument Mining: understanding the WHY

Opinion minin
think about a

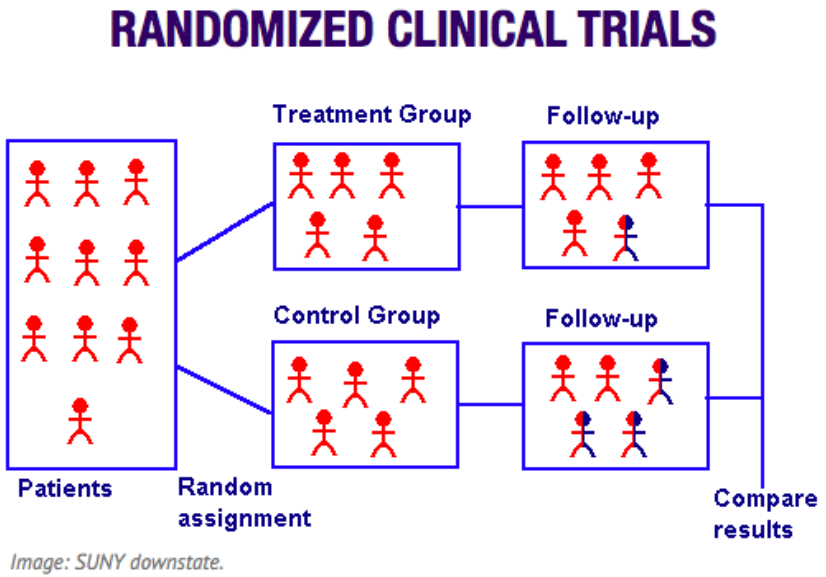
Argument mir
certain produc

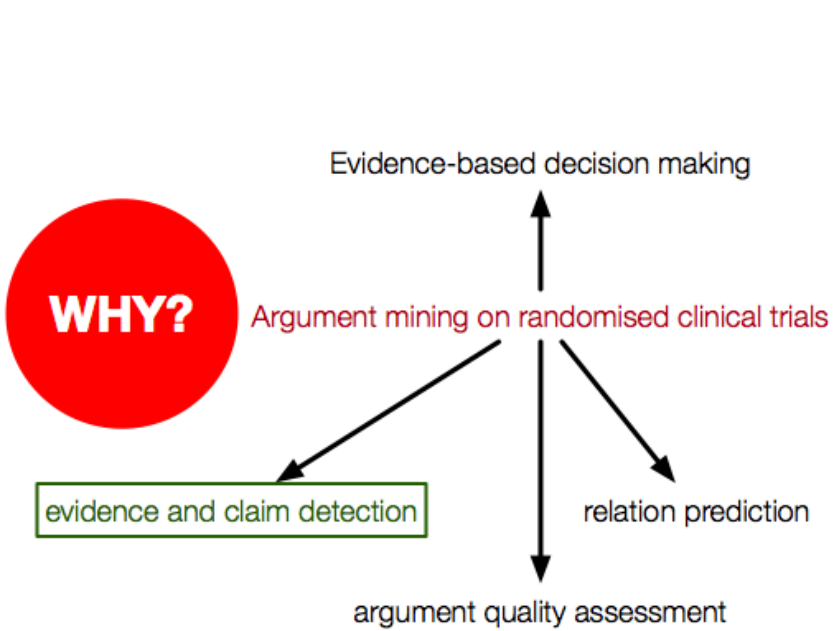
Applications:

- Structured summaries (main claims highlighted linked by relations)
- Inconsistencies detection
- Explainable AI
- …



Evidence-based decision making: clinical trials

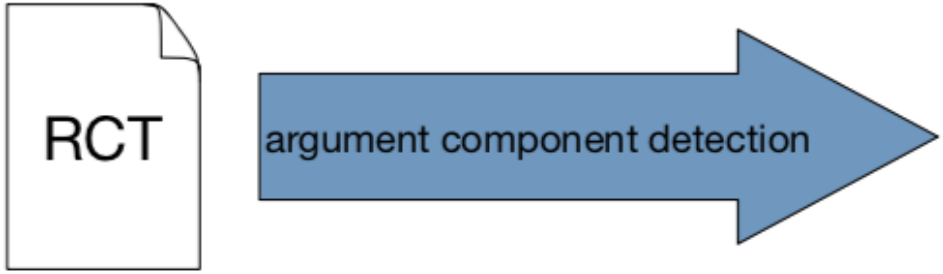




An example on glaucoma

To compare the intraocular pressure-lowering effect of latanoprost with that of dorzolamide when added to timolol. [...] [*The diurnal intraocular pressure reduction was significant in both groups ($P < 0.001$)*]$_1$. [*The mean intraocular pressure reduction from baseline was 32% for the latanoprost plus timolol group and 20% for the dorzolamide plus timolol group*]$_2$. [*The least square estimate of the mean diurnal intraocular pressure reduction after 3 months was -7.06 mm Hg in the latanoprost plus timolol group and -4.44 mm Hg in the dorzolamide plus timolol group ($P < 0.001$)*]$_3$. Drugs administered in both treatment groups were well tolerated. This study clearly showed that **[the additive diurnal intraocular pressure-lowering effect of latanoprost is superior to that of dorzolamide in patients treated with timolol]$_1$.**[3]

Our result (I)



Brimonidine-treated subjects showed an overall mean peak reduction in intraocular pressure (IOP) of 6.5 mm Hg; timolol-treated subjects had a mean peak reduction in IOP of 6.1 mm Hg.

Brimonidine lowered mean peak IOP significantly more than timolol at week 2 and month 3 ($P < .03$);

Allergy was seen in 9% of subjects treated with brimonidine.

No evidence of tachyphylaxis was seen in either group.

Dry mouth was more common in the brimonidine-treated group than in the timolol-treated group (33.0% vs 19.4%),

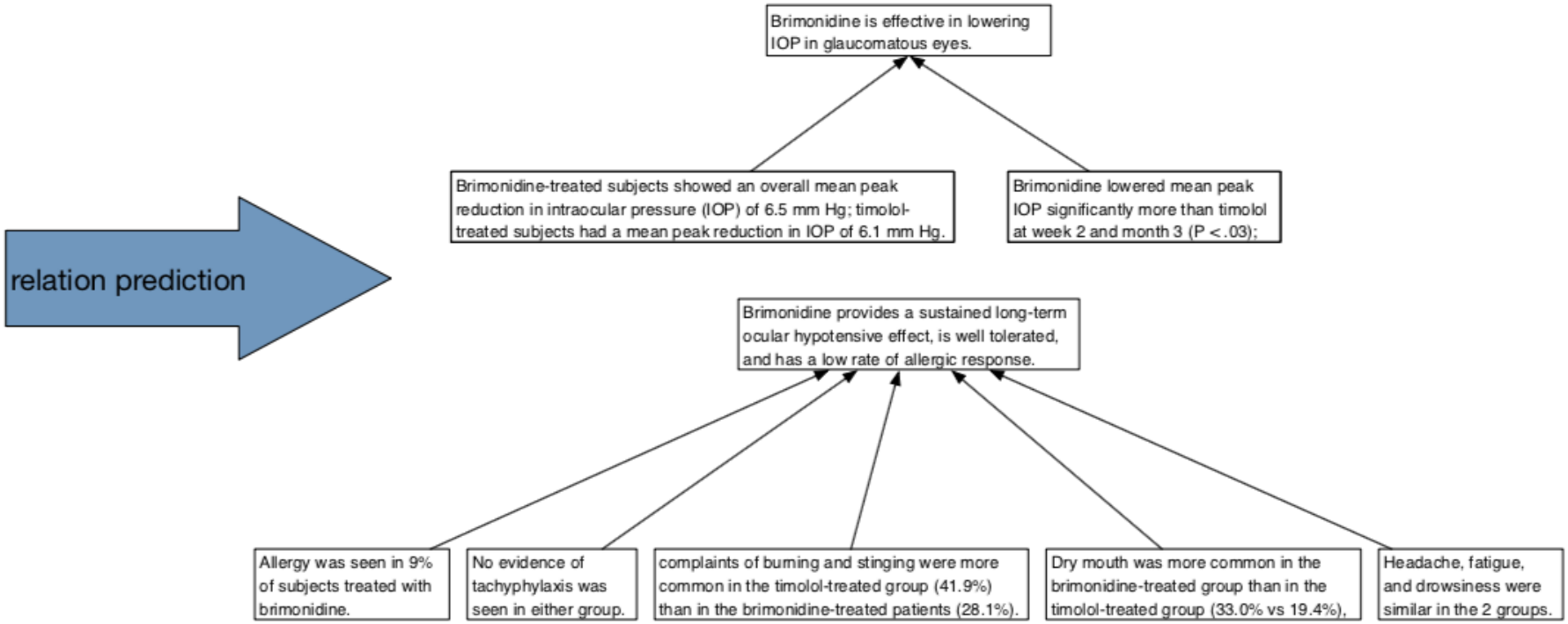
Headache, fatigue, and drowsiness were similar in the 2 groups.

complaints of burning and stinging were more common in the timolol-treated group (41.9%) than in the brimonidine-treated patients (28.1%).

Brimonidine is effective in lowering IOP in glaucomatous eyes.

Brimonidine provides a sustained long-term ocular hypotensive effect, is well tolerated, and has a low rate of allergic response.

Our result (II)



Cyberbullying phenomena detection

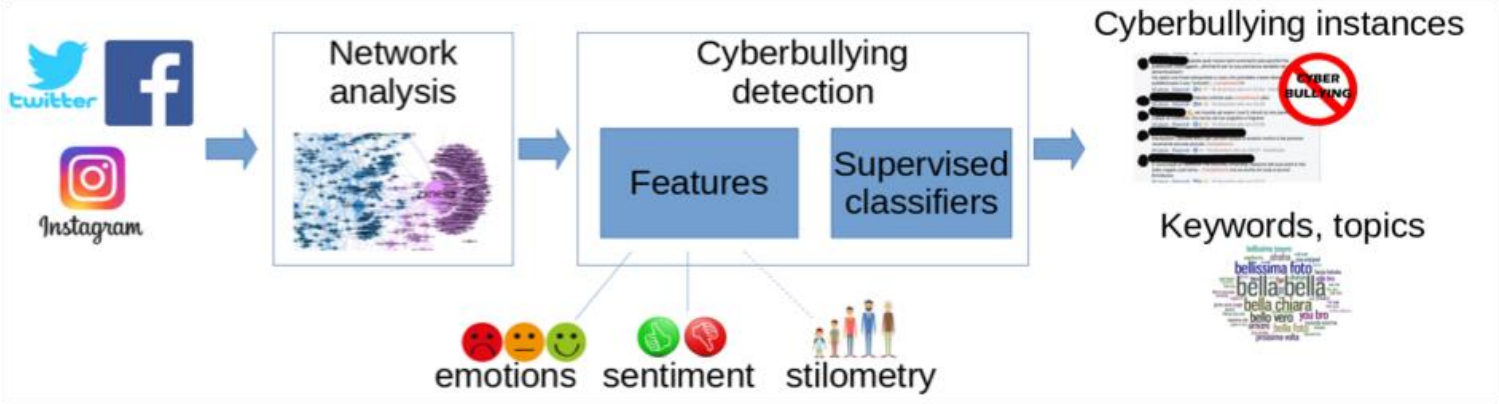
Cyberbullying is a form of bullying or harassment using electronic forms of contact. Around 70% of those who bully/have been bullied offline state they have also bullied/have been bullied online. In 2016, one million teenagers were harassed, threatened or subjected to other forms of cyberbullying only on Facebook.

There is increasing evidence that cyberbullying can influence depression, mental health, substance use, or suicide-related behaviours, mainly among young people. Virtual coaches integrating chatbots tailored to teens' needs can play a major role in the market of preventative and personalized health.

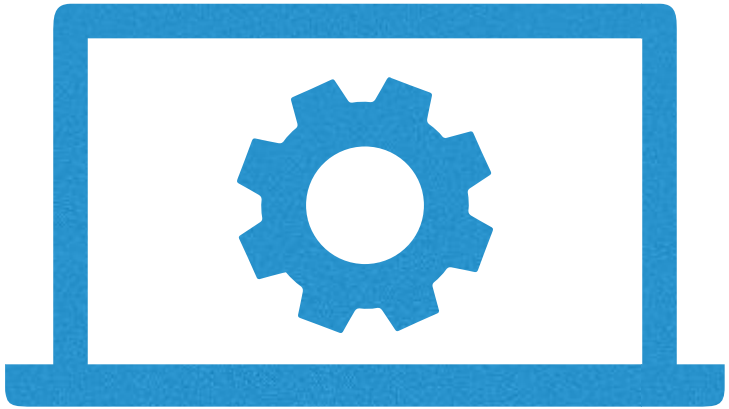
CREEP semantic technology

CREEP will monitor the wellbeing of young victims of cyberbullying, in full respect of privacy and data protection, **detecting risky situations emerging from web and social media**.

CREEP semantic technology offers a text mining, argumentation and sentiment analysis solution to detect cyberbullying activities (**topic-based social media monitoring**).



CREEP semantic technology DEMO

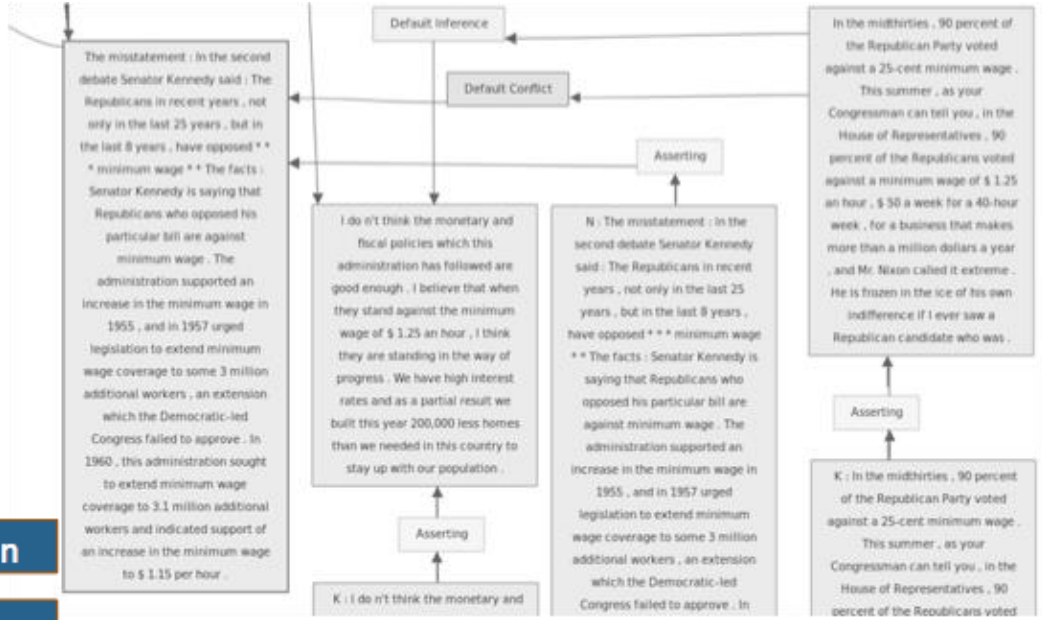


Ongoing work: Connection between arguments and emotions in online debates

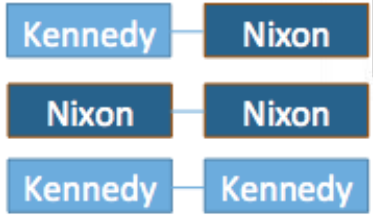
- Emotion detection (Heron Lab, University of Montreal)
  - webcams for facial expressions analysis [FACEREADER 6.0]
  - physiological sensors (EEG) for cognitive states
- Real-time engagement
  - engagement index [Pope et al.,1995]
  - EEG frequency bands
- Real-time facial analysis
  - classifying 500 key points in facial muscles
  - neural network
    - happy, sad, angry, surprised, scared, disgusted.
    - valence, arousal
    - neutral probability.
- Argumentation and Persuasion

Ongoing work: Never Retreat, Never Retract: Argumentation Analysis of Political Speeches

1960 Presidential Campaign
http://www.presidency.ucsb.edu/1960_election.php



5 Topics:
- Cuba
- Disarmament
- Health care
- Minimum Wage
- Unemployment



Relations: support / attack / neutral

Credits:

Parte del materiale di questa lezione proviene da:

- Isabella Chiari "*Introduzione alla Linguistica Computazionale*", Laterza, Roma-Bari, 2007.
- Sara Tonelli, slides introduttive al corso di CL.
- Collaborazione con Serena Villata