

## TLN [Radicioni]



Questo è un riassunto / dispensa della 2° parte del professore Radicioni del corso di Tecnologie del Linguaggio Naturale (2022). Questo documento è stato definito facendo un *merge* tra le cose dette a lezione dal prof ed altri vari riassunti trovati nel repo.



Preciso che potrebbero esserci errori di battitura e/o eventuali parti non ben scritte



Questo documento è stato fatto con Notion, se lo usi e vuoi direttamente la pagina (per fare modifiche, ecc..), sono disponibile a passarla. Scrivimi → <https://t.me/Aleeeee96>

## ▼ Introduzione

### Semantica Lessicale

La semantica lessicale è lo studio dei significati delle **parole** e delle loro relazioni.

In particolare ci si chiede

- cosa vogliono dire i singoli termini
- perché hanno quel significato

- e come rappresentiamo la conoscenza e le relazioni di questi termini

Ci sono due principali problemi (per noi banali) nella semantica lessicale che si vogliono risolvere

1. **Polisemia** (vogliamo scoprire quale sia il significato delle parole in una frase)

**Esempio** *Ho acquistato un sacco (uno zaino)*

*Si è infilato in un sacco di guai (molti guai)*

2. **Semantica delle frasi** (idiomi) (vogliamo capire qual è il significato della frase a partire dal significato delle singole parole)

**Esempio** *Il ladro ha vuotato il sacco*

Inoltre è importante il concetto di **contesto** che è l'insieme degli elementi adiacenti a una parola e può essere

- **sintattico** insieme degli elementi adiacenti a una parola dal punto di vista delle loro proprietà sintattiche (può essere nominale, verbale, aggettivale)
- **semantico** elementi adiacenti alla parola, dal punto di vista delle loro proprietà semantiche
  - **Esempio** “Saltare un fosso” vs “Saltare un pasto”
- **situazionale**

**Esempio** Nell'enunciato “*il tuo amico è forte*”, il significato di *forte* può essere energico, simpatico, e può divenire chiaro solo nel contesto specifico in cui è utilizzato

Per risolvere questi problemi abbiamo parlato del **Principio di composizionalità** che spiega come il significato degli enunciati si formi a partire dal significato degli elementi lessicali che li compongono.

Ci sono due approci

- **enumerazione dei sensi** i diversi sensi associati a un elemento lessicale sono elencati nella parola, insieme alle restrizioni lessicali che specificano i contesti in cui i diversi significati possono attivarsi.

- **concezione dinamica del significato** il significato di ciascuna parola interagisce con il significato delle parole adiacenti

Fra gli elementi che possiamo considerare ci sono

- **co-composizione** il significato del verbo potrebbe essere determinato da quello dei suoi argomenti

**Esempio** “*luca taglia il pane*” (affettare) “*luca taglia l’erba*” (falciare)

- **forzatura (o coercion) del tipo semantico** un verbo in combinazione con un nome specifico lo “spinge” a significare ciò che la semantica del verbo richiede, eventualmente variandone il tipo semantico

**Esempio** “*inizio un articolo*” vuol dire che inizio a leggerlo o scriverlo mentre “*iniziare la lezione*” significa iniziare a

- **legamento selettivo** l’aggettivo può selezionare una specifica porzione del significato del nome

**Esempio** “*buon coltello*” taglia bene “*buon medico*” è un medico competente

## Ontologia

Una ontologia definisce un insieme di primitive (**classi** e **relazioni**) che permettono di modellare un dominio di conoscenza. Ciascuna classe e relazione includono informazioni sul loro significato e vincoli su come applicarle in maniera logicamente consistente.

**Esempio** il verbo “*mangiare*” deve avere un soggetto vivente e un oggetto categorizzato come “*edibile*”

## Perche si usa l'ontologia?

Ci permette di condividere la descrizione di entità e relazioni di un certo dominio per finalità di ricerca, o semplicemente di riutilizzo dei dati

Ogni rappresentazione approssima in maniera imperfetta la realtà (non ho un punto di vista completo con una sola rappresentazione). Quindi scegliere una rappresentazione piuttosto che un'altra significa decidere cosa vedere e decidere cosa escludere, questo concetto si chiama

## ▼ Commitments Ontologico

una rappresentazione che coglie alcuni elementi e ne trascura altri. Tutte le rappresentazioni della realtà sono imperfette ed approssimate, poiché quando si sceglie come rappresentare qualcosa, tralasciamo inevitabilmente alcuni aspetti dell'oggetto che si sta rappresentando.

Il commitment ontologico non è un effetto collaterale, ma è l'essenza che fa parte di un'ontologia

L'ontologia permette quindi di condividere questa conoscenza, in funzione del riutilizzo dei dati e dell'informazione.

Tuttavia questa definizione è troppo generale, in quanto ammette che tutto possa essere considerato un'ontologia.

**Esempio** “*Paolo mangia la mela*” “*La macchina mangia la strada*”  
non sono la stessa cosa

Possiamo vedere un'ontologia come una esplicita specificazione di **concettualizzazioni**.

Con questa visione dell'ontologia è possibile avere situazioni diverse descritte da vocabolari differenti che però condividono la stessa concettualizzazione. Il concetto torna molto utile nella semantica, perché il significato di una frase, ad esempio, rimane uguale indipendentemente dalla lingua (e il suo vocabolario) presa in esame.

**Esempio** diversi termini (*mela* e *apple*) che coinvolgono gli stessi oggetti, possono avere la stessa concettualizzazione

Ecco quindi che si viene a creare una relazione stretta fra le ontologie e la semantica, con aspetti

diversi del linguaggio catturati tramite l'uso di ontologie specifiche per il ruolo

## Ontologia vs Semantica

È importante non confondere **ontologie** e **semantica**. Nell'ontologia studiamo cosa c'è; nella semantica studiamo a cosa si riferisce a cosa c'è. Infatti le parole che hanno determinati ruoli semantici ricopriranno determinate componenti all'interno dell'ontologia.

## Ontologia e KB

Le ontologie possono essere viste anche come parte di una base di conoscenza (KB). Queste infatti

sono formate da due parti

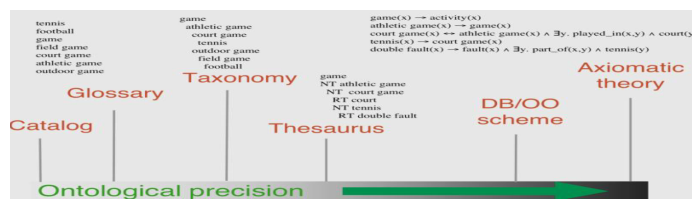
- **Terminological component (T-box)**, rappresenta proprio la parter terminologica, proprio dell'ontologia e indipendente da un particolare stato
- **Assertional component (A-box)**, riflette uno specifico stato della realtà

**Esempio** “*Gli uccelli volano*” fa parte della componente terminologica. “*Tweety vola*” fa parte della componente asserzionale, poiché un determinato uccello sta volando

## Livelli di ontologie

Possiamo costruire diversi tipi di ontologie: più un'ontologia è complessa, tanto più sarà precisa e informativa nel rappresentare la realtà.

Si parte dall'insieme di termini senza relazioni fra di loro (catalogo) per poi passare alla tassonomia che istituisce una relazione gerarchica. Successivamente abbiamo il thesaurus che fornisce informazioni semantiche, ecc..



## Ontologia e lessico

Il lessico è lo strumento mediante il quale rappresentiamo dei concetti.

Fra gli elementi del lessico ci possono essere diverse relazioni

- **Omonimia** due lessemi con la stessa forma, ma con sensi diversi (significati diversi).

**Esempio** *pesca* (verbo) e *pesca* (frutto).

- **Polisemia** stesso lessema con due sensi diversi (significati simili).

**Esempio** *banca* e *banca del sangue*. Stesso lessema, ma i sensi sono diversi, sebbene correlati.

- **Sinonimia** due lessemi con forma diversa, ma stesso significato.

**Esempio** *big* e *large*

- **Iponimia e iperonimia** due lessemi di cui uno denota una sottoclasse dell'altro.

**Esempio** automobile è un iponimo (più specifico) di veicolo, mentre veicolo è un iperonimo (più generale) di automobile.

- **Meronymia** e **onymia** due lessemi i cui significati sono uno la parte dell'altro.

**Esempio** *dito* è meronimo di *mano* e *mano* è olonimo di *dito*.

- **Antinomia** due lessemi con significato opposto

**Esempio** *ricco* e *povero*.

Tali relazioni sono utili per costruire un'ontologia, istituendo così delle relazioni fra concetti. Ad esempio, la relazione di iponimia è molto simile alla relazione **subclass-of**, mentre la meronymia è molto simile alla relazione **has-part**.

## Ontology design

Una prima, grande, distinzione fra gli elementi di un'ontologia è quella di **entità** ed **eventi** dove

- **entità** (**endurant**), sono oggetti che continuano per un periodo di tempo mantenendo la propria identità

**Esempio** *cellula*, *nucleo*

- **eventi** (**perdurant**), sono oggetti che accadono, si svolgono o si sviluppano nel tempo

**Esempio** *replicazione del DNA*, *mitosi*, *divisione della cellula*

Un'ontologia costruita su questo concetto viene definita **ontologia fondazionale** (upper ontology). Si tratta di un'ontologia che cattura distinzioni di base valide in più domini; in alcuni ambiti questo tipo di ontologie tornano molto utili, in altri no. Il campo dell'elaborazione del linguaggio naturale è uno dei campi in cui tornano utili.

**Esempio** un'ontologia che categorizza le verdure sarà un'ontologia di dominio botanico, mentre un'ontologia che cerca di rappresentare il concetto di tempo, spazio e materia sarà un'ontologia fondazione (di base)

Di questa tipologia (fondazionale) abbiamo visto **DOLCE**

## DOLCE

Si tratta di un'ontologia fondazionale che nasce da tre aspetti

## ▼ Distinguere endurants e perdurants

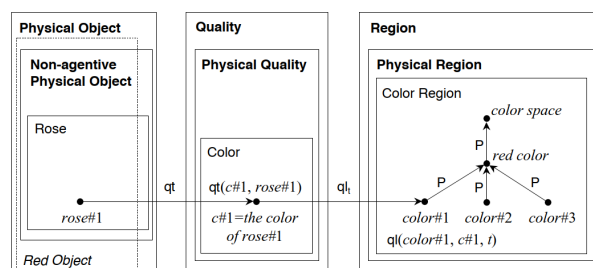
dicendo che

- gli **endurants** esistono nel tempo, possono cambiare naturalmente nel tempo, possono essere costituiti da parti non essenziali e tutte le parti essenziali che li compongono sono sempre presenti e non decadono
- i **perdurants** accadono in una finestra temporale, non cambiano nel tempo, ogni parte è essenziale, ma non è detto che tutte le parti essenziali siano presenti nel tempo

**Esempio** In DOLCE la relazione tra endurants e perdurants è quella di partecipazione. Ad esempio una persona (endurant) può partecipare ad una discussione (perdurant)

## ▼ Teoria delle qualità

dove le qualità servono a descrivere gli oggetti, e i valori delle qualità vengono presi da un quality space (spazio che contiene tutte le qualità (qualia) di una certa entità. “La penna rosso scuro” ha uno spazio concettuale che è quello dei colori) associato.



**Esempio** Quando due rose hanno lo stesso colore intendiamo che le loro qualità, che sono distinte, hanno la stessa posizione nel quality space dei colori possibili, ma non esistono due rose con la stessa identica qualità. Questo ci serve per capire la relazione tra “rosso” inteso come aggettivo e “rosso” inteso come sostantivo.

## ▼ Approccio moltiplicativo

afferma che oggetti/eventi diversi possono trovarsi nella stessa posizione spaziotemporale mantenendo però proprietà distinte

**Esempio** l'argilla di cui un vaso è costituito si troverà fisicamente nello stesso posto del vaso, ma sarà considerata un'entità diversa. Le due entità restano separate perché possiedono delle caratteristiche tipiche che non si trovano nell'altra entità e viceversa

## Knowledge Representation

La KR tramite logica del prim'ordine, (primo approccio storico), rende imprescindibile l'uso di regole per poter legare fra di loro le formule e derivare della conoscenza implicita (inferenza). L'inferenza però ha un costo piuttosto elevato e sono quindi nate rappresentazioni alternative più dirette (le reti semantiche)

### Reti semantiche

Le reti semantiche sono dei grafi in cui i **nodi** rappresentano i concetti e gli **archi** le relazioni fra concetti o proprietà dei concetti stessi

#### ▼ Grafi relazionali

La versione più semplice è quella dei **grafi relazionali** in cui vengono descritte le relazioni fra

le entità del grafo. Una delle relazioni più importanti è quella di tipo **isA**.

Questi grafi sono adatti a rappresentare i problemi semplici (mondo dei blocchi), ma hanno dei limiti

- nell'espressività tutti gli archi di un nodo sono infatti visibili come in AND fra di loro, ma manca un formalismo per rappresentare la disgiunzione, l'implicazione, i quantificatori ecc..
- quantificazione è più difficile da rappresentare in questa struttura
- l'arietà limitata a 2, perchè con questo tipo di grafi si possono fare solo relazioni binarie, per averne una n-aria bisogna introdurre dei nodi aggiuntivi, questo porta però ad una crescita delle dimensioni e della complessità del grafo relazionale.

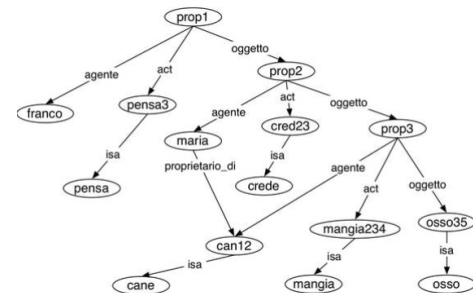
#### ▼ Reti proposizionali

Tramite i grafi relazionali risulta difficile rappresentare la frase "*Franco pensa che Maria creda che il suo cane stia mangiando un osso*". Ricorriamo quindi alle **reti proposizionali**, che sono strutture più complesse in grado di rappresentare intere



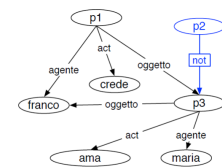
proposizioni. L'introduzione di nodi proposizionali ha portato alla generazione di tipi impliciti.

**Esempio** “pensa3” è di tipo isA. Il nodo prop1 indica che Franco pensa qualcosa, ecc..



La possibilità di avere nodi proporzionali aumenta l'espressività del linguaggio. Infatti è possibile introdurre i connettivi logici e dei contesti all'interno dei quali far operare i quantificatori.

La negazione può essere rappresentata tramite un arco che collega il risultato della negazione con la proposizione negata. Essendoci la negazione possiamo anche rappresentare la disgiunzione grazie all'uso delle leggi di De Morgan, trasformando  $A \vee B$  in  $\neg(\neg A \wedge \neg B)$  e applicando di conseguenza la relazione not.



Franco crede che Maria lo ami, ma non è vero.

Una caratteristica importante è la rappresentazione di **conoscenze gerarchiche** infatti molte entità possono essere raggruppate in classi, che a loro volta possono essere raggruppate in superclassi di ordine sempre più elevato

**Esempio** Vogliamo rappresentare che gli elefanti sono mammiferi. Si utilizza un nodo per gli elefanti e uno per i mammiferi, collegati da un arco con una relazione **isA**. Se volessimo esprimere il fatto che Clyde è un elefante, basta aggiungere il nodo Clyde e collegarlo mediante una relazione isA



Ci sono due approcci per le relazioni gerarchiche

- **Relazione di copertura** si rappresentano solo i legami essenziali, evitando di rappresentare i legami che possono essere inferiti navigando il grafo. Il vantaggio di questo approccio è un utilizzo efficiente della memoria
- **Chiusura transitiva** si rappresentano tutti i possibili legami isA fra i vari nodi della rete  
Il vantaggio di questo approccio è un utilizzo efficiente del tempo, poiché per stabilire se una relazione isA basta andare a vedere se è stata definita.

Nelle reti semantiche manca la semantica formale ma esiste una **semantica procedurale** dove bisogna presupporre l'esistenza di procedure in grado di utilizzare le informazioni in esse contenute, quindi il significato è attribuito a da chi accede alla rete stessa.

## Frame

Si tratta di un formalismo con aspetti in comune alle reti semantiche.

Il concetto fondamentale è che le persone usano un insieme strutturato di conoscenze pregresse (frame), rappresentanti genericamente una situazione o problema.

**Esempio** frame del ristorante: entrando ci si aspetta di trovare dei tavoli, di essere raggiunti da un cameriere, ecc. Questo porta a interpretare i nuovi fatti all'interno di un contesto, aiutando a ridurre le ambiguità e facilitando il recupero delle informazioni.

Non esiste un formalismo per la rappresentazione dei frame, ogni frame è identificato univocamente da un nome e le sue proprietà sono rappresentate mediante slot. I valori degli slot possono essere noti o meno e quando non lo sono, è comunque a volte possibile fare delle supposizioni accettabili sul valore fino a prova contraria. Tuttavia è possibile dire che esiste sempre una struttura gerarchica a tre livelli

- **di base**, una rappresentazione degli oggetti/entità di cui è formato il nostro mondo
- **super-ordinato**, è costituito da una generalizzazione di quello di base
- **subordinato**, specializzazione del livello base

**Esempio** *sedia* potrebbe essere un concetto di base, *mobile* il concetto superordinato e *sedia a dondolo* il concetto subordinato

Una caratteristica importante dei frame è che l'appartenenza ad una categoria non è rappresentata da un insieme di attributi ma dalla (maggiore o minore) somiglianza rispetto a un **prototipo** (elemento esemplare di una categoria)

**Esempio** un passero è un buon prototipo di uccello, mentre un pinguino lo è meno

Questo significa che i sistemi a frame permettono di ragionare intorno a classi di oggetti utilizzando appunto i prototipi (quindi un po' come se fosse un centroide) che devono essere adattati per tener conto della complessità del mondo.

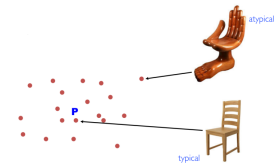
## ▼ Teorie e approcci alla KR

Esiste la

### ▼ Teoria dei prototipi e degli esemplari

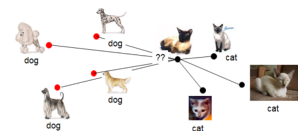
appartiene agli sviluppi della teoria mentalista. Il prototipo è un “*elemento esemplare*” di una categoria

- dei **prototipi** è più flessibile, afferma che un prototipo è un'approssimazione (astrazione) rappresentativa di una categoria.



**Esempio** posso rappresentare cani a 3 zampe, che non abbaiano, ecc... (quindi più tipi di cani)

- degli esemplari** è meno flessibile, afferma che la rappresentazione di un concetto è l'insieme delle rappresentazioni dei prototipi di quella categoria che sono stati incontrati durante la vita.



**Esempio** se dico che per rappresentare un cane, questo dovrà avere per forza 4 zampe, allora scarterò i cani con 3 zampie

Si tratta in entrambi i casi di modelli in cui si va a categorizzare un elemento sconosciuto sulla base di confronti rispetto a quanto presente nella nostra conoscenza, scegliendo la categoria a cui l'elemento più si avvicina. (cioè quanto abbiamo visto con i frame).

### ▼ Teoria referenziale del significato

le parole sono lo strumento attraverso il quale facciamo riferimento a ciò che esiste.

**Esempio** *sedia* è un riferimento a un oggetto; *acquistare* è un riferimento a un evento.

### ▼ Teoria mentalista o concettuale

arricchisce la precedente teoria con l'ipotesi che il riferimento fra parole e realtà non è diretto, ma mediato dall'immagine mentale che costruiamo di queste entità.

**Esempio** quando parliamo di “*sedia*” non facciamo riferimento direttamente alla nozione di “*sedia*”, ma alla rappresentazione mentale che abbiamo della classe di oggetti che rientrano nel tipo “*sedia*”

### ▼ Teoria strutturale -> synset

il significato dei termini non consiste nel suo riferirsi per esempio a un oggetto (ipotesi referenziale), o nel rimando all'immagine mentale (ipotesi mentalista), ma nel *valore* che la parola assume in relazione alle altre parole presenti nella lingua che fanno parte dello stesso campo semantico e designano oggetti analoghi.

**Esempio** nel caso di *sedia*: sgabello, sedile, trono, poltrona, panca, ecc..

### ▼ Teoria distribuzionale

il significato delle parole è determinato in larga misura dell'insieme di altre parole con cui queste co-occorrono. Due parole sono tanto più simili quanto più sono simili i contesti in cui queste occorrono.

### ▼ Teoria dei processi duali

teoria per capire come combiniamo l'approccio a ontologie con quello a prototipi/esemplari. Stabilisce una nuova divisione nei processi cognitivi

- **System 1**

relativo a ciò che è implicito i meccanismi che scattano istintivamente, quindi l'insieme delle capacità che hanno permesso

- **System 2**

relativo a ciò che è esplicito e a tutti i meccanismi sequenziali, deliberativi. è quello che serve per tirare fuori formalmente quello che

lo sviluppo evolutivo  
dell'uomo (è un processo inconscio)

compiamo, è un processo più  
conscio/consapevole

System 1 (Implicit)	System 2 (Explicit)
Unconscious	Conscious
Automatic	Controllable
Evolved early	Evolved late
Language Independent	Related to Language
Parallel, Fast	Sequential, Slow
Pragmatic/contextualized	Logical/Abstract

È possibile costruire sistemi che uniscano System 1 e System 2 in modo che si completino a vicenda

---

## ▼ WordNet

Si tratta di un sistema di riferimenti lessicali online basato sul concetto di synonym set (**synset**), ciascuno dei quali rappresenta un concetto lessicale. Il suo compito è quello di associare dei possibili sensi ad un termine.

### Struttura

Ogni sezione è organizzata in **synset**, che sono gruppi di termini lessicali che rappresentano lo stesso concetto.

WordNet è organizzato in base alle relazioni semantiche, dove i nodi sono l'insieme dei sinonimi e le relazioni sono gli archi che collegano i nodi.

Il principio su cui si fonda WordNet è la **sinonimia**.

WordNet suddivide il lessico in quattro categorie, ciascuna delle quali rappresentata con una struttura adeguata:

- **sostantivi** sono organizzati come gerarchie
- **verbi** sono organizzati in base alle relazioni che hanno
- **aggettivi e avverbi** sono organizzati come degli iperspazi n-dimensionali

### Matrice lessicale (Termine vs Senso)

Un concetto è composto da una veste lessicale (termine) e da uno o più significati (sensi).

Quindi la semantica lessicale distingue il

- **word meaning** il termine/concetto in forma lemmatizzata
- **word form (utterance)** come la parola può comparire in tutte le sue forme

Il nostro obiettivo è di rappresentare termini e sensi per studiare le loro relazioni, e lo facciamo con la **matrice lessicale** dove sulle righe abbiamo i sensi e sulle colonne i termini.

Word Meanings	Word Forms				
	F <sub>1</sub>	F <sub>2</sub>	F <sub>3</sub>	. . .	F <sub>n</sub>
M <sub>1</sub>	E <sub>1,1</sub>	E <sub>1,2</sub>			
M <sub>2</sub>		E <sub>2,2</sub>			
M <sub>3</sub>			E <sub>3,3</sub>		
⋮				⋮	
M <sub>m</sub>					E <sub>m,n</sub>

Quindi, se due entry hanno la stessa colonna, ma riga diversa, la loro forma è **polisemica**. Se invece due entry hanno la stessa riga, ma una colonna diversa le loro forme sono **sinonimi**

## Cos'è il synset

Analizziamo com'è fatto un singolo nodo: un synset che è un insieme di termini sinonimi.

I synset possono essere utilizzati per fare word sense disambiguation.

**Esempio** il termine *board* può significare “*tavola di legno*”, ma anche “*insieme di persone riunite per uno scopo*”. Tale termine apparirà in almeno due synset: {*board, plank, ...*} per il primo significato e {*board, committee, ...*} per il secondo.

Se disponiamo di un contesto possiamo confrontare le parole del contesto con quelle dei due synset per disambiguare il significato.

Quindi i synset non spiegano cosa sia un concetto, ma ne identificano i sinonimi.

I synset includono anche una **gloss** (commento/descrizione) per descrivere meglio il concetto e, una o più brevi frasi d'esempio.

## ▼ Relazioni semantiche

- **Sinonimia** è la relazione più importante, ed è definita come la capacità di una parola di essere sostituita ad un'altra, in un certo contesto, senza alterare il valore di verità della frase.



Dato che sinonimi puri sono rari rilassiamo la relazione vincolandoli al contesto della frase e diciamo che una parola *x* è sinonimo di una parola *y* se, all'interno di un contesto *C*, in una frase dove appare *x* è possibile sostituire *y* pur mantenendo il senso d'origine della frase

**Esempio** nel contesto della carpenteria sostituire *plank* con *board* non ne altera il valore di verità

- **Antonimia** è la relazione di significati opposti fra parole. Risulta molto difficile da definire, perché non sempre equivale a una negazione netta

**Esempio** *ricco* e *povero* sono antonimi, ma non essere ricco non equivale per forza ad essere povero (posso essere benestante)

- **Iponimia** dà vita a una struttura gerarchica, dove il subordinato eredita tutte le caratteristiche del superordinato aggiungendone di proprie.

**Esempio**, acero è un iponimo di albero

- **Meronymia** possiamo vedere la meronymia come una relazione part-of, e di solito se una cosa è meronimo di un'altra allora quell'altra è olonimo della prima

**Esempio** dito è *meronimo* di mano, in quanto un dito è parte della mano e mano è *olonimo* di dito

## Nomi in WordNet

### ▼ Organizzazione dei nomi in WordNet

Al top della struttura di WordNet sono presenti 25 categorie molto generali, dette **primitive o supersensi**. Dove ci sono queste categorie viene detto **basic-level** e al di sopra si avranno descrizioni più corte e generiche, mentre al di sotto le parole avranno caratteristiche aggiuntive che aggiungono meno al concetto basilare.

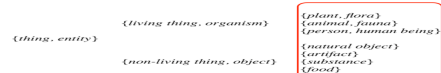
25 unique beginners (nouns)

{act, action, activity}	{natural object}
{animal, fauna}	{natural phenomenon}
{artifact}	{person, human being}
{attribute, property}	{plant, flora}
{body, corpus}	{possession}
{cognition, knowledge}	{process}
{communication}	{quantity, amount}
{event, happening}	{relation}
{feeling, emotion}	{shape}
{food}	{state, condition}
{group, collection}	{substance}
{location, place}	{time}
{motive}	

Quest 25 categorie sono stati definite così da avere, per ogni nome, un iperonimo che lo rappresenti, quindi le parole ereditano le caratteristiche dei supersensi.

Le gerarchie che si dipartono da questi 25 supersensi sono mutuamente esclusive, anche se a volte ci sono dei cross-references .

Ad esempio, 7 dei 25 superconcetti fungono da specificazione (iponimi) per gli esseri viventi e gli essere non viventi, e quindi sono gerarchizzati a loro volta



I termini (nomi) in WordNet quindi sono rappresentati in una struttura gerarchica sulla base delle relazioni di iponimia, in cui ogni termine ha

- **un suo superiore** (padre / superordinato) dal quale ne eridita le caratteristiche
- **e poi aggiunge delle caratteristiche proprie** che contraddistinguono il termine stesso ovvero
  - **attributi /modifieri** espressi come *aggettivi*, e sono dei possibili modificatori di quel termine. Alcuni aggettivi non possono essere intesi come modificatori effettivi, ma come modificatori metaforici

**Esempio** Un canarino generoso può essere interpretato solo metaforicamente in quanto la generosità non è una feature dei canarini

- **parti** sono rappresentate tramite relazioni di meronimia che si possono ereditare.

**Esempio** becco e ala sono parti di bird. Il canarino, essendo iponimo di bird avrà come parti anche becco e ala.

La meronimia è asimmetrica e transitiva, anche se a volte la transitività può essere forzata: è vero che una porta ha una maniglia e una casa ha una porta, ma è strano sostenere che una casa abbia una maniglia.

- **funzioni** (verbi) una caratteristica funzionale di un sostantivo, cioè è la descrizione di qualcosa che svolge o si può svolgere con/su di esso. Un



oggetto che non è un x, se viene definito come un buon x allora svolge bene le funzionalità che x normalmente svolge.

**Esempio** una persona che si siede su una scatola può dire che la scatola è una buona sedia, indicando che la scatola svolge bene la funzione che ci si aspetta che svolgano che sedie, non le scatole

**Esempio** il concetto *canary* avrà:

- **attributes** *yellow, small*
- **parts** *beak, wings*
- **functions** *sing, fly*

## Verbi in WordNet

I verbi sono per molti linguaggi una componente imprescindibile di una frase perchè aiutano a definire il contesto della frase.

Ci sono meno verbi rispetto ai nomi e i verbi sono più polisemici quindi questo significa che i verbi possono cambiare il loro significato a seconda dei nomi con i quali co-occorrono, mentre i nomi tendono ad avere lo stesso significato anche con verbi differenti

**Esempio** *correre in un campo vs correre un rischio*

Anche i verbi, in WordNet, sono suddivisi in macro-categorie rappresentanti domini semantici

diversi. Le distinzioni semantiche fra verbi non sono però le stesse dei nomi

Infatti non si può parlare di iponimia fra verbi, ma serve usare una relazione diversa. Essa prende il nome di **troponimia** e descrive il fatto che un il verbo Y è un troponimo del verbo X se nel fare l'attività Y si fa anche la X

**Esempio** *passeggiare e camminare* non fa funzionare l'iponimia perchè "*passeggiare è un tipo di camminare*" non ha molto senso come concetto.. mentre c'è una troponimia tra *mormorare* rispetto a *parlare* (mentre mormoro sto parlando)

## Wordnet vs Dizionario



WordNet non è un dizionario perchè un dizionario standard non può ricercare significati in quanto i termini sono organizzati in ordine lessicografico; noi vogliamo organizzare il lessico in gruppi di significati in modo da ottenere da un certo termine il suo significato e dal suo significato tutti i significati correlati (complesso da realizzare con un dizionario)

Infatti nei dizionari

- **la specifica non è totale** non tutte le caratteristiche di un albero sono specificate (non viene detto che un albero ha le radici)
- **nessuna informazione sui termini coordinati** esistono tanti tipi di alberi, ma il dizionario non ci dà alcuna informazione. Dovremmo scorrere l'intero dizionario per trovare questo tipo di informazione.
- **rischio di ricorrenze infinite** un certo nome può essere definito in termini di un altro nome, il quale a sua volta potrebbe essere definito con il nome iniziale. La computazione di tali definizioni porterebbe ad una ricorsione infinita (praticamente sarebbe il meme degli spiderman che si guardano)

Tutte queste problematiche vengono risolte da WordNet che ha una funzione totalmente diversa rispetto ad un semplice dizionario

## Assunzioni psicolinguistiche

WordNet è ispirato da teorie psicolinguistiche, infatti i nomi sono stati organizzati seguendo un sistema gerarchico, quindi un modello lessicale adottato dalla mente umana. Le evidenze psicolinguistiche che ci indicano la presenza di tale struttura usano come prova il modo col quale siano in grado di gestire le anafore. L'anafora è una figura retorica che consiste nel ripetere una o più parole all'inizio di più frasi.

**Esempio** *“Possedeva una pistola ma l'arma non aveva sparato”,*  
l'*arma* è un'anafora per riferirsi alla *pistola*, quindi con gli iponimi possiamo capirlo

## Evidenze da esperimenti

È stato osservato che i tempi di risposta (tempo impiegato da un essere umano per dire che la frase è vera) sono minori per la frase *“a canary sing”* rispetto a *“a canary can fly”*. Ancora più tempo è necessario per rispondere alla frase *“a canary has skin”*.

Grazie a questo esempio possiamo assumere che all'interno della mente dell'umano “sing” è memorizzato come features dei canary, precisamente come function; *can fly* come feature di “bird” e “has skin” come feature di animal.

Se tutte le informazioni relative a canary fossero conservate insieme al nome, i tempi di reazione sarebbero gli stessi per tutte e tre le frasi, ma poiché le evidenze mostrano che i tempi sono diversi, possiamo immaginare la presenza di una struttura gerarchica che recupera informazioni aggiuntive su un concetto soltanto quando necessario (quindi prediligiamo prima lo specifico per poi scalare sul generale nella gerarchia)

## ▼ FrameNet

FrameNet consiste in un ulteriore passo in avanti rispetto a WordNet, in quanto permette di assegnare il senso non più soltanto alle singole parole, ma a intere frasi.

Abbiamo già menzionato i frame per la rappresentazione della conoscenza per cui l'ipotesi è quindi che nel nostro modo di caratterizzare i concetti ci basiamo largamente sulla nostra esperienza personale. I blocchi di informazioni alla base di questa ipotesi sono appunto i **frame**, ovvero un quadro di riferimento per un certo contesto

**Esempio** se vado ad un nuovo ristorante dato che conosco come funziona un ristorante classico allora so già qualcosa di quel nuovo ristorante.

## Cos'è e a cosa serve

è un database lessicale basato sui **frame semantici**. Il compito di FrameNet è quello di definire i frame, riempirli con i termini che ben si incastrano e descrivere il concetto espresso dal frame. Per cui si tenta di mettere insieme due tipi di elementi: **termini e significati**.

Le attività di framenet sono

- caratterizzare i frame capendo quali sono gli elementi che li caratterizzano
- una volta definiti i ruoli e le funzioni principali all'interno di un frame si cercano le parole che li rappresentano meglio
- per ogni frame vengono costruite delle frasi di esempio che devono essere rappresentative dei termini che abbiamo individuato nei passi precedenti

- dopo avere ottenuto un numero sufficiente di esempi li annotiamo (attività fatta dagli esperti)

## WordNet vs FrameNet

In WordNet dato un termine polisemico e un contesto in cui questo occorre ci chiediamo qual è il senso di quel termine in quel contesto. In FrameNet invece il task è diverso, perchè noi abbiamo un frame che descrive uno scenario stereotipato e si cerca di capire come gli elementi della frase che stiamo analizzando possono essere inseriti nel frame in questione.

## Cos'è un Frame

è una descrizione astratta, quindi è una stereotipizzazione di cose, stati o eventi.

I frame sono una struttura dati costituita da slot, dove uno slot può essere a sua volta un frame. Inoltre i frame possono essere organizzati secondo una gerarchia.

Gli elementi che utilizziamo per descrivere un frame sono relativi al determinato frame che trattiamo, ci sono alcuni termini, che sono evocativi di frame, cioè sono in grado di rimandare direttamente ad un certo frame (si dice che queste parole evocano il frame).

**Esempio** “cameriere” ci potrebbe ricordare il frame di un ristorante

Quindi l'evocazione di un frame può essere vista come l'attivazione di una regola: ad una certa parola (fatto) corrisponde l'attivazione di un certo frame (regola).

### ▼ Nel caso di termini ambigui?

I termini polisemici risultano problematici, in quanto possono attivare diversi frame. Ciò è dovuto all'ambiguità del linguaggio. Per cui invece di lavorare con le parole lavoriamo con le **unità lessicali** (lexicall unit - LU) che sono una coppia {parola e suo significato}.

In WN sensi diversi appartengono a synset diversi mentre in FM sensi diversi appartengono a LU diversi.

## Particolarità che possiamo cogliere in FN

### ▼ I verbi

I verbi causano alcuni problemi, tale problema viene gestito facendo un'analisi sintattica e a partire dal verbo si applicano dei pattern per cercare di comprendere in quale situazione ci troviamo. In pratica sfruttiamo i Frame per capire il contesto corrente e quindi il contesto d'uso del verbo.

**Esempio**

- “*She earns a lot less than she deserves*”
- “*I made a lot of money, but I earned it*”

Il verbo **earn** ha due significati diversi, anche se i contesti sono molto simili. Nella prima frase ci riferiamo ad un vero e proprio guadagno, mentre nella seconda frase ci riferiamo ad un guadagno come merito. FrameNet ci aiuta a risolvere questa ambiguità

### ▼ Subject as..

In genere il soggetto di una frase che compie un’azione è la persona/oggetto che propriamente compie l’azione. In questo caso si parla di **Subject as Speaker**. Tuttavia ci sono dei casi in cui il soggetto non compie l’azione indicata dal verbo, ma è soltanto metaforico, in questi casi si parla di **Subject as Medium**. Nel caso di Subject as Medium si fa uso di metonimie e quindi si può tradurre la frase in una frase di più semplice interpretazione. Anche in questo caso FN ci può aiutare in questo processo di traduzione.

### ▼ Nominalizzazione

ossia quando il verbo cambia significato a seconda del sostantivo con cui appare.

**Esempio** “*aderire alla pelle*” (aderenza) o “*aderire ad un partito*” (fedeltà/adesione);

Anche in questo caso, FN ci aiuta a comprendere il significato di un verbo.

## Esempio (Frame Vendetta)

Come si costruisce questo frame

- si inizia con il definire un **vocabolario del frame**, cioè un elenco dei termini (nomi, verbi, locuzioni, ecc..) rilevanti che possono far parte di questo topic
- Dopodichè si fornisce un **frame definition** [FD] (usabile come le gloss) e una lista di **frame elements** [FE] (servono per etichettare le parti di una frase che tratta di quel particolare frame, devono essere molto precisi e meno generici possibile, perchè così almeno definiamo un frame il più possibile aderente alla situazione che vogliamo descrivere)

Nel nostro caso, il frame può contenere questi FE {avenger; offender; injury, ecc..}

- successivamente si procede nella costruzione di un **corpus di frasi di esempio** in cui sono presenti i FE del frame
- ora si parte con l'**annotazione** delle frasi, quindi per esempio si prendono tutte le frasi in cui occorre il termine *avenge* e da qui abbiamo due tipi di target
  - **predicati**, che sono parole che invocano frame
  - **fillers** che sono parole che soddisfano ruoli semantici di frame

| **Esempio** Chi il vendicatore nella frase?

## Dov'è utile usare FN?

È possibile sfruttare FN per ottenere informazioni di diverso tipo

- un primo esempio è la richiesta di come sintatticamente venga realizzato il ruolo del FE "*offender*"
- è anche possibile interrogare FN per ottenere tutte le parole associate a un dato frame o i pattern sintattici in cui ogni parola è usata per esprimere il frame.

## Quali sono le sue applicazioni?

FN è una risorsa lessicale che mira a fornire un corpus di frasi annotate, in questo modo è possibile estrarre informazioni sui modi in cui i termini vengono utilizzati. Lo scopo di FrameNet è di fornire un inventario di situazioni ed eventi per abilitare la comprensione di un testo mediante tecniche di elaborazione del linguaggio naturale che siano in grado di associare automaticamente uno o più frame ad ogni frase del testo stesso, per cui in sintesi le applicazioni di FN sono

- **WSD** tramite ricerca diretta di lexical unit e guardando a che frame ognuno appartiene
- information extraction
- **comprensione del testo**

## ▼ Vediamo un caso d'uso di applicazione di FN

Un uso tipico di FN è quello di fare **Information Extraction** per la comprensione del testo.

Abbiamo preso l'esempio la sezione di cronaca nera di un giornale, dove

- vengono selezionati i termini che appaiono più frequentemente
- poi vogliamo risolvere le disambiguazioni dei termini (WSD), i riferimenti anaforici
- a questo punto l'obiettivo è, associare un certo numero di frame alle frasi, ma a priori non possiamo sapere di quanti frame consti la situazione descritta in un articolo, quindi proviamo ad associare un frame diverso ad ogni frase.

- Vediamo nelle frasi le parole (a partire da quelle più rilevanti) per capire quali frame possono essere attivati da quelle parole.
- L'analisi non può essere limitata ad un approccio parola per parola, in quanto entità come *White House*, *the Secret Service*, ecc. devono essere considerate in toto. (visti come sintagmi)

## Conclusione

Per concludere FrameNet è una risorsa bipartita

- da una parte c'è un **insieme di Frame** che descrivono situazioni stereotipate
- dall'altra parte c'è un **corpus** annotato che ci permette di stabilire dei mezzi tra elementi che sono stati annotati dagli esperti in vari livelli (annotazione morfologica, sintattica e semantica) che ci permette di fare delle analisi su quali componenti realizzano alcuni ruoli semantici abbiamo visto come il verbo *avere* fa giocare un ruolo semantico diverso nelle frasi

## ▼ ConceptNet

Qui introduciamo un altro tassello ossia: il **senso comune**; infatti nessuno strumento visto fino ad ora si è posto il problema di rappresentare il common sense. WordNet parla di significati, BabelNet è un WordNet multilingua potenziato e FrameNet descrive situazioni tipiche grazie alle quali individuare un contesto

## Cos'è il senso comune?

è l'insieme dei fatti e delle conoscenze di base possedute dalla maggior parte della popolazione, come ad esempio il fatto che per aprire una porta bisogna girare la maniglia. Questa conoscenza è di vario tipo (spaziale, temporale, psicologico, ecc...).

La difficoltà nel definire il senso comune sta nel fatto che non è mai espresso esplicitamente, ma si dà per scontata la sua presenza in ognuno di noi. Invece per le macchine è sempre stato difficile possederla e utilizzarla. Infatti per comprendere un testo abbiamo bisogno di una grande dose di common sense, che ogni persona possiede a differenza delle macchine

Nessuno strumento visto fino ad ora si è posto il problema di rappresentare il common sense. **ConceptNet** nasce appunto con l'idea di fornirlo alle macchine.

## Cyc - un primo approccio

Un primo approccio alla rappresentazione del commonsense è stato Cyc, un tentativo di organizzare la conoscenza comune in un framework logico. Nonostante Cyc sia molto ricca, dal punto di vista terminologico, il suo approccio si è rivelato fallimentare poiché per

poter utilizzare Cyc è necessario prima mappare, il testo in input, ad una struttura logica (linguaggio CycL). Tale operazione è complessa e poco precisa a causa dell'ambiguità del linguaggio

## Cos'è ConceptNet?

è una rete semantica che contiene molte cose che i computer dovrebbero conoscere del mondo,  
specialmente quando si tratta di comprendere testi scritti da umani.



Lo scopo di ConceptNet è quello di cablare informazioni come “*per aprire una porta devi girare la maniglia*” o come “*se dimentichi il compleanno di qualcuno, questo potrebbe essere arrabbiato con te*”.

Il task non è semplice!

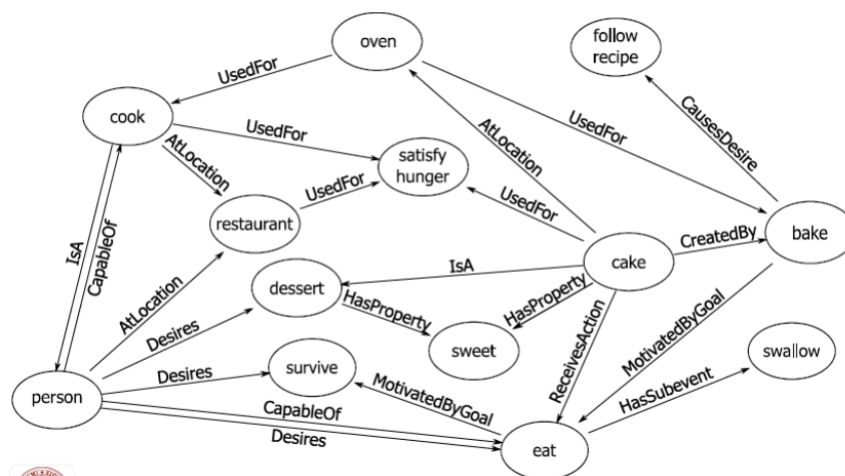
## Struttura di CN

La struttura interna di ConceptNet è un grafo costituito da più di 300 mila nodi messi in relazione tra loro, da 1.6 milioni di archi, al fine di rappresentare il senso comune. I nodi sono frammenti frasi (o frammenti di esse) e gli archi rappresentano una fra 20 relazioni semantiche possibili che un nodo può assumere.



I nodi di ConceptNet sono anche chiamati **concetti**, tuttavia non sono da confondere con i concetti visto fino ad ora, (visto che si tratta di frammenti di testo), infatti qui si supera quindi la definizione di nodo di WordNet come elemento puramente lessicale per permettere di rappresentare più concetti incontrati nella vita quotidiana (nessi temporali, spaziali, di capacità, ecc..).

Così facendo si perde però l'annotazione del senso delle parole.





La prima osservazione che possiamo fare è che nell'immagine i nodi sono delle *entità particolari* tipo “*satisfy hungry*”, cioè non è una primitiva sto nodo; e poi non c'è più una gerarchia (non esiste un nodo padre di un nodo figlio) questi sono detti **compound concept**: elementi come *comprare cibo*, *guidare fino al supermercato* sono nodi che rappresentano eventi di vita quotidiana.

Quindi si viene a creare una conoscenza molto più informale, di natura pratica, che viene detta conoscenza **rivedibile** ossia che è possibile esprimere qualcosa di vero in molti casi, ma non sempre

**Esempio** un effetto della caduta dalla bici è farsi male: spesso è vera, ma in alcuni casi no.

## Applicazioni di ConceptNet

ConceptNet è particolarmente adeguato al ragionamento su di un **contesto**, per cui è ottimo per

- Data una storia che descrive una serie di eventi, da dove prendono luogo questi eventi? Qual è il mood della storia e quali sono i prossimi eventi plausibili? (proiezione spaziale, affettiva e temporale)
- Data una query di ricerca dove uno dei termini può avere significati multipli, qual è il significato più probabile? (disambiguazione del contesto)
- Se si ripresenta un concetto all'interno di una storia, quale dei concetti già conosciuti si approssima più fedelmente a quello appena trovato (analogie)

Questo perché la maggior parte dei suoi collegamenti fra concetti sono piuttosto generici (detti **k-lines [knowledge lines]**), che permettono di aumentare la connessione della rete e quindi permettono di comprendere le collocazioni spaziali / temporali di eventi, fino ad arrivare al riconoscere forme di ironia, sarcasmo, ecc..



Per quanto riguarda la WSD è necessaria una precisazione: visto che i nodi di ConceptNet sono stringhe è impossibile effettuare una *pura* WSD in quanto i significati non sono presenti in CN. Ciò che è possibile fare è esplorare le adiacenze di un certo nodo che fa match con parte di un testo fornito in input e sfruttare le relazioni trovate per dedurre informazioni in merito alla query. Infatti i frammenti di testo hanno un senso univoco e quindi non esiste alcun problema di ambiguità.

**Esempio** *play guitar* non è ambiguo rispetto a *play tennis*; *play* invece sarebbe ambiguo.

Il problema dell'ambiguità quindi riguarda termini singoli e non locuzioni.

## Relazioni semantiche

In ConceptNet abbiamo 20 relazioni semantiche (ad oggi 40) tra cui

- causality (EffectOf)
- gerarchia di eventi (SubeventOf)
- abilità di un agente (AbilityOf)
- affect (PropertyOf, LocationOf, MotivationOf), ecc..

## Costruzione di CN

La costruzione di questa risorsa parte con il sito per l'Open Mind Common Sense (OMCS), dove a volontari viene chiesto di rappresentare vari tipi di common sense attraverso attività come la compilazione di spazi vuoti in frasi del tipo "*A knife is used to \_*". Vengono così raccolte frasi che rappresentano la conoscenza comune.

Da queste frasi il processo di costruzione si articola in due fasi

- **Estrazione** dove vengono estratte (con regex) delle asserzioni attraverso un insieme di regole
- **Rilassamento** per migliorare la connessione della rete e diminuire i buchi semantici.

In questa fase

- si **uniscono** le asserzioni duplicate
- si fa **lifting** delle informazioni a partire dalla relazione isA

**Esempio** se la banana è un frutto e il frutto è dolce allora la banana è dolce.

- si creano anche **generalizzazioni** lessicali

**Esempio** se la maggior parte dei frutti hanno la proprietà *sweet*, allora tale proprietà viene **liftata**, ossia passata alla classe padre come *propertyOf(fruit, sweet)*.

- si fa **riconciliazione** delle discrepanze morfologiche

**Esempio** *bici* viene collegata *abbicicletta*

## Valutazione

Dare un giudizio sul commonsense è generalmente difficile, perché bisogna prima stabilire cosa sia e cosa non sia conoscenza comune.

- un metodo per avere un'idea generale della bontà è la verifica di come i concetti contenuti nella rete siano espressi: se sono brevi, ovvero formati da poche parole, sono ritenuti facili. In CN circa il 70% dei nodi ha una lunghezza al massimo di tre parole. Confrontando questo risultato con il fatto che un sintagma verbo-nome richiede almeno quattro parole ("*take a dog for walk*"), si capisce che la maggior parte dei nodi di CommonNet sono molto semplici.
- Un altro metro di valutazione è la ripetizione delle asserzioni: dato che sono semplici, è legittimo aspettarsi che capitino di frequente. È dimostrabile però che la maggior parte dei concetti di CN appaiano una sola volta nelle frasi di OMCS e un'altra grande percentuale non compaia proprio e sia stata quindi inferita (sono k-lines), dimostrando ancora la bontà di questa rete semantica.

## ▼ Cover

Cover è una risorsa linguistica sviluppata ad UniTO, che tratta dati estratti da ConceptNet, NASARI e BabelNet. Lo scopo dell'algoritmo è di generare in modo automatico una rappresentazione vettoriale del senso comune per un dato concetto in input.

## Funzionamento

L'input quindi è un concetto (**esempio** bn:00035902n), poi in ordine eseguiremo

- estrazione semantica che consiste in due fasi
  1. dove individuiamo (all'interno di ConceptNet) dei termini rilevanti che sono legati al nostro concetto dato in input

2. da questi termini usiamo NASARI per capire quanto sia importante il singolo termine in relazione al concetto scelto in questo modo
  - a.  $t$  è un termine di un synset di uno degli elementi contenuti nel vettore NASARI di input
  - b. se almeno 3 elementi direttamente collegati a  $t$  sono elementi presenti all'interno del vettore NASARI
- a questo punto abbiamo ottenuto un insieme di termini rilevanti  $T$  e ora dobbiamo passare ad un insieme di concetti rilevanti  $C$  che saranno estratti da BabelNet. Qui è importante capire come sono stati ottenuti i termini al passo precedente
  - se il termine era presente direttamente nel vettore NASARI (*condizione a*), allora in concetto sarà direttamente presente nell'insieme dei concetti.
  - se il termine era stato inserito tra i termini rilevanti mediante la *condizione b*, allora recuperiamo tutti i possibili BabelNet synset per quel termine (e poi quindi i loro vettori NASARI) e calcoliamo la similarità di ognuno di essi con il vettore NASARI di input.  
Se il valore di similarità del senso più simile supera una certa soglia, allora il senso è stato trovato.
- Alla fine si inseriscono i concetti dell'insieme  $C$  all'interno di un vettore, che ha una caratteristica: è un elenco di liste dove ciascuna contiene i termini che sono legati ad una certa relazione

Example of the 'fork' vector

<i>relatedTo</i>	<i>tool, food, utensil, cutlery, eating</i>
<i>isA</i>	<i>tool, cutlery, utensil</i>
<i>atLocation</i>	<i>table, desk, plate</i>
<i>usedFor</i>	<i>eating</i>
<i>madeOf</i>	<i>metal</i>

la forchetta è associata a certi termini rilevanti che sono aggregati in certe relazioni

## Concept similarity

In Cover abbiamo un modo alternativo per calcolare la similarità tra concetti, infatti il concetto di similarità deriva da quanti elementi in comune hanno quei due vettori.

**Esempio** *Bird* e *Cock* sono *relatedTo* a 44 elementi il primo e 6 l'altro e ora cerchiamo i termini in comune, quindi ne troviamo 4 su 6 e

applichiamo il rapporto simmetrico di Tversky così da assegnare un punteggio a quella feature. Alla fine si fa una media complessiva tra le feature trovate e otteniamo un indice di correlazione pari a 0.63, che è ottimo, considerano lo score di un umano.

Dimension	Dim. Score	V1   V2 count	Shared values
<i>relatedTo</i>	0.68	44   06	<i>feather, chicken, roosting, vertebrate</i>
<i>isA</i>	0.58	07   04	<i>animal</i>

## ▼ BabelNet

BabelNet nasce dalla fusione fra WordNet e Wikipedia.

Le risorse WordNet sono poco sviluppate per le altre lingue. Per questo vogliamo sfruttare le informazioni contenute in Wikipedia per ottenere una risorsa simile a quella restituita da WordNet. La fusione dei due progetti quindi porta ad avere una rete semantica molto estesa e allo stesso tempo dettagliata in un insieme di lingue molto ampio.

### ▼ Comè formata Wikipedia?

Wikipedia è una enciclopedia multilingua online. È un'opera collaborativa opensource creata e corretta da volontari di tutto il mondo al fine di ottenere un'enciclopedia di largo spettro

Il punto forte di Wikipedia è la sua ricchezza semantica: è formata da pagine riferite a un concetto o una **named entity** (l'azienda Apple, per esempio). Il titolo di ogni pagina ha opzionalmente un'etichetta fra parentesi che specifica il significato del lemma, se ambiguo (**esempio** Play (theatre)). Certe pagine hanno una infobox che rappresenta un riassunto degli attributi principali dell'entità e sono memorizzate tramite grafi RDF. Esistono poi relazioni fra pagine di vario tipo:

- **pagine di redirect** inoltrano a pagine che contengono effettivamente le informazioni sul concetto, modellando quindi la sinonimia
- **pagine di disambiguazione** raggruppano link alle pagine per i possibili concetti rappresentati da un termine, modellando quindi la polisemia
- **link interni** collegano altre pagine di concetti correlati a una o più pagine
- **link inter-lingua** collegano a pagine in lingue diverse per lo stesso concetto

- **categorie** raggruppano pagine per argomento.

Tuttavia la natura enciclopedica di Wikipedia rappresenta un limite, in quanto non tutti i significati sono rappresentati da una pagina ad hoc

**Esempio** “apple” come frutto e albero si trovano nella stessa pagina; questo tipo di copertura però ce lo risolve WordNet.

Inoltre in Wikipedia una cosa importante che manca sono i **verbi**

Sia WordNet che Wikipedia sono rappresentabili come due grafi, per cui possono essere combinati per ottenere il dizionario enciclopedico multilingua **BabelNet** (*un WordNet multilinguistico*)

## Cos'è BabelNet

BabelNet è un grafo direzionato dove nei nodi abbiamo i **Babel synset**: che sono **concetti** (*play*) e **named entities** (*Shakespeare*) e gli archi connettono coppie di concetti, ad esempio “play is-a dramatic composition”. Ogni arco è etichettato con una relazione semantica  $R$  (is-a, part-of,  $\epsilon$ ,) che denota una relazione.

Esiste anche la possibile relazione  $\epsilon$  che indica un collegamento semantico non specificato.

Ogni nodo quindi contiene un insieme di lessicalizzazioni multilingue del concetto che rappresenta, **esempio** {playen, dramait, obraes, ecc...}. Gli elementi di un Babel synset sono tutti i sinonimi presi da WordNet per un dato concetto, a cui vengono aggiunti i lemmi di altre lingue.

## Costruzione di BN

La costruzione del grafo di BabelNet avviene in tre passi:

1. Hanno preso da WN tutti i sensi (quindi preso e importato in toto) e poi da Wikipedia hanno preso tutte le entry (pagine) e tutte le relazioni (link) nelle varie pagine.
  - a. I concetti recuperati da Wikipedia e da WordNet possono sovrapporsi sia in termini di concetti che di relazioni, quindi è necessaria una unificazione dei dati rimuovendo gli eventuali doppi
2. Le realizzazioni lessicali dei concetti disponibili in più lingue sono raccolte nei Babel synset da Wikipedia (se disponibili) o da traduzioni automatiche di frasi in cui il concetto compare.
3. I Babel synset vengono infine collegati fra di loro stabilendo relazioni semantiche. Queste relazioni sono create sulla base delle relazioni recuperate da WordNet e Wikipedia

valutate  
secondo una certa misura di affinità.

---

## ▼ Nasari

### Introduzione

NASARI è la controparte vettoriale di BabelNet, quindi non è una rete semantica, ed è la rappresentazione vettoriale di un synset BabelNet

Nasce con lo scopo di creare una rappresentazione vettoriale del senso di una parola che non dipenda dalla semantica distribuzionale, ovvero da quanto quella parola appaia vicino a delle

co-occorrenze. Quindi mette insieme significati diversi di un termine in un vettore unico.

L'approccio distribuzionale non riesce infatti a isolare il singolo senso della parola, ma forma dei vettori che racchiudono più sensi.

L'approccio di NASARI, essendo basato su BabelNet, sfrutta i gloss di WordNet per riuscire a

creare vettori per un singolo senso.

### Muffin

Le rappresentazioni vettoriali hanno diversi problemi, ad esempio non sono in grado di modellare il senso individuale di una parola o di un concetto, visto che fanno confluire diversi significati in un'unica rappresentazione vettoriale. Un prima soluzione a questo problema è data dal MUFFIN.

Quindi si raccoglie informazione sul BabelSynset in questione (sfruttando i gloss di WordNet per riuscire a recuperare informazione) e la si raggruppa in un **subcorpus**: questo subcorpus, contenente l'informazione contestuale, viene confrontato con l'intero corpus di Wikipedia, calcolandone la specificità.



L'intero procedimento è fatto con l'ausilio di Wikipedia perchè un problema di questo approccio (se non usiamo Wikipedia) è la generazione di un vettore che rappresenti bene i concetti, anche nei casi ambigui.

**Esempio**, il vettore per il concetto "*piano*" deve permetterci di disambiguare tra l'aggettivo e il sostantivo.

Sulla base della specificità ottenuta, si ottiene una rappresentazione vettoriale del senso in esame che può essere di due tipi

- **vettori lessicali** rappresenta un concetto, attraverso un vettore di item lessicali

**Esempio** “*tavolo*” ha come dimensioni del suo vettore lessicale “*gamba*”, “*piano*”, ecc.);

```
BabelSynsetId  WikipediaPageTitle  lemma1_weight1  lemma2_weight2
```

- **vettori unified** rappresenta un concetto come un vettore di altri sensi (babel synset) dove i primi elementi sono più importanti per descrivere il senso e gli ultimi sono meno importanti

**Esempio** se parliamo di *play(drama)* il termine *letteratura* sarà nelle prime celle del vettore

```
BabelSynsetId  WikipediaPageTitle  synset1_weight1
```

I due tipi di vettori sono memorizzati come sequenze separate da tabulazioni, il cui primo elemento denota l’ID del BabelSynset, il secondo il titolo della pagina di Wikipedia associata al BabelSynset (se esistente) e i successivi elementi i lemmi o i synset (in base al tipo di vettore) con il loro peso associato dopo il carattere di underscore.

La versione che è venuta dopo è quella dei **vettori embedded** che sono rappresentati da una sequenza di numeri e non sono leggibili dall’uomo.

## Weighted Overlap

Dato che i due vettori (lexical e unified) non contengono valori numerici è necessario introdurre una metrica apposita.

$$WO(v_1, v_2) = \frac{\sum_{q \in O} \left( \text{rank}(q, v_1) + \text{rank}(q, v_2) \right)^{-1}}{\sum_{i=1}^{|O|} (2i)^{-1}}$$

Per cui è stata identificata questa metrica dove di questa formula la parte più interessante è il numeratore dove vediamo che la WO tra due vettori è costituita dalla sommatoria su tutte le feature che sono in comune tra i due vettori in questione, e per ogni feature in comune si calcola il rango della feature sul primo vettore e anche sul secondo vettore, tutto questo alla meno 1





Il **rango** è la posizione della feature all'interno del vettore.

**Esempio** *rango=1* significa semplicemente che la feature è in cima al vettore

## Summarization

è una delle più importanti applicazioni di Nasari, abbiamo due tipi di riassunti

- **estrattivo** sono costruiti riutilizzando parti importanti del testo in input come frasi o paragrafi
- **astrattivo** generano il riassunto sulla base delle informazioni raccolte, senza riutilizzarle in modo diretto quindi rielaborando il discorso

## Caratteristiche di un riassunto

Il riassunto deve essere

- **indicativo**, ossia deve fornire un'idea sul contenuto del documento
- **informativo**, cioè deve contenere tutti gli elementi principali e rilevanti
- **critico**, ossia valutare il documento esprimendo una visione sulla qualità del lavoro dell'autore

Un riassunto è inoltre caratterizzato da

- un **fattore di compressione** (ovvero quanto più corto è il riassunto del testo di partenza)
- il **pubblico** a cui si rivolge (più o meno approfondito)
- la **coerenza** del testo prodotto (quanto bene si è riusciti a risolvere le anafore, quante frasi si ripetono, ecc.)
- il **numero di documenti** da usare per il riassunto (documento singolo o multiplo).

## Approcci

- **approccio shallow**, l'approccio è statistico e opera a livello lessicale contando le co-occorrenze dei termini. Lo svantaggio principale è la bassa qualità del riassunto prodotto, inoltre si necessita di grandi quantità di testo per fornire buoni risultati.
- **approccio profondo** prova a capire il testo e necessita di regole per l'analisi e la manutenzione di tesi. Il vantaggio è che il riassunto prodotto è di elevata qualità, tuttavia il processo è lento

## Criteri di rilevanza

Devono essere stabiliti dei criteri per estrapolare dal testo di partenza le parti salienti.

Esistono

numerosi approcci

1. **Posizione nel testo** consiste nel selezionare le frasi che appaiono in specifiche posizioni,  
come ad esempio l'introduzione e la conclusione.
2. **Metodo del titolo** si estraggono dal titolo del documento le parole che costituiranno le keyword per individuare delle frasi importanti nel testo.
3. **Optimum Position Policy (OPP)**: si basa sull'assunzione che le frasi importanti siano posizionate in punti chiave conosciuti in anticipo (da un esperto del dominio)
4. **Cue Phrases** le frasi nel testo più importanti contengono delle parole/sintagmi che fungono da  
indizio allora gli diamo un punteggio positivo oppure fuorvianti allora punteggio negativo

| **Esempio** tipo “*per questo motivo...*”, “*quindi in sostanza...*” ecc..

5. **Metodi basati sulla coesione** sfruttano approcci diversi per creare delle strutture semantiche  
per le frasi che indicano quanto connesse siano fra loro.



---

## ▼ VerbNet

### Introduzione

Consideriamo questa domanda e le sue due risposte

- Q: Who **created** the first effective polio vaccine?

-  - A1: [Becton Dickinson<sub>agent</sub>] **created** [the first disposable syringe<sub>theme</sub>] for use with the mass administration of the first first effective polio vaccine
-  - A2: [The first effective polio vaccine<sub>theme</sub>] was **created** in 1952 by [Jonas Salk<sub>agent</sub>] at the University of Pittsburgh

Potremmo essere tentati di vedere quanto termini della domanda sono contenuti nelle due risposte e riportare la risposta che ha più termini rilevanti. La risposta è nella seconda frase, e quindi questo approccio risulta inefficiente perchè se cerchiamo il verbo “*created*” questo verbo è in entrambe le risposte, quindi siamo punto e a capo. Allora usiamo la rappresentazione semantica.

## Perchè usiamo la rappresentazione semantica?

Perchè così facendo noi cerchiamo dei pezzi di informazione inerenti alla domanda, noi stiamo cercando chi ha creato *cosa*, dove il *cosa* è il vaccino non the “*first disposable syringe*”. Quindi non ci basiamo tanto sui termini rilevanti ma proprio di individuare dei **ruoli semantici**, e capire chi è l'agente nella frase (il vaccino in quato caso)



Individuare il **tema** e l'**agente** di una frase è lo scopo di questa risorsa.

**Ipotesi** il comportamento di un verbo è in larga misura determinato da suo significato, il livello sintattico riflette l'insieme dei significati e la semantica complessiva di quel verbo.

## Cos'è VerbNet?

È una rete che unisce la sintassi e il significato semantico dei verbi (che sono rappresentati gerarchicamente). È indipendente dal dominio (è utilizzabile nei contesti più disparati) e contiene collegamenti anche ad altre risorse lessicali come WordNet, FrameNet e PropBank.

### ▼ Differenza tra livello semantico e sintattico

**Esempio** “*John ha lasciato la stanza*” (uscire da) vs “*John ha lasciato una fortuna*” (perso)

A livello sintattico è difficile capire il verbo, per cui si cerca di rilevare a livello semantico gli agenti a cui il verbo si riferisce

## Struttura

È organizzata in **classi verbali**: insieme di verbi che mostrano un comportamento semantico simile e pattern sintattici simili.

Per raggruppare in classi i linguisti hanno usato un metodo che si chiama **alternation** che possono essere di tipo locativo, temporale, ecc.. e questo metodo ci aiuta a disambiguare il senso di un verbo

**Esempio** “*left*” in presenza di un complemento di luogo avrà significato di “*lasciare*”, mentre in sua assenza potrà significare “*uscire*”.

Oppure per capire in due frasi “uguali” chi è l'agente, il paziente

**Esempio** “*Sharon ha dato l'acqua alle piante*” vs “*Sharon alle piante ha dato l'aqua*”

A volte però queste differenze non bastano e serve analizzare le relazioni tra il predicato e i suoi argomenti; per questo in VerbNet le classi sono basate sulla **classificazione di Levin** dove ogni classe

- ha un **referimento univoco**
- un **insieme di verbi membri** (verbi correlati a quel verbo insomma) e delle **proprietà** (le alternation possibili per la classe, ovvero le forme diverse con cui gli argomenti del verbo possono essere espressi).
  - esistono delle alternation particolari, contraddistinte da un \*, che denotano delle realizzazioni lessicali errate per affermare che un certo verbo della classe non può usare quell'alternation.

**Esempio** “*John cut the bread / \*The bread cut*”

John può tagliare il pane ma il pane non può tagliare John

- **ruoli semantici** l'agente, il paziente, ecc.. (come abbiamo visto nell'esempio del vaccino)
  - caso particolare è l'actor che è un agente che non subisce l'azione

**Esempio** “*Marco ha incontrato Luca*” in teoria avremmo due agenti, ma con l'actor possiamo distinguere meglio il ruolo dei due nella frase (Luca sarà l'actor)

- **selectional preferences**, ciascun ruolo ha degli elementi che possono/devono/non devono caratterizzarlo

**Esempio** *mangiare* richiede un agente vivo e senziente e un paziente commestibile)

- **frames** descrivono ogni possibile forma sintattica utilizzabile per i verbi quindi
  - forme transitive, intransitive
  - molte alternation
  - e gli eventi (start, during, end, result) [indicato nel paragrafo sotto].

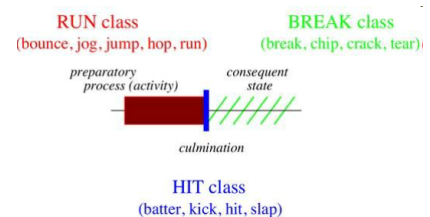
Le classi di Levin però presentavano dei problemi, fra cui la mancanza di omogeneità semantica fra

i verbi contenuti e il fatto che lo stesso verbo potesse essere contenuto in più classi.

Si sono apportate opportune modifiche per risolvere questi problemi introducendo anche una rappresentazione della parte dell'evento a cui quella classe si riferisce.

Infatti, una teoria dice che i verbi si riferiscano a degli **eventi** che possono essere suddivisi in tre parti

- un **processo preparatorio** che ha un inizio, una continuazione e una fine (*run*)
- la **culminazione** ovvero un'azione puntuale (*hit*)
- lo **stato conseguente** uno stato che avviene dopo la culminazione, una conseguenza (*break*)



Così otteniamo appunto quelle che sono le **HIT class**

## ▼ PropBank

### Cos'è?

È un corpus annotato con proposizioni verbali ed argomenti del Penn Treebank.

### Struttura

PropBank usa il concetto di **RoleSet**, che descrive un insieme di ruoli derivanti da un uso particolare

di un verbo, associabile a sua volta a un insieme di frame sintattici che descrivono le variazioni

sintattiche ammesse per quell'insieme di ruoli.

I ruoli di un RoleSet sono detti **argomenti** e sono etichettati con un numero da 0 a 9.

In genere **ARG0** (l'argomento zero) denota l'agente (un proto-agent) e **ARG1** il paziente, ma non è una certezza e più sale il numero dell'argomento e meno è possibile fare questa generalizzazione.

Esistono poi degli argomenti particolari che svolgono il ruolo di **modificatori** per aggiungere

informazione aggiuntiva sull'evento per capire se qualcosa è *causa di...*, *finalità*, *quando si è svolto*, ecc..

<b>ArgM-LOC:</b>	location	<b>ArgM-CAU:</b>	cause
<b>ArgM-EXT*:</b>	extent	<b>ArgM-TMP:</b>	time
<b>ArgM-DIS:</b>	discourse connectives	<b>ArgM-PNC:</b>	purpose
<b>ArgM-ADV:</b>	miscellaneous	<b>ArgM-MNR:</b>	manner
<b>ArgM-NEG:</b>	negation marker	<b>ArgM-DIR:</b>	direction
<b>ArgM-MOD:</b>	modal verb	<b>ArgM-GOL:</b>	goal

Un esempio è ArgM-TMP che denota il tempo dell'evento.

Ad un RoleSet è possibile associare dei frame sintattici: in tal caso il costrutto prende il nome di **Frameset**.



È importante tenere presente che un verbo polisemico può avere più Frameset se le sue differenze di significato sono abbastanza diverse e che i Frameset di PropBank sono diversi da quelli di FrameNet.

Un Frameset infatti rappresenta una distinzione di senso molto grezza, a cui corrispondono molteplici sensi di WordNet. In una frase vengono individuati gli argomenti nella realtà la frase è rappresentata ad albero (deriva da un albero sintattico del Penn Treebank)

### Esempio

Ex: [Arg0 He] [ArgM-MOD would][ArgM-NEG n't] accept [Arg1 anything of value] [Arg2 from those he was writing about]. (wsj 0186)

frase annotata con Frameset

## Mega riassunto di ProbBank

è un tentativo diverso rispetto a VerbNet, invece di menzionare esplicitamente i semantic role, il set di elementi che caratterizzano ciascun verbo è costituito da un insieme di dipendenti, quindi non classi verbali, ma verbi che possono essere raggruppati sulla base di sistemi di dipendenti; questi dipendenti sono una decina di argomenti (arg0, ecc.. poi altri modificatori per specificare: luogo, durata, tempo ecc..)

## ▼ LessLex

### Introduzione

### Reti Semantiche VS Distributional resource (Word embeddings)

Nel mondo esistono due macro-famiglie per rappresentare i sensi delle parole

- **Reti semantiche** (**esempio** Wordnet dove la conoscenza è sotto forma di grafo e i nodi che sono i concetti sono connessi da relazioni semantiche)
- **Risorse distribuzionali** i sensi delle parole sono rappresentati come vettori numerici

Ma ora vediamoli a confronto

### ▼ Reti semantiche

#### Vantaggi

- alta specificità
- multi-lingua

#### Svantaggi

- non sono semplici da usare perchè per calcolare un punteggi di similarità (non è rapido calcolare la distanza su un grafo di due concetti)

### ▼ Risorse distribuzionali

#### Vantaggi

- facilità nel comparare sensi

#### Svantaggi

- perdiamo specificità perchè tutti i sensi di una parola sono rappresentati numericamente in un vettore
- non multi-lingua

## Cos'è LessLex

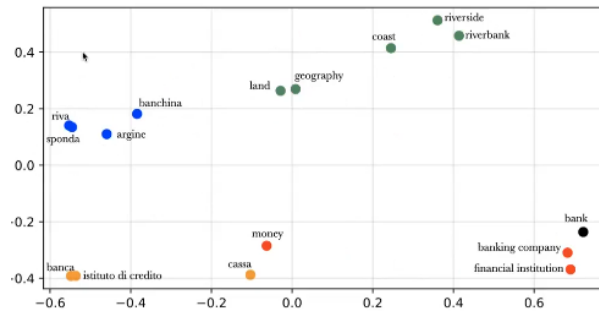
una risorsa che nasce perchè vuole trarre i vantaggi da entrambe le famiglie di risorse (quelle sopra) in cui le unità lessicali sono i synset a cui associamo un vettore numerico, questo perchè vogliamo mantenere la facilità di applicazione delle risorse distribuzionali



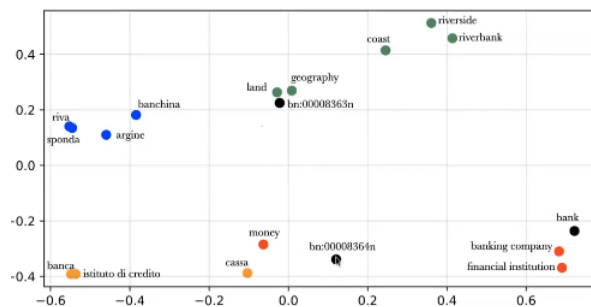
Quindi usiamo la specificità di BabelNet + facilità di confronto di vettori distribuzionali

## Come fare?

- Accedo a BabelNet con un termine (*bank*) e ottengo i suoi significati (synset)
- Si estraggono i termini interessanti dai synset (dalla gloss, dal titolo di Wikipedia, ecc) e cerchiamo un insieme di questi termini interessanti per ogni synset, quindi un senso sarà rappresentato da un set di termini rilevanti
- Ora cerchiamo questi termini dell'insieme su **ConceptNet Numberbatch** e prendiamo il vettore di quel termine, quindi ora un senso sarà rappresentato da un insieme di vettori dei termini rilevanti
- A questo punto ogni vettore definisce un punto nello spazio



- Infine facciamo la media di tutti questi vettori per cogliere il vettore inerente a quel senso (il centroide in sostanza) per cui quando inseriremo “bank” nel grafico noteremo che sarà più vicino al centroide del cluster riguardo alla banca



## Valutazione

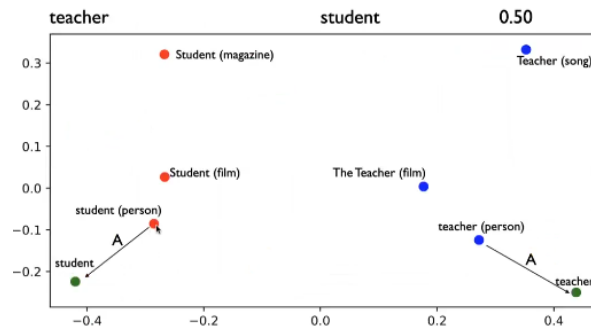
Si è usata inizialmente la **Semantic similarity** in base a questa valutazione i dottorandi hanno notato che quando noi essere umani facciamo similarity, per dare un valore di similarity associamo ai termini, dei significati

**Esempio** *student* e *teacher* noi li associamo al contesto scuola e gli diamo in punteggio di 0.5 mentre per le stesse parole l'algoritmo ritorna 0.81 questo perché i punti di *student* e *teacher* nel piano cartesiano, erano vicini perché si riferivano a dei film

Quindi si è cambiata prospettiva andando ad individuare il significato associato ai termini, non di emulare lo score di similarity

Per cui si è usata la **Rank similarity** che funziona meglio perché considera nel piano cartesiano un **senso** e il **termine da cui siamo partiti**. Infatti la parola *student* e il senso di *student* associato a persona (cioè *student[person]*) è più vicina rispetto a *student[film]* e *student [magazine]*.





Quindi questa metrica riproduce risultati più fedeli in comparazione ai valori dati dagli umani, infatti non dava più 0.81 ma 0.55 (simile a 0.5 degli umani)

## ▼ Principio di composizionalità

è un principio che stabilisce che il significato di un'espressione linguistica è determinato solo dal significato delle sue espressioni costituenti e dalle modalità con cui sono combinate.

**Esempio** il significato della frase "*Socrate è un uomo*" è completamente dato dal significato delle parole "*Socrate*" e "*uomo*" e della connessione logica "*è un*".



Il principio quindi asserisce che vale per le espressioni linguistiche quello che vale, per esempio, per le espressioni aritmetiche, in cui il valore di un'espressione dipende dal valore degli operandi e dal significato dell'operatore con cui sono combinati.



La prima formulazione esplicita di questo principio è normalmente attribuita a Gottlob Frege.

## Perché la composizionalità?

La vogliamo perché possiamo teoricamente elaborare infinite stringhe considerando che abbiamo capacità finite di memoria, per cui vorremmo avere un meccanismo semplice che spieghi molti fenomeni

## Struttura

ci basiamo sulle seguenti considerazioni

- **Productivity argument** dato che esistono finite parole nel dizionario e finiti modi di combinarle la composizionalità ci rende funzionante questo principio. Vogliamo essere produttivi e la composizionalità è ottimale.

Ha solo il problema che non riesce a distinguere espressioni idiomatiche

**Esempio** ‘*flat tyre*’ (gomma a terra) , ‘*flat beer*’ (birra sgasata) in questi casi non riusciremmo ad attribuire il senso corrente per ogni espressione *adj-noun*.



La composizionalità può essere preservata se esiste una KB dove abbiamo già queste espressioni etichettate con il loro senso

- **Systematicity argument** la capacità di comprendere alcune frasi che sono simili a quelle che già sappiamo comprendere

Il problema è che il linguaggio spesso non è sistematico

**Esempio** Tutti quelli che capiscono “*entro un'ora*” e “*senza orologio*” capirebbero anche “*entro un orologio*” e “*senza un'ora*”?

Allora il passo successivo è stato quello di ipotizzare che esistesse un argomento modulare

- **Modularity argument**: individuiamo due tipi di modularità
  - **Modularità di Chomsky e Fodor** il potere generativo di una grammatica può essere catturato da un solo modulo (con conoscenze lessicali + regole sintattiche)
  - **Modularità rilassata** basta un modulo per l'analisi sintattica e un modulo per significati e regole di combinazione

## Composizionalità incrementale

Significa che esistono degli step della computazione che procedono linearmente seguendo il flusso dell'input che percepiamo quando sentiamo una frase, in cui il significato di un'espressione complessa ad uno stage di processing è computata basandosi sulle espressioni costituenti a quello stage e sulla struttura sintattica costruita fino a quello stage. La composizionalità deve essere passata in maniera incrementale.

L'ipotesi **semplice** dice che noi elaboriamo le frasi mantenendo un'ipotesi corrente in base a quello che abbiamo sentito e che stiamo sentendo ma, non spiega come sia rivedibile questa ipotesi che ci costruiamo quando nella frase che stiamo ascoltando arrivano

elementi non compatibili con quello che abbiamo ipotizzato al passo precedente ( *sigma* )

Ci sono due elementi fondamentali di questa ipotesi

- i significati delle espressioni non hanno a che fare con il contesto in cui occorrono

**Esempio** se dico *sedia* intendo sempre sedia

- principio inside-out, l'analisi della frase procede in maniera bottom up

Ma non è in grado di gestire le costruzioni progressive (dove il significato dipende dal contesto in realtà)

**Esempio** "*Lei assomiglia a sua madre*" vs "*lei sta assomigliando a sua madre*"

## Principio del contesto

è opposto a quello della composizionalità: dice che dobbiamo memorizzare già tante informazioni perchè il significato non può essere costruito e dedotto da significati di partenza e man mano come nella composizionalità

## Evidenze sperimentali

Esistono evidenze sperimentali per avere delle idee su come funziona la composizionalità e quali sono i meccanismi che la mettono in sofferenza

- **Semantic illusions** selezioniamo soltanto alcuni tratti significativi di una frase

**Esempio** se uno chiede "*quanti animali Mosè ha caricato sull'arca?*", siamo tentati di rispondere "*due per ogni razza*", ma in realtà Mosè non c'entra un cazzo con Noè

- **Misinterpretations** frasi cortocircuitanti

**Esempio** "*Mentre anna vestiva il bambino giocava nel letto*" invece è più comprensibile "*Mentre anna vestiva il bambino, il bambino giocava nel letto*". Infatti è il ruolo del bambino che crea un *cortocircuito* perchè all'inizio è parsificato come l'oggetto di "*vestiva*" e poi parsificato come soggetto di "*giocava nel letto*"

- **Event-Related brain potentials (ERP)** risposte cerebrali che diamo a certi stimoli con una frase, un'immagine, ecc.. I neuroscienziati distinguono principalmente due onde
  - N400 impulsi che produciamo dal processing semantico di una frase (sia positivo che negativo). L'ampiezza di questa onda dipende dal termine che l'ha prodotta con il contesto in cui compare.
  - P600 è associata a violazioni di vincoli sintattici, tipo quando ci troviamo davanti alle semantic illusion, o misinterpretations

### Esempio

*I treni olandesi sono gialli e molto affollati (vera)*

*I treni olandesi sono bianchi e molto affollati (falsa)*

*I treni olandesi sono acidi e molto affollati (anomala)*

Quello che osserviamo proponendo queste frasi e misurando gli stimoli delle persone, è che la N400 fornisce delle evidenze convergenti (in comune tra la gente) riguardo alle frasi falsa e anomala. Questo perché le persone sanno che un treno non può essere acido anche se non conosco i treni olandesi, e poi esiste la **word knowledge** cioè le persone che sono state in Olanda sanno che non sono bianchi ma sono gialli

- **Coercion** ha luogo quando ci sono entità la cui interpretazione sintattica viene forzata a diventare eventi

**Esempio** "il giornalista ha iniziato l'articolo", l'articolo non si inizia, sarebbe come dire "ha iniziato a scrivere"

Tutte queste sperimentazioni forniscono evidenze neuro-scientifiche sul fatto che il principio della composizionalità non funziona bene

## Conclusioni

Per salvare il principio di composizionalità è necessario ipotizzare che esistano delle KB esterne che aiutano nel processo. Ma allora c'è un problema di fondo: non esistono meccanismi semplici per spiegare la variabilità del linguaggio e la pluralità di fenomeni che caratterizzano l'incontro di vari pezzi lessicali.

L'ipotesi di composizionalità è importante, ma gli strumenti attuali non ci consentono di considerare quest'ipotesi come valida ed attendibile (in pratica: coming soon)



Questo documento è stato fatto con Notion, se lo usi e vuoi direttamente la pagina (per fare modifiche, ecc..), sono disponibile a passarla. Scrivimi → <https://t.me/Aleeeee96>