

ontologico vs logico
l'oggetto + intuire
le relaz. di appartenenza

Tecnologie del Linguaggio Naturale

Parte Prima

Lezione n. 02

I livelli linguistici

Ama (PIF) → rappresentazione
del significato

↑
architettura a
casse

2 Marzo 2021



Dalla superficie alla profondità

Il problema ...

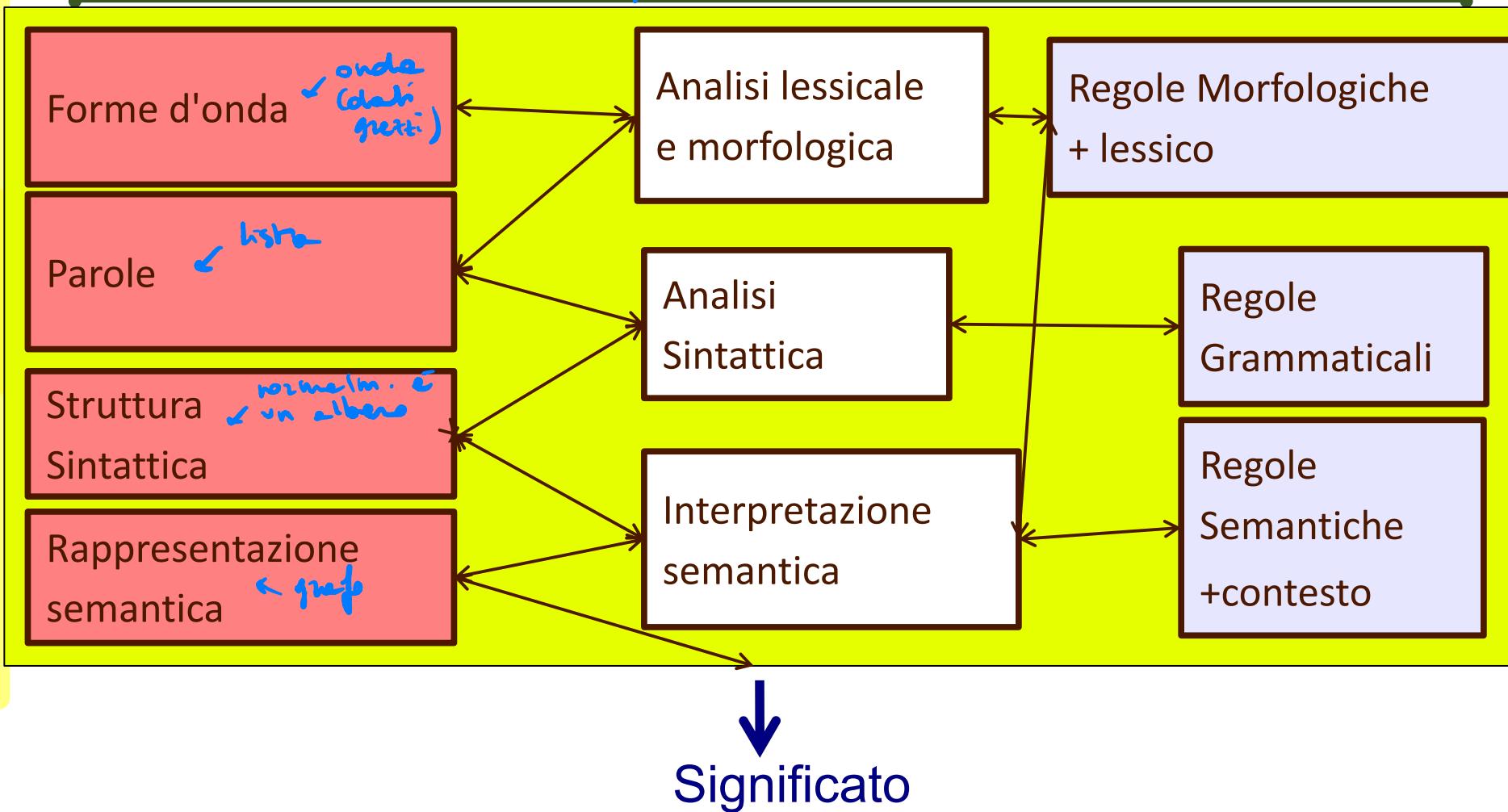
Convertire una frase o un testo in una forma che
permetta l'applicazione di meccanismi di
ragionamento automatico

strutture dati

Frase

livello linguistico

basi di conoscenza



... ma la soluzione non è così
semplice

- Come funzionano questi moduli?
- Come i moduli comunicano con le basi di conoscenza?

Prologo

- La grande guerra tra regole e statistica
 - Rules-driven
 - Data-driven
- ← quali sono? dipende dai domini!*

On Becoming a Discipline, M. Steedman

- <http://www.aclweb.org/anthology/J08-1008>
- Teoria quantistica vs. Relatività Generale

Per capire la guerra, una semplice domanda:
quando finisce una frase?

t sentence splitting

Sentence splitting (segmentation)

- “!” , “?” -> OK
- “.” ->
 - fine frase
 - abbreviazione: Doc., Mr.
 - numeri: 0.2% or 4.5

Sentence splitting (segmentation)

Come costruire un classificatore binario che decida EoS (End of Sentence) o Not EoS?

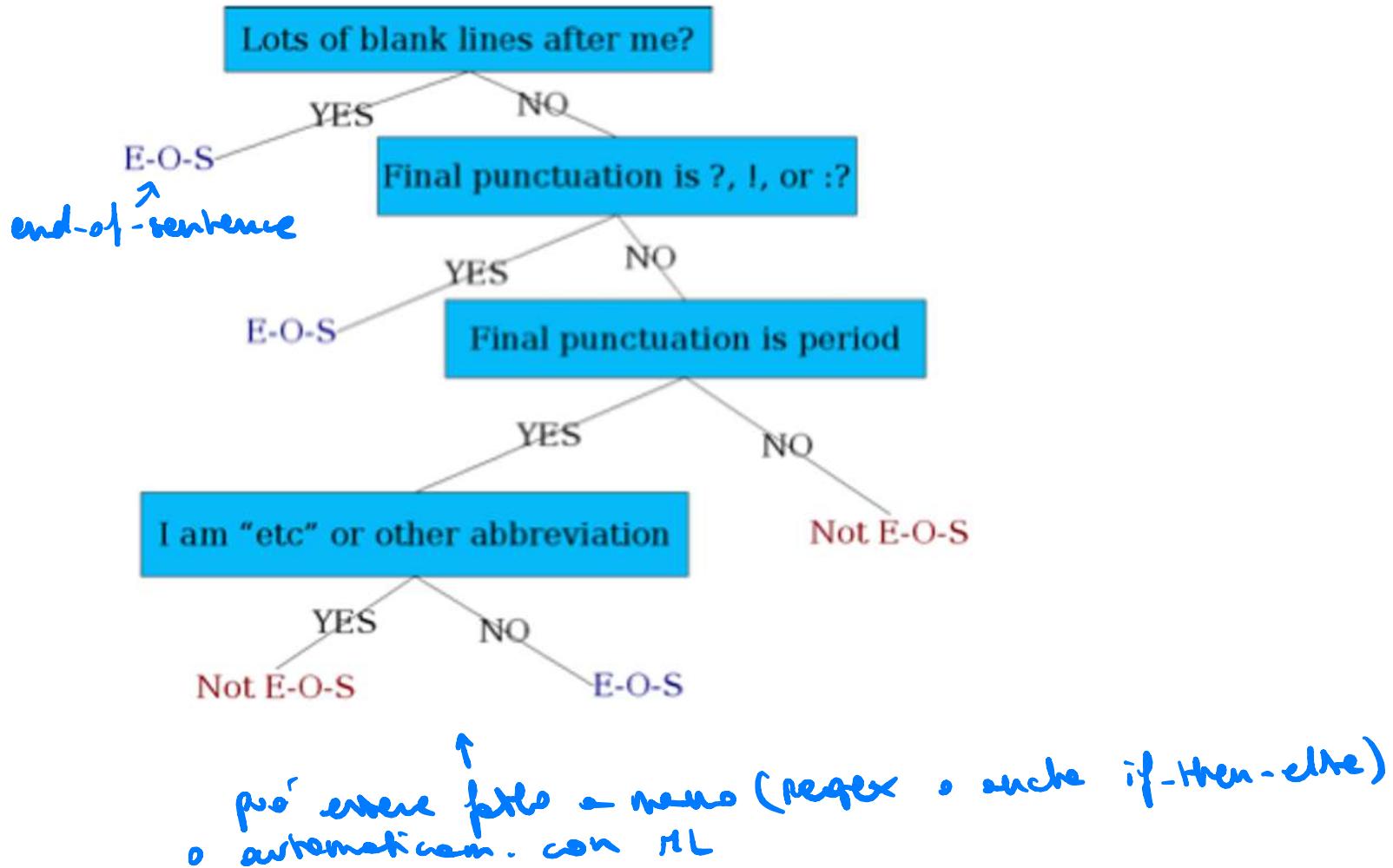
○ Regole scritte a mano

- regular-expressions
- tokenizer (FA) + rules
automata finito

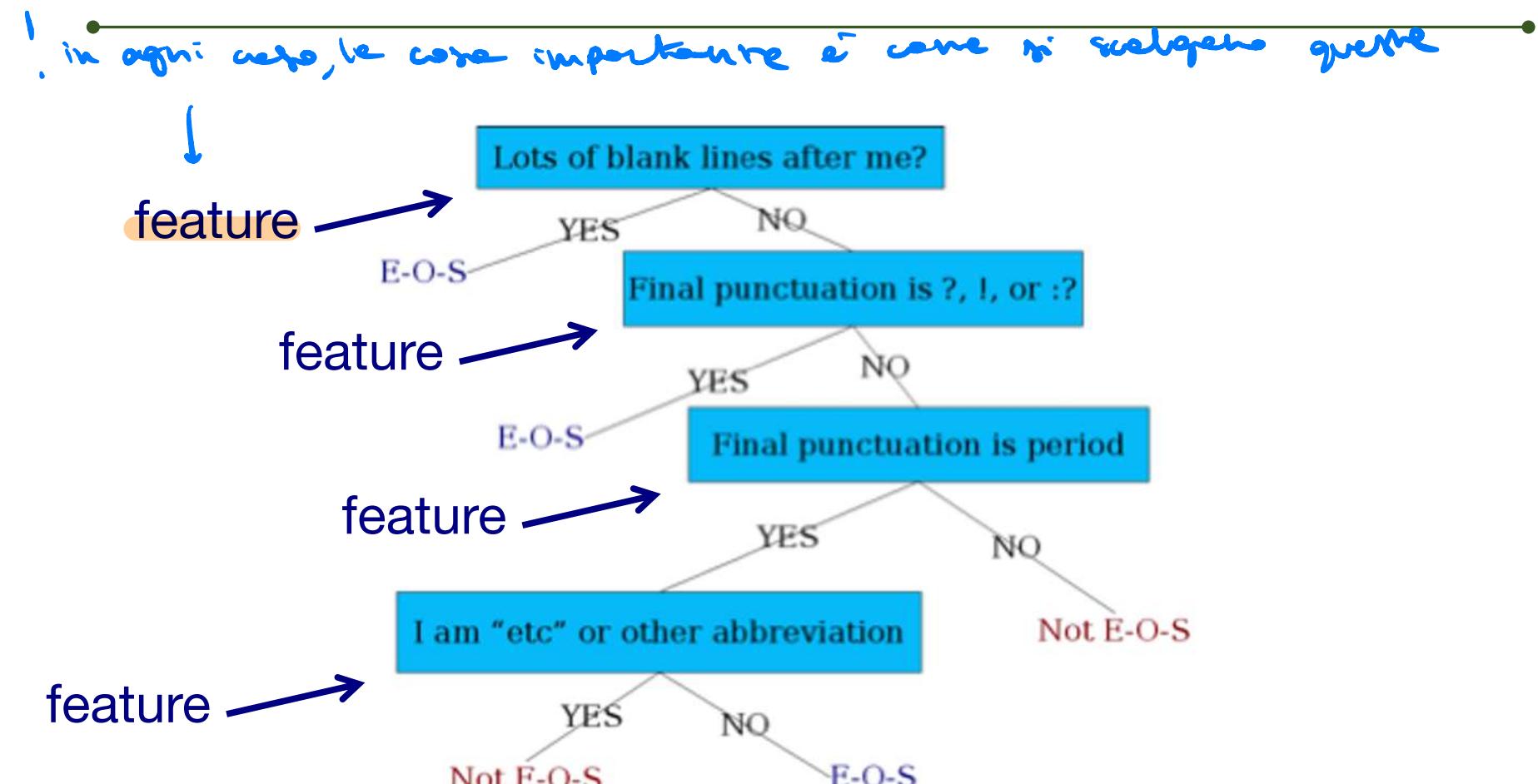
○ Machine learning ← classificatore binario statistico



Esempio: albero di decisione



Esempio: albero di decisione



Feature più complesse

- Case of word with “.”: Upper, Lower, Cap, Number
- Case of word after “.”: Upper, Lower, Cap, Number
- Features numeriche
 - Length of word with “.”
 - Probability that the word with “.” occurs at end-of-s
 - Probability that the word after “.” occurs at beginning-of-s

Cosa è davvero un albero di decisione?

- Una serie di IF-THEN-ELSE encapsulati
- Ho due possibilità per strutturare l'albero:
 - by-hand
 - solo in contesti semplici
 - valori numerici?
 - machine learning su un training corpus (es. C4.5)
- Il punto cruciale comunque, in entrambi i casi, è nella scelta delle features

! comunque è necessario del lavoro: qualcuno deve scegliere il corpus

Modelli del ML

- Decision trees
- Logistic regression
- SVM
- Perceptron
- Neural Nets
- etc.

"ma le reti neurali non hanno bisogno di features". Né anche le reti end2end hanno comunque bisogno d'strutturare il corpus in input.
Autoapprendono le features in base a come viene costruito l'input.
In qualche modo si è quindi comunque fatto un lavoro iniziale di selezione pensando a questo
il problema "si sposta + o rimane"
in sistema di traduz. automatica

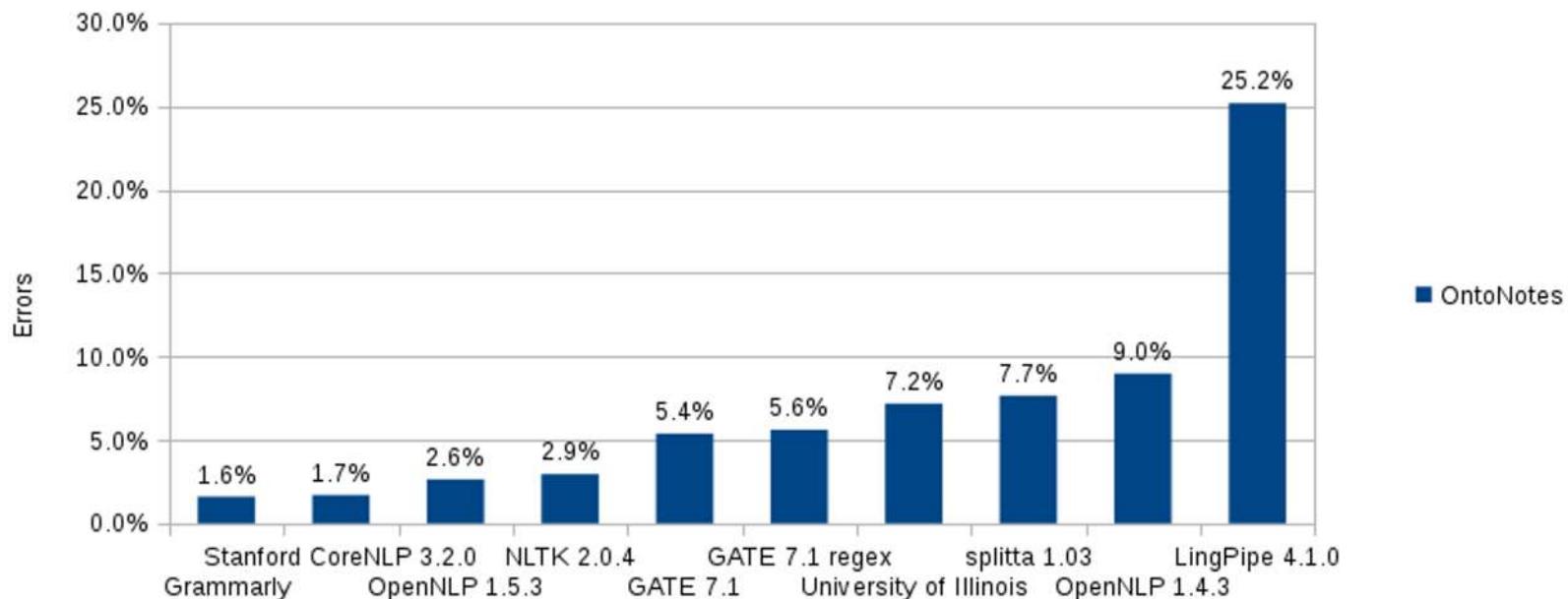
Il punto cruciale è sempre nella scelta delle features

Features linguistiche

- In questo corso ci concentreremo sullo studio delle feature linguistiche
- In alcuni casi l'approccio by-hand verrà privilegiato
 - poiché è didatticamente più chiaro/semplice e poiché è più semplice verificarne la fondatezza cognitiva mediante introspezione
- Le feature linguistiche possono sempre essere usate in sistemi di ML

State of art for sentence splitter

<http://tech.grammarly.com/blog/posts/How-to-Split-Sentences.html>



nel frattempo un'altra guerra ...

- Computational Linguistics and Deep Learning, Christopher D. Manning
 - http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00239
 - “The Deep Learning Tsunami”
→ veritabilmente × le computer visibili
→ abbastanza vero × NLP, ci sono anche
ottimi sistemi rule-based
 - Neural networks, Vectors and automatic feature inductions
 - Sequences vs. Trees and Graphs
- “It would be good to return some emphasis within NLP to cognitive and scientific investigation of language rather than almost exclusively using an engineering model of research.”
probabilmente
è vero e è +
completo, le
strutture sono
+ profonde
+ sofisticate

Outline

1 • Il livello morfologico e l'analisi lessicale

- Parole di contenuto e di funzione
- L'analizzatore morfologico e il PoS tagging

2 • Il livello sintattico

- Sintassi e grammatiche
- I costituenti e le CFGs
- Le grammatiche a dipendenze

3 • Il livello semantico

- Semantica lessicale
- Semantica guidata dalla sintassi

1

Il livello morfologico e l'analisi lessicale

- Il lessico è fondato sul concetto di parola
- Che cos'è una parola? Intuitivamente è una sequenza di caratteri delimitata da spazi o punteggiatura

me ..
↓

Il livello morfologico e l'analisi lessicale

- Sequenze di più parole. Es. passammela = passa a me essa
- Le parole hanno un significato unitario (semantica lessicale), ma volte sequenze di parole hanno un significato unitario. Es. di corsa, by the way
↳ "police rende"
- In altre lingue il problema è più grave
 - In tedesco: *Lebenversicherungsgesellschaftangestellter* = *impiegato di una società di assicurazione sulla vita*
 - In inglese: *Wouldn't?* = *Would not*

Ricerca sul dizionario: l'analizzatore morfologico

Presenza di suffissi

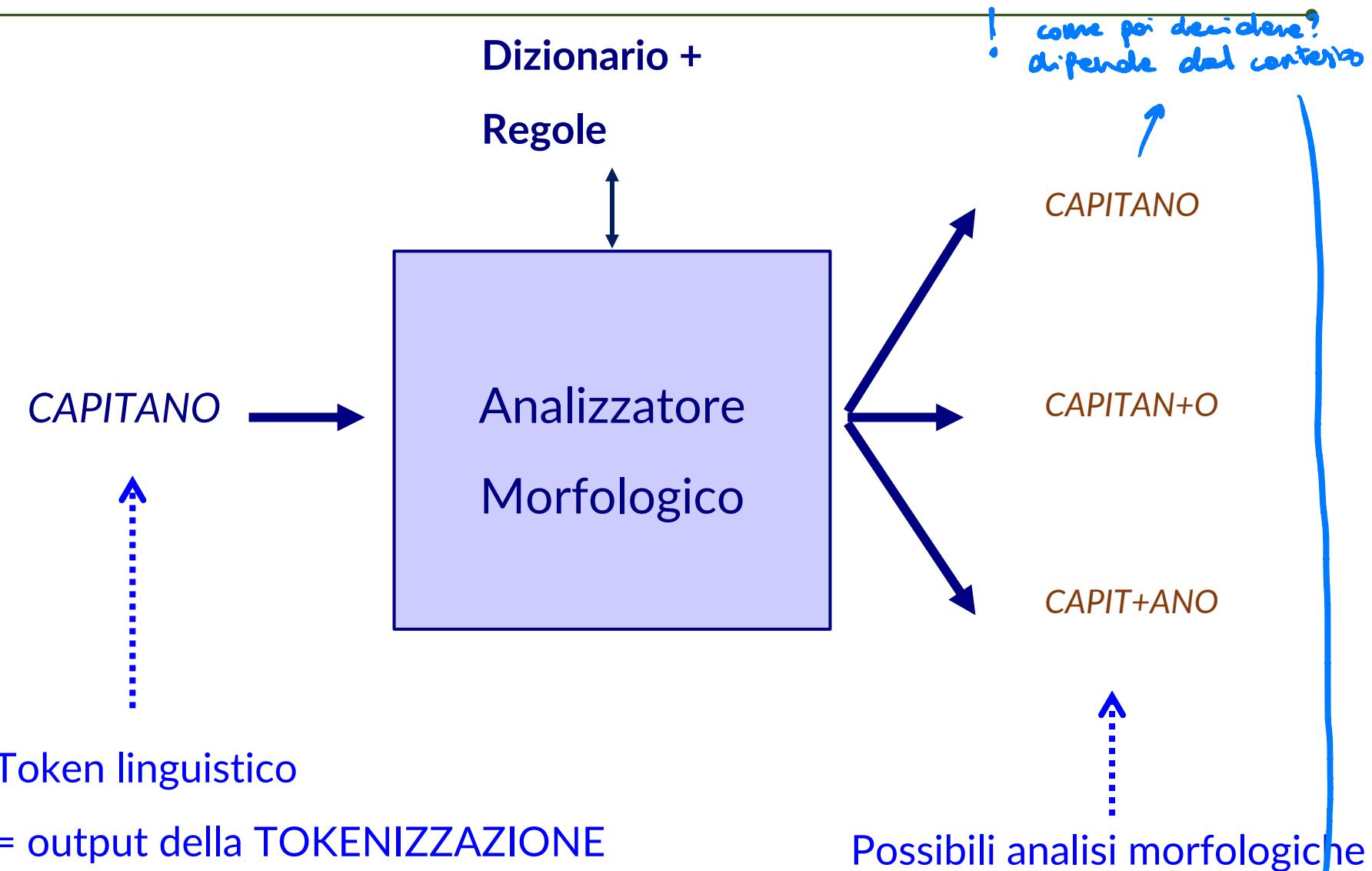
- Esempio 1: *capitano*

↙ es: *Tot*o**

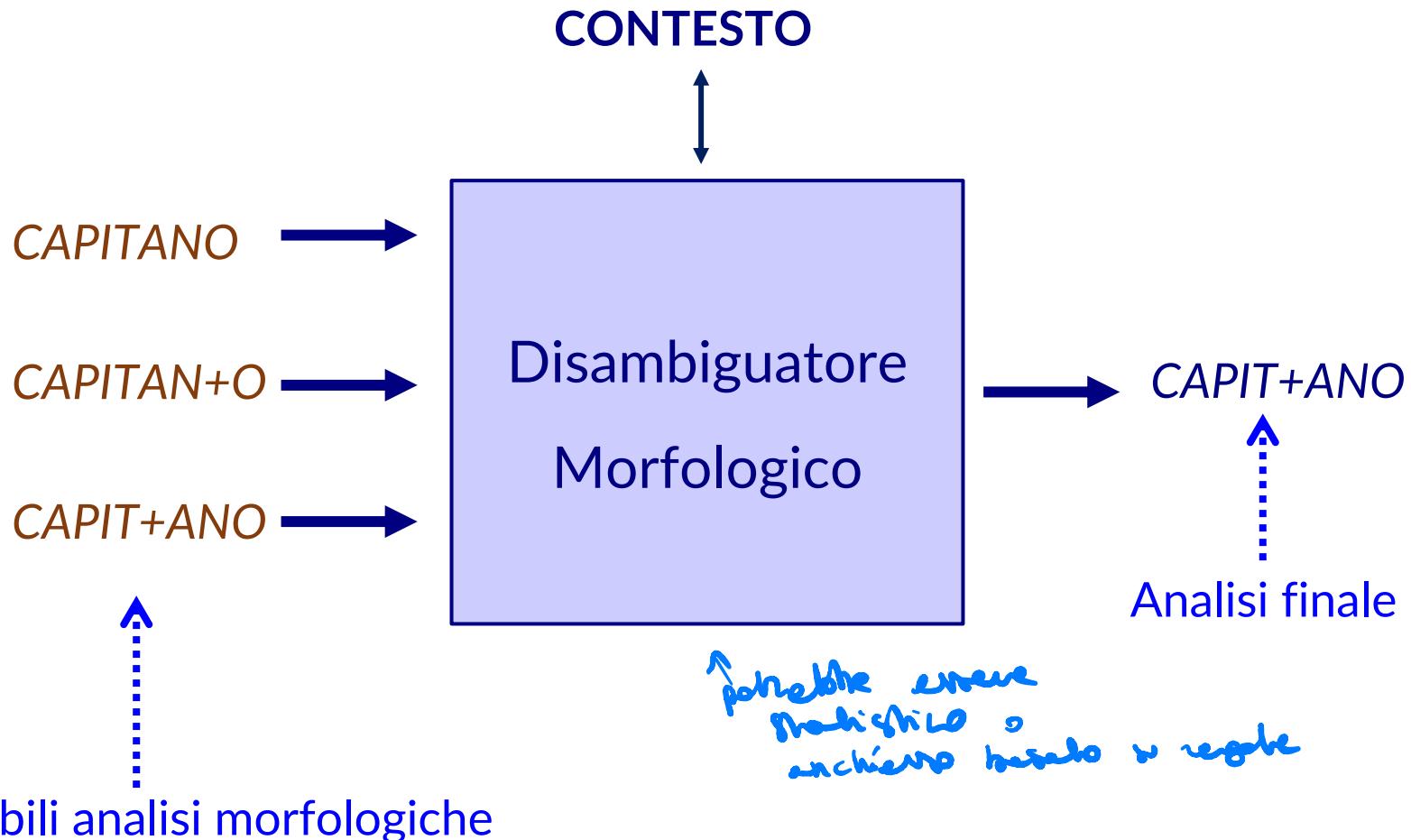
- CAPITANO (forma non declinabile)
- CAPITAN+O (nome o aggettivo o forma del verbo *capitanare*)
- CAPIT+ANO (forma del verbo *capitare*)
regole morfologiche

Altri suffissi: *mente* (avverbi), *one* (maggiorativo), ecc.

Analizzatore morfologico



Analizzatore morfologico



Ricerca sul dizionario: l'analizzatore morfologico

Presenza di suffissi

- Esempio 1: *barpile*

- *BARPILE* (forma non declinabile)
- *BARPIL+E* (nome o aggettivo o forma del verbo *barpilare*)
- *BARP+I+LE* (imperativo del verbo *barpere* con suffisso polinomiale)

Analizzatore morfologico e forme composte e multiple

Forme composte = generalmente una parola contenuto più una (o più) parole funzione

parte di parola che contiene i significato

- STAMPAMELO

orfem

- STAMP è una radice verbale
 - ↑ stemma : morfema "principale", + significativo
- A è un suffisso verbale
- ME e LO sono forme pronominali

↑ ↑
anche questi sono orfemi

se volesti fermarmi a questo livello, potrei fare informazioni relativi basato su stemmi

di fatto produrre in output un array [stamp, a, me, lo] e link a informazioni esterne per ogni segmento

è radice di "stampare"

• lemma → la forma normale di uno stemma o parola

Analizzatore morfologico e forme composte e multiple

Forme multiple = le diversi componenti sono nel dizionario ma la semantica non è compositiva

- più o meno: puntatore tra le parole per recuperare la giusta semantica
- prendere un abbaglio: (frasi fatte) rimandare all'interprete semantico

parola

- lemma \leftarrow dizionario : abbatteremo \rightarrow abbattere
↳ copiare "meglio": se lo uso con wordnet
- stemma \leftarrow morph-it : stampemmo \rightarrow stamp
↳ lo uso - copiare "di cosa parla" un documento; - di calio? di moglie?
il dominio
↳ ha a che fare con la semantica in maniera esplicativa

Analisi Morfologica -> Risorse

- Finite-state toolkit [https://en.wikipedia.org/wiki/Foma_\(software\)](https://en.wikipedia.org/wiki/Foma_(software))
- Foma <https://code.google.com/p/foma/wiki/GettingStarted>
- Morph-it! <https://docs.sslmit.unibo.it/doku.php?id=resources:morph-it>
 - SFST <http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html>
 - FSA utilities <http://www.jandaciuk.pl/fsa.html>
- TULE: <http://www.di.unito.it/~mazzei/software/tule/tule-jan-2012.tar.gz>
- Lemmatizzazione vs. Stemming

Parti del Discorso - Part of Speech

Aristotele, a.k.a lexical categories, word classes, "tags", PoS

School grammar:

- nome
- verbo
- aggettivo
- avverbio

sono
conjugati

sono open class/
category
CONTENT WORDS

- preposizione
- articolo
- pronome
- congiunzione
- interiezioni

hanno
una cardinalità
che varia molto molto poco

FUNCTION WORDS

- noun
- verb
- adjective
- adverb
- preposition
- determiner
- pronoun
- conjunction
- interjection

Parti del Discorso - Part of Speech

← ci riferiamo solo ai paradigmi

- Semantica

"la casa che vedendo" (S)
"la teoria che credevano" (O)
"la mangiare che credevano" (X)

La linguitica classifica individua queste a
categorie attraverso le relazioni

"parlo con la francese"

sono in relaz. sinategmatica

- Paradigmatico vs. Sintagmantico ->

Treccani: In linguistica, rapporti p., i rapporti che intercorrono tra gli elementi della frase e gli elementi che virtualmente potrebbero alternarsi con essi nella frase, distinti dai rapporti sintagmatici che intercorrono tra gli elementi che si succedono nella frase (per es., data la frase mangio una pera, tra mangio e mangiai, o mordo, vedo, dipingo, ecc., tra una e la o questa, ecc. intercorrono rapporti paradigmatici, mentre tra mangio e una pera, tra una e pera intercorrono rapporti sintagmatici).

Nome

- “Che cosa c'è in un nome? Ciò che noi chiamiamo con il nome di rosa, anche se lo chiamassimo con un altro nome, serberebbe pur sempre lo stesso dolce profumo.”
- Persone, oggetti, luoghi *← dal punto di vista semantico*
- Proprietà sintagmatiche:
 - its ability to occur with determiners (a goat, its bandwidth, Plato's Republic)
 - to take possessives (IBM's annual revenue)
 - to occur in the plural form (goats, abaci)
plate
- Nomi comuni, propri, di massa, contabili

Verbo

- Eventi, azioni, processi
- Molte forme morfologiche
 - tempo
 - modo
 - numero
- Tante categorie (e.g. ausiliari, modali, copula)

Aggettivi e avverbi

Aggettivi

- Proprietà
- Koreano

Avverbi

- “Modificano qualcosa”,
spesso verbi, ma anche
altri avverbi o intere frasi

Open vs. Closed classes

- Open vs. Closed classes

- Closed:

- determiners: *a, an, the*
- pronouns: *she, he, I* ↪ *cambiano i significati
dei secoli (es: "thou")*
- prepositions: *on, under, over, near, by, ...*
- Why “closed”?

- Open:

- | • Nouns, Verbs, Adjectives, Adverbs.

Open vs. Closed classes

Open class (**lexical**) words

Nouns

Proper

*IBM
Italy*

Common

*cat / cats
snow*

Verbs

Main

*see
registered*

Adjectives

old older oldest

Adverbs

slowly

Numbers

*122,312
one*

... more

Stanno + mehi. I numeri sono infatti, ma ad es. gli ordinamenti

Closed class (**functional**)

Determiners

the some

Conjunctions

and or

Pronouns

he its

Modals

*can
had*

Prepositions

to with

Particles

off up

... more

Interjections

Ow Eh



shanno + noi closed >=> questi non cambiano

STOP WORDS: preposizioni, avverbi, (alcune volte usati come sinonimi di FUNCTIONAL)

Parti del Discorso - Part of Speech

Quali strutture debbi usare? varie proposte ↗

- Penn PoS: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

- ISDT PoS: http://medialab.di.unipi.it/wiki/POS_and_morphology

↗ "Tutin University TreeBank"

- TUT PoS: <http://www.di.unito.it/~tutreeb/syntcat-22-7-02.doc>

- Google universal PoS: <http://www.petrovi.de/data/universal.pdf>

↳ 12 PoS: **NOUN** (nouns), **VERB** (verbs), **ADJ** (adjectives), **ADV** (adverbs), **PRON** (pronouns), **DET** (determiners and particles), **ADP** (prepositions and postpositions), **NUM** (numerals), **CONJ** (conjunctions), **PRT** (particles), '.' (punctuation marks) and **X** (a catch-all, e.g. abbreviations and foreign words).

PoS Tagging

è un problema di sequence labeling
è importante, perché do un tag
a ogni parola TENERE conto
di quanto c'è prima/dopo
la posizione come

- Words often have more than one PoS: back
 - *The back door* = JJ ← *adjective*
 - *On my back* = NN
 - *Promised to back the bill* = VB
- The PoS tagging problem is to determine the PoS tag for a particular instance of a word.

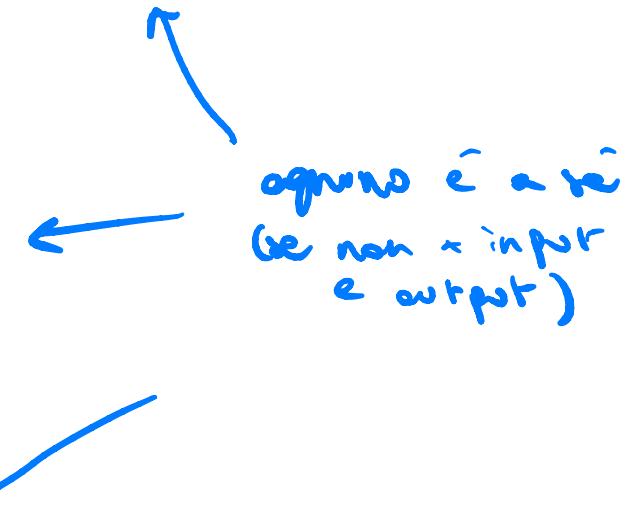
non si usano regole, ma modelli statistici

lo si fa ai tempi → es è possibile che "passiamo"
assegnare un PoS a "passato", poi a "me", poi a "lo"

Outline

- Il livello morfologico e l'analisi lessicale

- Parole di contenuto e di funzione
- L'analizzatore morfologico e il PoS tagging



- Il livello sintattico

- Sintassi e grammatiche
- I costituenti e le CFGs
- Le grammatiche a dipendenze

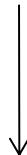
- Il livello semantico

- Semantica lessicale
- Semantica guidata dalla sintassi



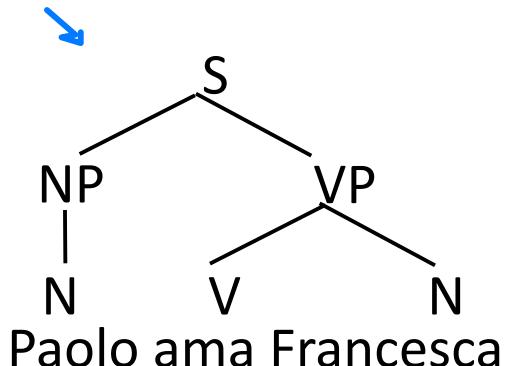
Natural Language Syntax

Paolo ama Francesca

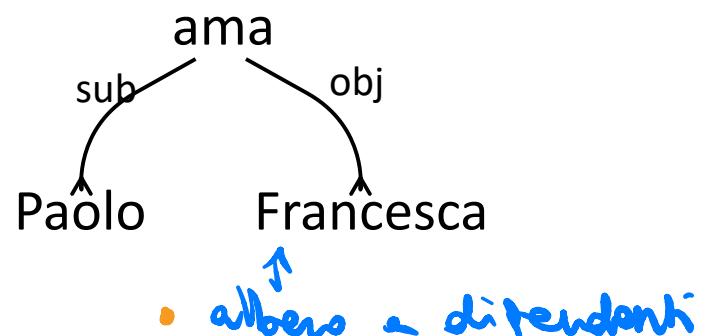


Syntactic Parsing: deriving a syntactic structure from
the word sequence

l'albero è una
molt. che si
fortemente
RICORDA



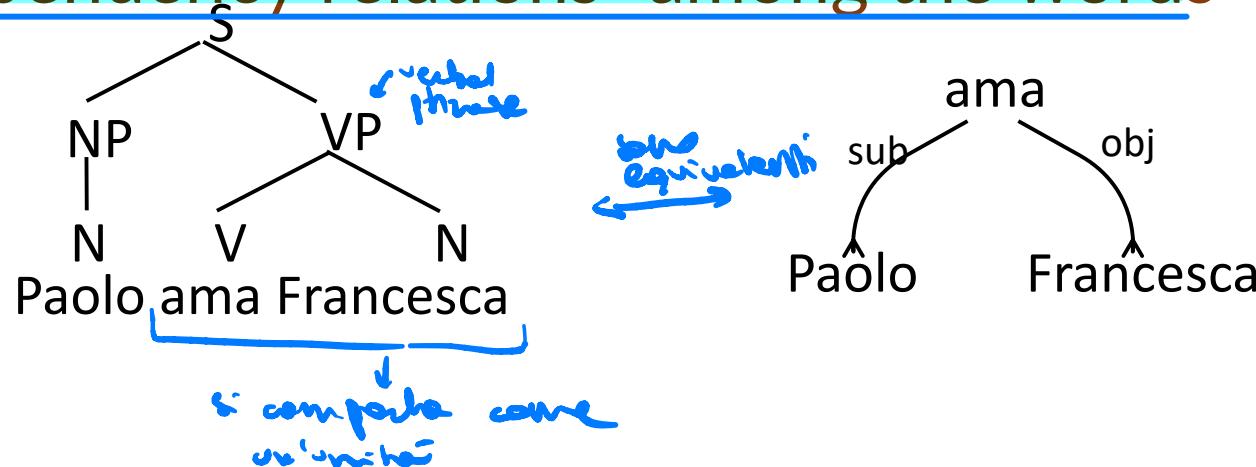
• albero è costituenti



A different syntactic structure: Constituency vs. Dependency

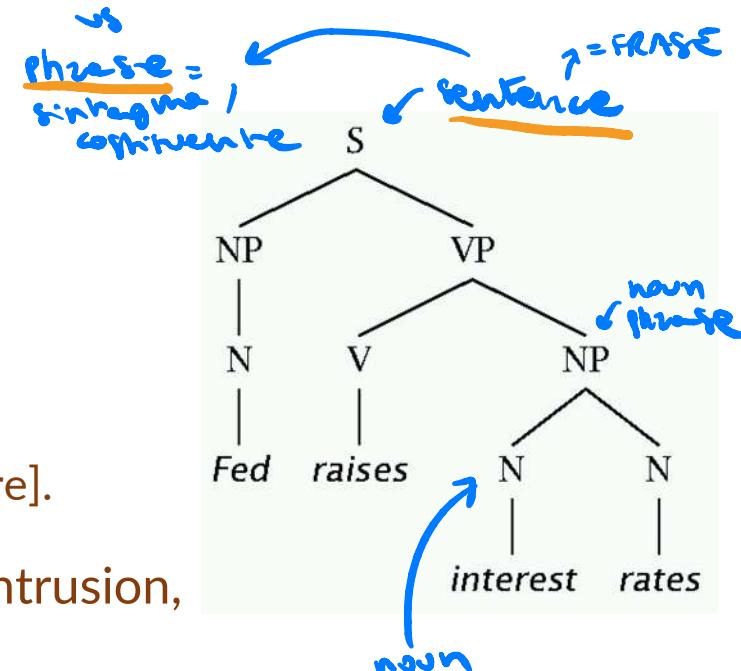
- a • Constituency structure represents the “grouping relations” among the words
- b • Dependency structure represents the “dependency relations” among the words

identifichi gli
gruppi di parole,
dei sotto-alberi

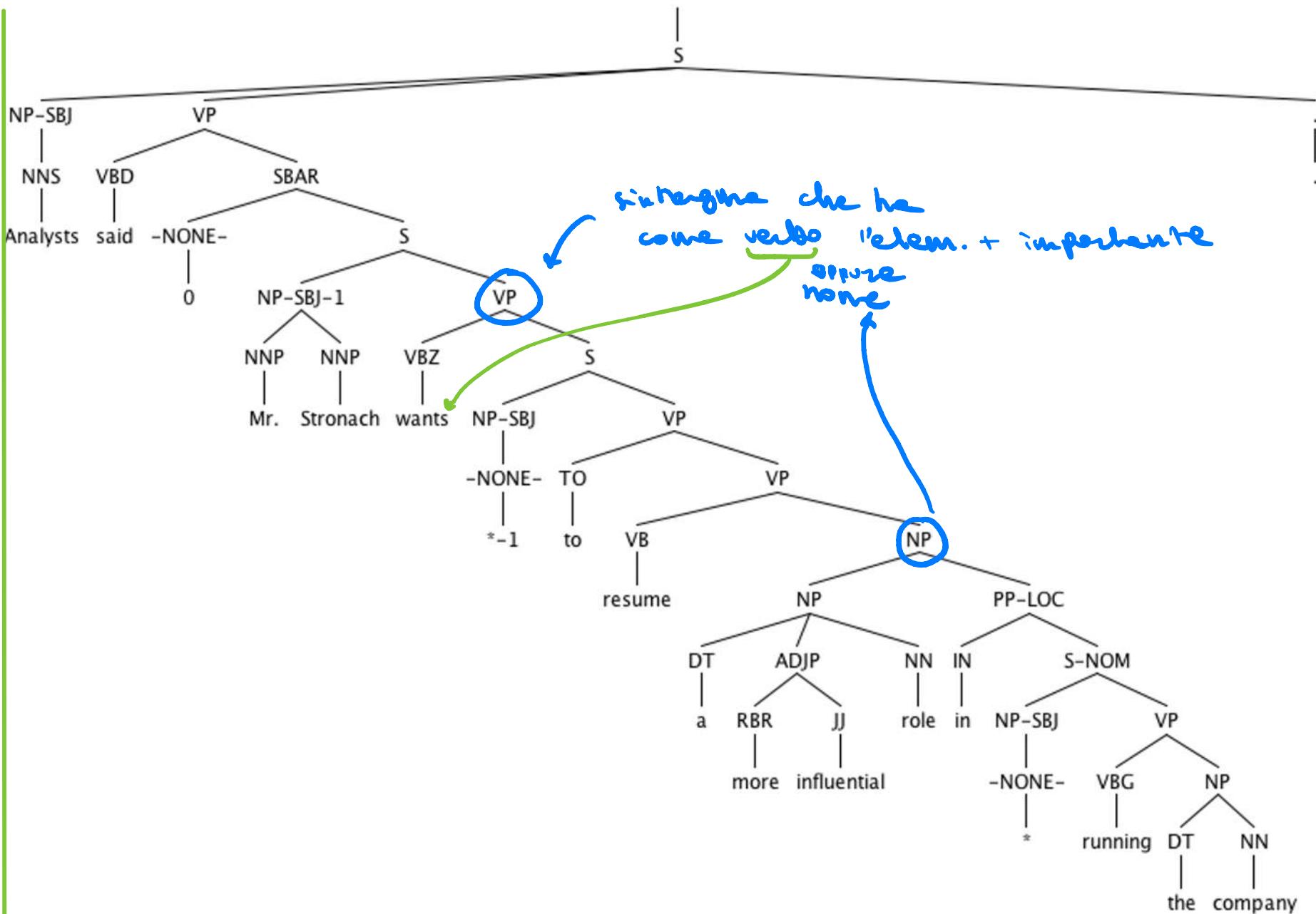


④ Costituenza

- Phrase structure organizes words into nested constituents.
- How do we know what is a constituent?
- Distribution: a constituent behaves as a unit that can appear in different places:
 - John talked [to the children] [about drugs].
 - John talked [about drugs] [to the children].
 - *John talked drugs to the children about
- Substitution/expansion/pro-forms:
 - I sat [on the box/right on top of the box/there].
- Coordination, regular internal structure, no intrusion, fragments, semantics, ...



gli NP formano una relaz. paradigmatica con tutti gli NP
VP VP



Headed phrase structure

- VP -> ... VB* ...
- NP -> ... NN* ...
- ADJP -> ... JJ* ...
↑ adjectives
- ADVP -> ... RB* ...
↑ adverbs
- SBAR(Q) -> S|SINV|SQ -> ... NP VP ...
- Plus minor phrase types:
 - QP (quantifier phrase in NP), CONJP (multi word constructions: as well as), INTJ (interjections), etc.

Costituenza

Costituente = gruppo di parole contigue

- che si comportano come un'unità [Fodor-Bever,Bock-Loebell] *che può apparire in posizioni diverse*
- che hanno delle proprietà sintattiche

Ex. preposed-postposed, substitutability.

Noun Phrases (NP), Verb Phrases (VP),...

- CFG: simboli non terminali \Leftrightarrow Constituenti (Chomsky)

context free grammar

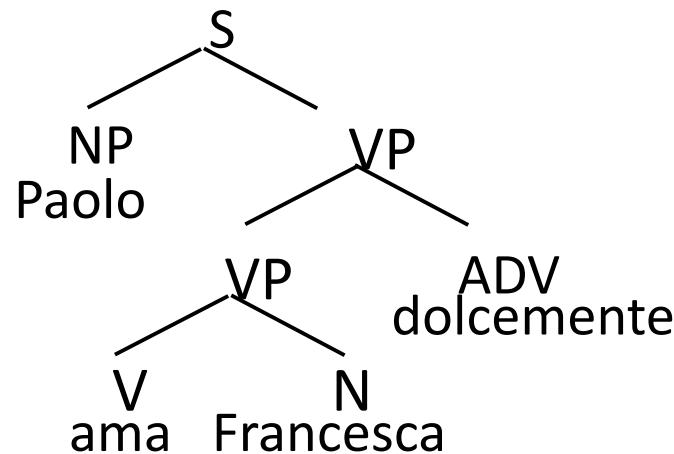
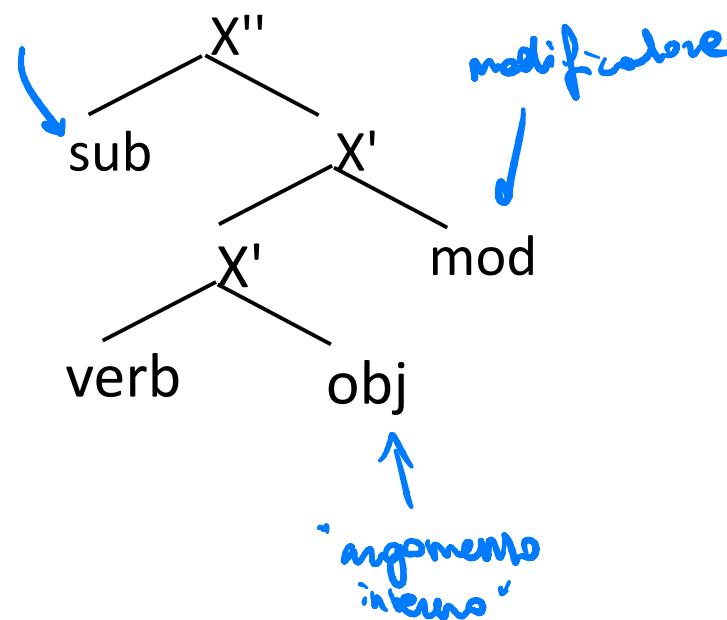
*ci dice ↑
"la grammatica -
costruttori può essere
associata a una CFG"*

Costituenza e relazioni grammaticali

La teoria X-barra ci permette di localizzare le

relazioni grammaticali nella struttura a costituenti

"argomento esterno"



! Chomsky ci dice che troviamo
il sogj. in alto a sinistra, il
cogg. in basso a destra

5

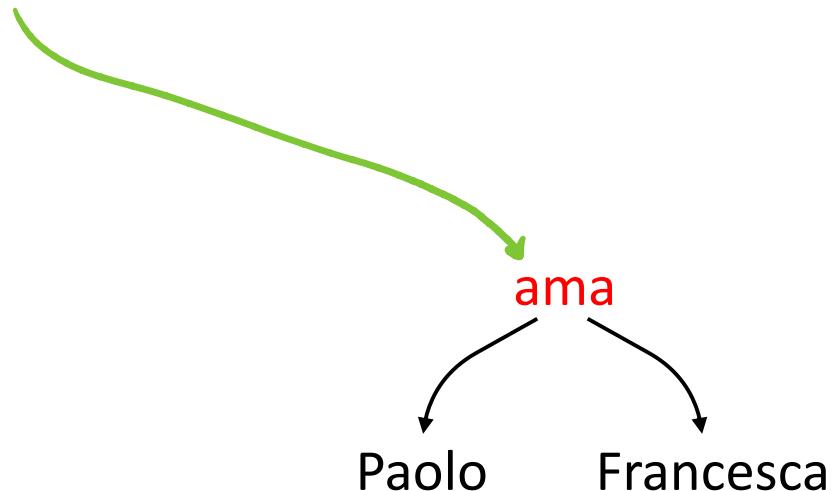
Dependency relation

'altra funzione

- Relation among two words:
 - Head: dominant word
 - Dependent: dominated word
- The head selects its dependents and
determines their properties
- Example: the verb determines the number of
its arguments

Dependency relation

- Head: dominant word



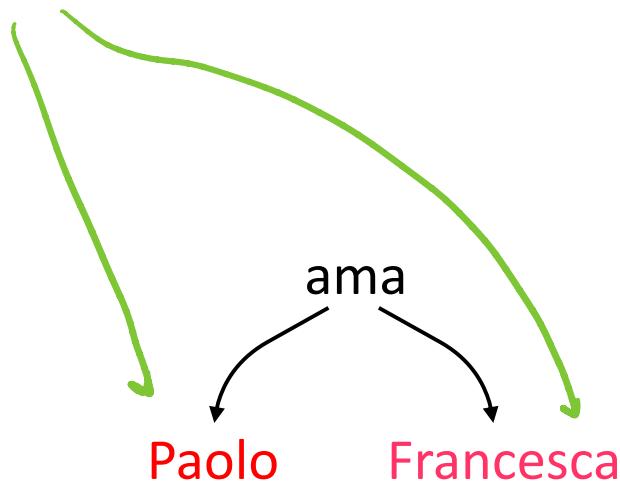
soglio rappresent.
delle relaz. gerarchiche,
di dominanza linguistica

NON ci sono nodi
intermedi
(cfr. albero coherenti)

relazione di
gruppo

Dependency relation

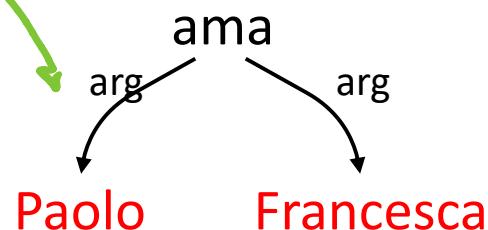
- Dependent: dominated word



Dependency relation

- Dependent 1

- argument

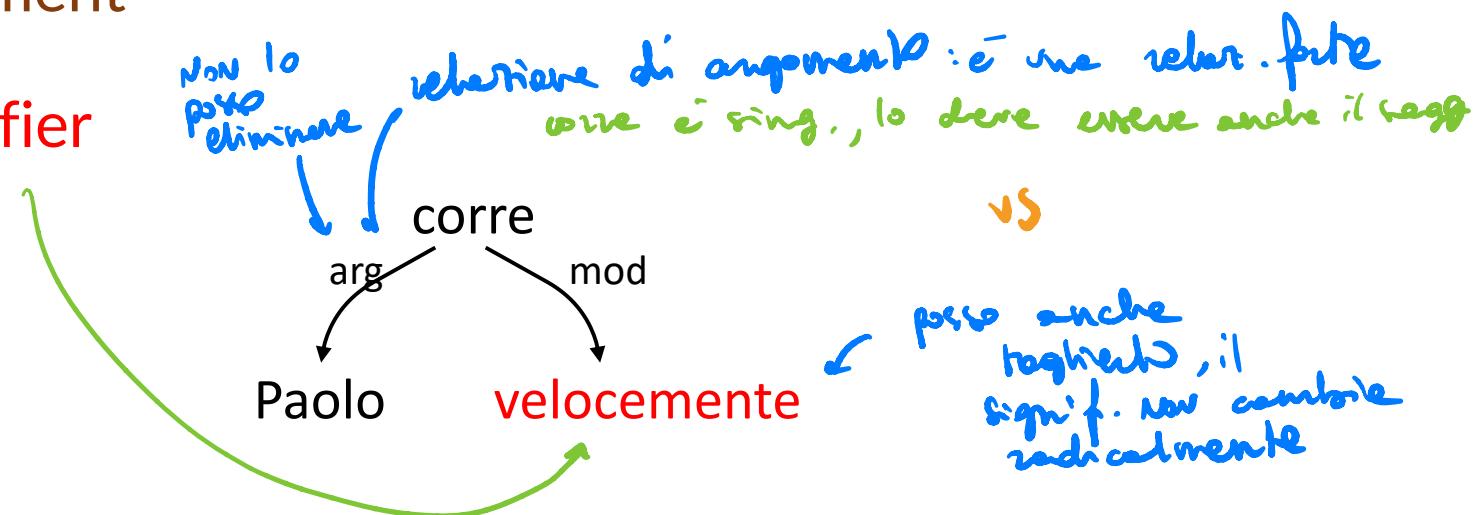


Dependency relation

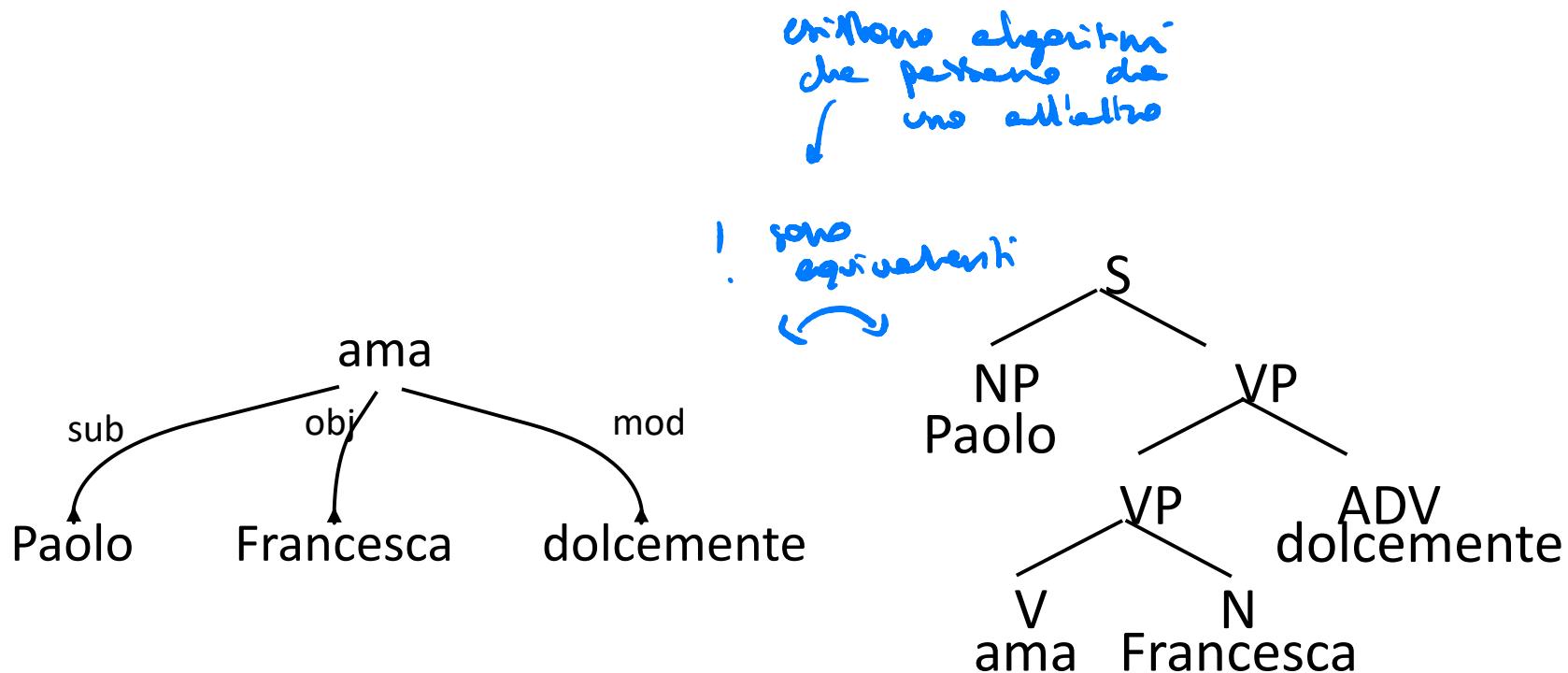
- Dependent 2

- argument

- modifier



Dependency and Constituency



Outline

• Il livello morfologico e l'analisi lessicale

- Parole di contenuto e di funzione
- L'analizzatore morfologico



• Il livello sintattico

- Sintassi e grammatiche
- I costituenti e le CFGs
- Le grammatiche a dipendenze

• Il livello semantico

- Semantica lessicale /atomica/ → riguarda le singole parole
- Semantica guidata dalla sintassi → "gruppi di parole"
↳ semantica formale

2 ipotesi fondamentali nello studio del linguaggio/NLP:
1) il linguaggio si può studiare attraverso i vari livelli (che sono "come stagne")
2) i livelli lavorano in maniera sequenziale

- Che cosa costruisce la gente quando comprende un discorso? Trasformazione dal linguaggio naturale in una qualche forma di rappresentazione della conoscenza
- Come viene eseguito il processo di costruzione?

Interpretazione semantica

Le strutture dati sono gli insiemi (e le loro interazioni)

① FASE 1: Semantica lessicale



- Connessioni legate al significato dei vari lessemi
- Struttura interna dei lessemi legata al significato

regole di associazione / "la stringa"

Lessema = coppia forma-significato = elemento del lessico

la rappresentaz. del testo e comprensione

- Forma ortografica e fonologica
- Senso

Quando un essere umano vuole sapere il significato di una parola, usa il vocabolario



Vocabolario

- **right** = *adj.* located nearer the right hand esp. being on the right when facing the same direction as the observer
- **left** = *adj.* located nearer to this side of the body than the right
- **red** = *n.* the color of blood or a ruby
- **blood** = *n.* the red liquid that circulates in the heart, arteries and veins of animals

quel e il problema?

Vocabolario

- **right** = *adj.* located nearer the **right** hand esp. being on the **right** when facing the same direction as the observer
- **left** = *adj.* located nearer to this side of the body than the **right**
- **red** = *n.* the color of **blood** or a ruby
- **blood** = *n.* the **red** liquid that circulates in the heart, arteries and veins of animals

ci sono delle circoscrizioni infinite! (bad revision)

non ci danno delle definizioni separate per il significato delle parole, ma c'è dare delle definizioni separate su INSIEME

Vocabolario

- **right** = *adj.* located nearer the **right** hand esp. being on the **right** when facing the same direction as the observer
- **left** = *adj.* located nearer to this side of the body than the **right**
- **red** = *n.* the color of **blood** or a ruby
- **blood** = *n.* the **red** liquid that circulates in the heart, arteries and veins of animals

QUINDI, INNECE, SÌ COMINCIO A ~

RELAZIONE TRA LESSEMI

← relazioni intrinseche

- Omonimia, polisemia, sinonimia, iponimia

"red" belongs to
"colors"

Homonymy and Polysemy

Omonimia: 2 lessemi con la stessa forma

(ortografica) hanno due sensi diversi

- A bank can hold the investments
- We can go on the right bank of the river



a volte si scrive infatti che esiste
bank² e bank² ← 2 lesseni

Homonymy and Polysemy

Polisemia: lo stesso lesema (stessa forma) ha due sensi diversi

← e' più una questione di spiegazione

- A bank can hold the investments
- He got the blood from the bank

In questo caso hanno la stessa etimologia: i sensi sono in qualche relazione

Homonymy and Polysemy

- Quanti sensi ci sono?
- In che relazione sono?
- Come possono essere distinti?

Zeugma:

- Giovanni ha aperto il gioco
- Maria ha aperto la finestra

sono omonimi

Giovanni ha aperto il gioco e la finestra

Synonymy

Sinonimia: Due lessemi con forma diversa hanno lo stesso senso (sostituibilità)

- How big is that plane?
- How large is that plane?

 Synset = {^{big, large}_{Bank, Deposit}}
condividono il significato

• Ma a volte ci sono delle differenze

- What is the cheapest first class *fare*?
- What is the cheapest first class *price*?

non le sostituisco con "large sister"



Polysemy (big sister), subtle shades of meaning (fare-price), collocational constraints (big mistake), register (academic slang).

Hyponymy

• **Iponimia:** due lessemi di cui uno denota una sottoclasse dell'altro

- Automobile è un iponimo di veicolo
- Veicolo è un iperonimo di automobile

veicolo
I
automobile

- Quella è un'automobile → quello è un veicolo
- (?) Quello è un veicolo → quella è un'automobile

le storie che fanno alberi. L'output del livello sintattico è un albero, e qui lo si arricchisce



FASE 2: Semantica compositiva

- Domanda:

predicato



- Giovanni ha acquistato un'automobile ->
- Giovanni ha comprato un veicolo?

! la semantica lessicale mi dice
che "acquistato" e "comprato" sono sinonimi e che "automobile" e "veicolo" sono iperonimi

- Traduzione dal linguaggio naturale in una qualche

forma di rappresentazione della conoscenza

- Ispirata alla semantica dei linguaggi formali

- significato di "X + Y" in aritmetica
- significato "X AND Y" in logica

Semantica compositiva

!

- La semantica di un sintagma è funzione della semantica dei sintagmi componenti; non dipende da altri sintagmi esterni al sintagma stesso.



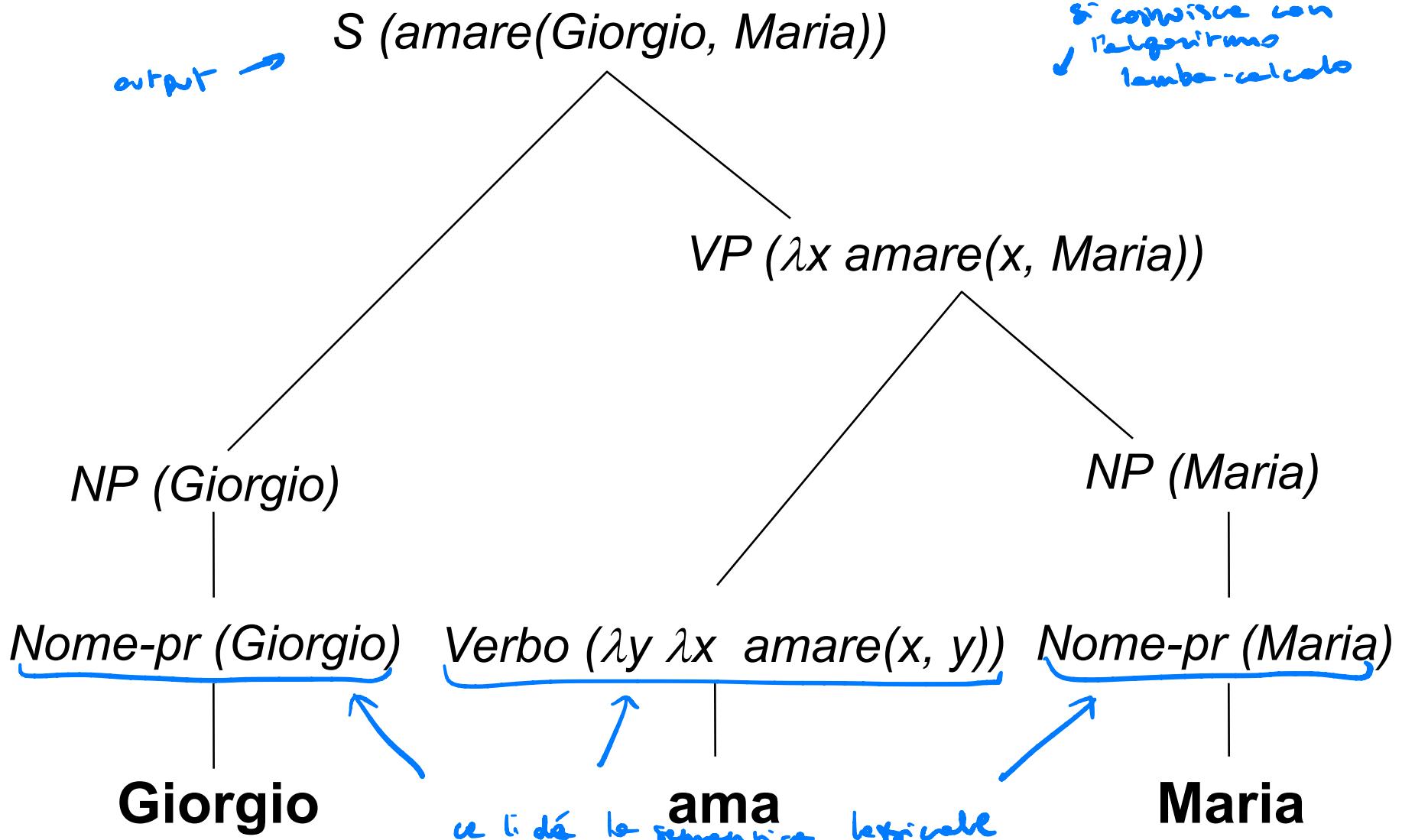
- Conoscendo il significato di X, Y, e +, possiamo comporre il significato “X+Y”
- Con la semantica compositiva, un numero finito di regole controlla un numero infinito di situazioni

Semantica del linguaggio naturale (Montague 1974)

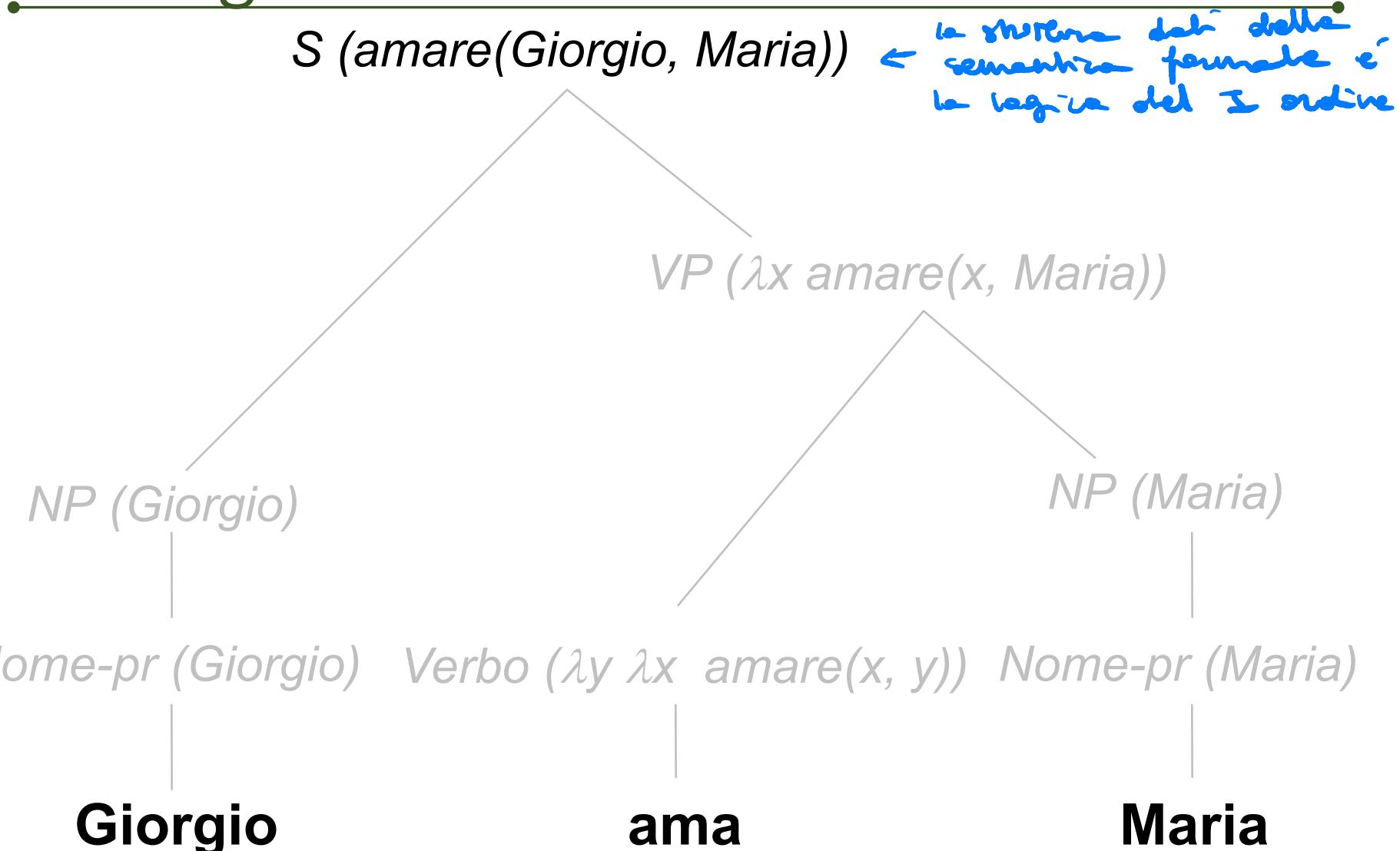
- Quali rappresentazioni semantiche si associano con quali sintagmi
- “Giorgio ama Maria”: Ama(Giorgio, Maria)

una qualche re è ciò forte molto simile
"loves" oppure
P19235ø

Interpretazione semantica di “Giorgio ama Maria”



Interpretazione semantica di “Giorgio ama Maria”



Reasoning!

permette di poter rispondere a domande di logica (question answering)

Deduzione

$$\forall X \text{ uomo}(X) \rightarrow \text{amare}(X, \text{Maria})$$

&

$$\text{uomo}(\text{Giorgio})$$

l'unica che preseva la verità

ex: Prolog

$$\Rightarrow \text{amare}(\text{Giorgio}, \text{Maria})$$

Induzione

("Generalizzazione", "Stereotipizzazione")

$$\text{uomo}(\text{Giorgio}) \& \text{amare}(\text{Giorgio}, \text{Maria})$$

$$\text{uomo}(\text{Paolo}) \& \text{amare}(\text{Paolo}, \text{Maria})$$

....

$$\Rightarrow \forall X \text{ uomo}(X) \rightarrow \text{amare}(X, \text{Maria})$$

Abduzione

$$\forall X \text{ uomo}(X) \rightarrow \text{amare}(X, \text{Maria})$$

&

$$\text{amare}(\text{Giorgio}, \text{Maria})$$

$$\Rightarrow$$

potrebbe essere falso
potrebbe essere un caso

$$\text{uomo}(\text{Giorgio})$$

Metasemantica

→ es: semantica diatributiva

- Semantica lessicale vs. semantica compositiva
- Fosco Maraini, "Gnosi delle Fànfole" (1978)
 - > https://www.youtube.com/watch?v=62I8A_kCNwA

Il Lonfo

Il Lonfo non vaterca né gluisce
e molto raramente barigatta,
ma quando soffia il bego a bisce bisce,
sdilena un poco e gnagio s'archipatta.
È frusco il Lonfo! È pieno di lupigna
arrafferia malversa e sofolenta!
Se cionfi ti sbiduglia e ti arrupigna
se lugri ti botalla e ti criventà.

Eppure il vecchio Lonfo ammargelluto
che bete e zugghia e fonca nei trombazzi
fa legica busia, fa gisbuto;
e quasi quasi in segno di sberdazzi
gli affarferesti un gniffo. Ma lui, zuto
t' alloppa, ti sberneccchia; e tu l'accazzi.

Il Lonfo

Il Lonfo non vaterca né gluisce

e molto raramente barigatta,

ma quando soffia il bego a bisce bisce,

sdilena un poco e gnagio s'archipatta.

È frusco il Lonfo! È pieno di lupigna

arrafferia malversa e sofolenta!

Se cionfi ti sbiduglia e ti arrupigna

se lugri ti botalla e ti criventà.

Eppure il vecchio Lonfo ammargelluto

che bete e zuggchia e fonca nei trombazzi

fa legica busia, fa gisbuto;

e quasi quasi in segno di sberdazzi

gli affarferesti un gniffo. Ma lui, zuto

t' alloppa, ti sbernecchia; e tu l'accazzi.



Semantica Distribuzionale (vettoriale)

il signific. di una parola è corretto solo nel contesto
in cui vive



- “You shall know a word by the company it keeps!” (Firth 1957), i.e. the meaning of a word is thus related to the distribution of words around it.

- *A bottle of tesguino is on the table.*
- *Everybody likes tesguino.*
- *Tesguino makes you drunk.*

word -> numerical vector -> embedding

→ word embedding → rappresentare i vettori che significano una parola

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

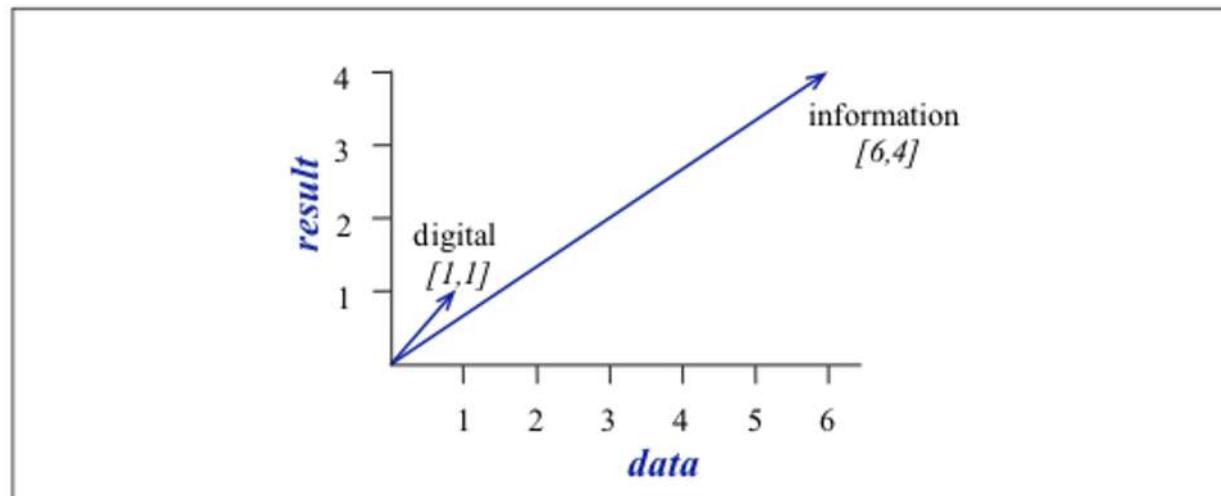


Figure 15.5 A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *result*.

-> Positive Pointwise Mutual Information (PPMI)

Sparse vs. dense vectors

- Vectors are

- annote*
 - long (length $|V|=20,000$ to $50,000$)
 - sparse (most elements are zero, cf. one-hot representation)

- Alternative, learn vectors which are

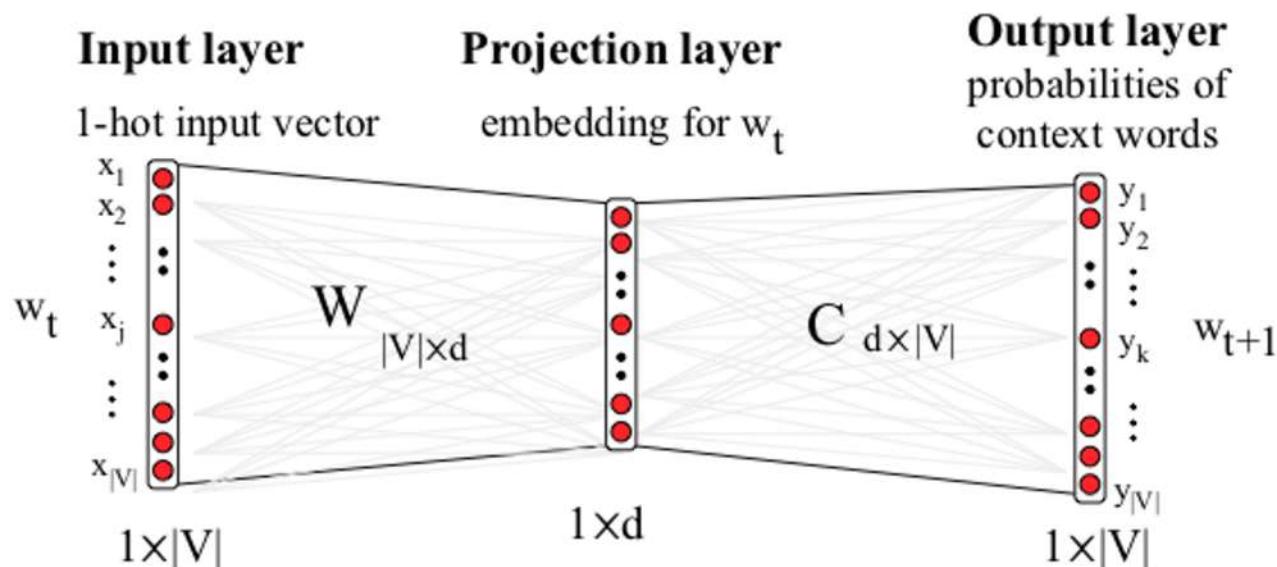
- one*
 - short (length 200-1000)
 - dense (most elements are non-zero)

- Short vectors may be easier to use as features in machine learning (less weights to tune)

Neural Embeddings

Non va bene = fare di di più.
(meglio wordNet)
← ottimo → avere le sue parole
rappresentate vicine e opposte

- The neural models therefore learn an embedding by starting with a random vector and then iteratively shifting a word's embeddings to be more like the embeddings of neighboring words, and less like the embeddings of words that don't occur nearby.
- Skip-Gram



Properties of Embeddings

Redmond	Havel	ninjutsu	graffiti	capitulate
Redmond Wash.	Vaclav Havel	ninja	spray paint	capitulation
Redmond Washington	president Vaclav Havel	martial arts	graffiti	capitulated
Microsoft	Velvet Revolution	swordsman ship	taggers	capitulating

- $\text{vector('king')} - \text{vector('man')} + \text{vector('woman')} \approx \text{vector('queen')}$
- $\text{vector('Paris')} - \text{vector('France')} + \text{vector('Italy')} \approx \text{vector('Rome')}$

↑
operazioni numeriche

va bene = un generatore
casuale di storie

non è
deduzione /
logica

Le semantica compositiva NON si riusce a fare coi vettori!

Un altro livello: la pragmatica

“Oggi mi trovo ad Alessandria”

2 tipi di significato: **presupposto linguistico**: il significato non cambia se lo dice un altro, o in un altro momento
grounded: se collega il signif. al mondo, cambia in base a chi lo dice (e determina se è vera o falsa)

ontologie di common sense

L'interpretazione di “io” (sottinteso) e “oggi” dipende da

chi enuncia la frase e quando, rispettivamente

enuncia il significato alla realtà

Pragmatica

■ **Anafora:** sintagmi che si riferiscono a oggetti precedentemente menzionati

- “La torta era sul tavolo. Giorgio **la** divorò”
- “In giardino c’erano il cane e il gatto che giocavano con un pezzo di stoffa. **Il felino** lacero’ la stoffa.
- “Dopo essersi fidanzati, Giorgio e Maria trovarono un prete e si sposarono. Per la luna di miele, **essi** andarono ai Caraibi.

Strutture vs. neuroni

- What innate priors should we build into the architecture of deep learning systems? <https://youtu.be/fKk9KhGRBdI>



R	semantiche lessicale insiemi (e loro interazioni)	vs	semantiche costruttive alberi (arlicchiti)
• sintesi di dati			
• obiettivo	• signif. alle singole parole		• signif. - alba frase/ sequenze di parole