

INTRODUZIONE ALLA LINGUISTICA COMPUTAZIONALE (STORIA, METODI, TERMINOLOGIA)

OLTRE LA REGOLA
Eleonora Litta

CHE COS'È LA LINGUISTICA COMPUTAZIONALE

- La linguistica computazionale si concentra sullo sviluppo di formalismi descrittivi del funzionamento di una **lingua naturale**, che siano tali da poter essere trasformati in **programmi eseguibili da computer**.
- È un campo **interdisciplinare** che coinvolge due o più materie o aree di conoscenza diverse.
- È **multidisciplinare**: diverse materie di studio in un'unica attività (Informatica e Linguistica, Studi Umanistici e Ingegneria).
- NON è la stessa cosa di:
 - ← Informatica Umanistica, Cultura Digitale (AIUCD)
 - ← Digital Humanities (DH)

LA MISSIONE DELLA LINGUISTICA COMPUTAZIONALE

- Descrivere **formalmente** il linguaggio naturale per sviluppare programmi e strumenti per vari compiti di **elaborazione del linguaggio naturale (NLP Natural Language Processing = TAL Trattamento Automatico del Linguaggio)**
=> è LINGUISTICA!
- Utilizzare strumenti NLP per **costruire, analizzare e interrogare** grandi quantità di dati linguistici (scritti, parlati, audio, video)
- Risultati pratici: T9, Google, Web Semantico, Chatbots, Siri, Alexa, Algoritmi Social Media...

L'IMPORTANZA DELLA MODELLIZZAZIONE

- **Modello:** rappresentazione astratta della conoscenza linguistica per simulare la performance linguistica umana (la ricerca come imitazione della natura)
- Una caratteristica importante di un modello è la sua capacità di **generalizzare** (e quindi di spiegare e/o predire) un fenomeno, oltre le sue singole istanze
- I modelli non sono mai neutrali rispetto alla realtà che descrivono

COMPONENTI DI UN MODELLO

- **Dati:** rappresentativi ed equilibrati. Il modello deve spiegare i dati → [vedi sezione su] Risorse Linguistiche
- **Categorie:** per raggruppare i dati che condividono alcune proprietà (es.: Nomi) → insiemi di tag e criteri di annotazione
- **Relazione** tra Dati e Categorie: tramite regole esplicite (rule-based) o un modello matematico implicito (data-driven). Es.: cat (dato) è un Nome (categoria) → metodi di NLP



LINGUISTIC DATA SCIENCE

COS'È LA LINGUISTIC DATA SCIENCE

Un campo della **Scienza dei Dati** che offre una base formale per l'**analisi**, la **rappresentazione**, l'**integrazione** e lo **sfruttamento** dei **dati linguistici** ai fini di:

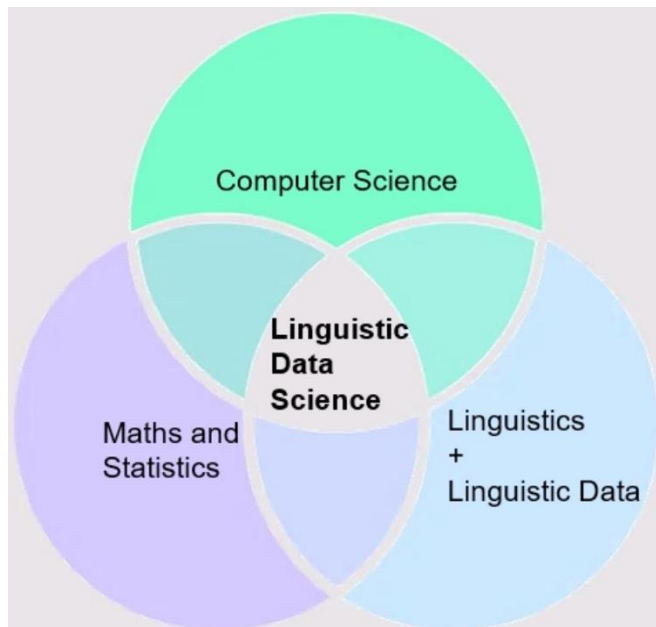
1. Analisi del Linguaggio (morfologia, sintassi, teminologia, ecc.)
2. Applicazioni del Linguaggio (traduzioni automatiche, riconoscimento del parlato, analisi del sentiment, ecc.)

Attraverso l'analisi dei dati e le scienze informatiche è possibile risolvere problemi teoretici ed empirici che non sono stati completamente spiegati da ricerca effettuata nel passato, è possibile sviluppare nuovi modi di fare ricerca linguistica, scoprire nuove teorie attraverso l'uso di una grande massa di dati linguistici.

SCIENZA DEI DATI LINGUISTICI

Di cosa ha bisogno la scienza dei dati Linguistici?

1. Rappresentazione implicita di caratteristiche linguistiche attraverso lo studio di per esempio semantica distribuzionale, embeddings, statistica ecc.
2. Rappresentazione esplicita di dati (linguistici): parole, categorie grammaticali, sensi, concetti, ontologie, dizionari, thesauri, terminologia ecc.



ALAN TURING (1912-1954)

- **Crittanalisi:** Bombe (UK) vs. Enigma (D)
- **Macchina di Turing:** una macchina astratta che manipola simboli su un nastro secondo una tabella di regole. La macchina opera su un nastro di memoria infinito diviso in celle, dove sono scritti i “simboli”
- Il primo computer digitale elettronico: una **macchina universale**, capace in linea di principio di elaborare qualsiasi tipo di dato trattato come simbolo.
- **Codifica Digitale:** Ogni lettera è assegnata a una sequenza di codici 0 e 1. A differenza della rappresentazione analogica (come un disco in vinile), dove il supporto è analogo all'informazione, nel digitale l'informazione è rappresentata da simboli standard (binari).

PRIMI PASSI

Tra gli anni '30 e '60:

- Guerra Fredda fu la ragione principale dei grandi finanziamenti alla ricerca
- 1966 – Interruzione dei finanziamenti. Paradosso tra fiducia e timore verso la tecnologia

Pilastri di quel periodo => tecnologia di oggi

- Crittanalisi => traduzione automatica
- Macchina di Turing => concetto di computer (macchina finita)
- Intelligenza artificiale => primi passi nel settore

La **Macchina di Turing** come strumento per **la costruzione della conoscenza**: l'idea di **formalizzare un oggetto informale** come il linguaggio **per comprenderlo** profondamente.

ALPAC REPORT

- *Automatic Language Processing Advisory Committee (1966)*
- David Hays (ALPAC member) dà un nome alla disciplina
- Stop alla Machine Translation → c'è bisogno di ricerca fondamentale
- AMTCL → ACL (1968)
- *Machine Translation and Computational Linguistics* → *American Journal of Computational Linguistics* (1974) → *Computational Linguistics* (1980)

IBM MEETING 1972 (PISA)

Il calcolatore, evidentemente, non può lemmatizzare le parole non previste nel vocabolario, ma il problema di più difficile soluzione è quello delle parole omografe; il calcolatore deve decidere, per esempio, se la parola "danno" in un certo contesto sia il sostantivo maschile sinonimo di "rovina" o non piuttosto una forma del verbo "dare", o addirittura il presente di "dannare"; e deve distinguere fra "filare" di alberi e il verbo "filare"; tra gli aggettivi "militare", "popolare", "salutare" e i verbi omografi. Gli esempi per ogni lingua sono, sia pure in misura diversa, purtroppo numerosi.

Purtroppo, perchè almeno fino a che non si troveranno degli algoritmi sicuri per operare lo scioglimento della omografia, l'intervento dello studioso costituisce un collo di bottiglia che rallenta inevitabilmente la serie automatica delle elaborazioni. Ostacola inoltre applicazioni che si propongono l'analisi automatica dei testi per fini di immediata utilità pratica, quali la traduzione meccanica di una lingua o la ricerca automatica delle informazioni in una serie di documenti.

La collaborazione tra i ricercatori e linguisti ha attualmente prodotto metodi per la analisi sintattica automatica delle frasi con i quali è possibile operare corrette distinzioni tra parole omografe appartenenti a diverse parti del discorso. Per esempio, un "programma" impiegato al CNUCE consente di analizzare automaticamente frasi come "lo zio legge il libro", nella quale appaiono ben tre parole omografe: "lo" pronome/articolo; "legge" sostantivo/verbo; "libro" sostantivo/verbo.

IBM MEETING 1972 (PISA)

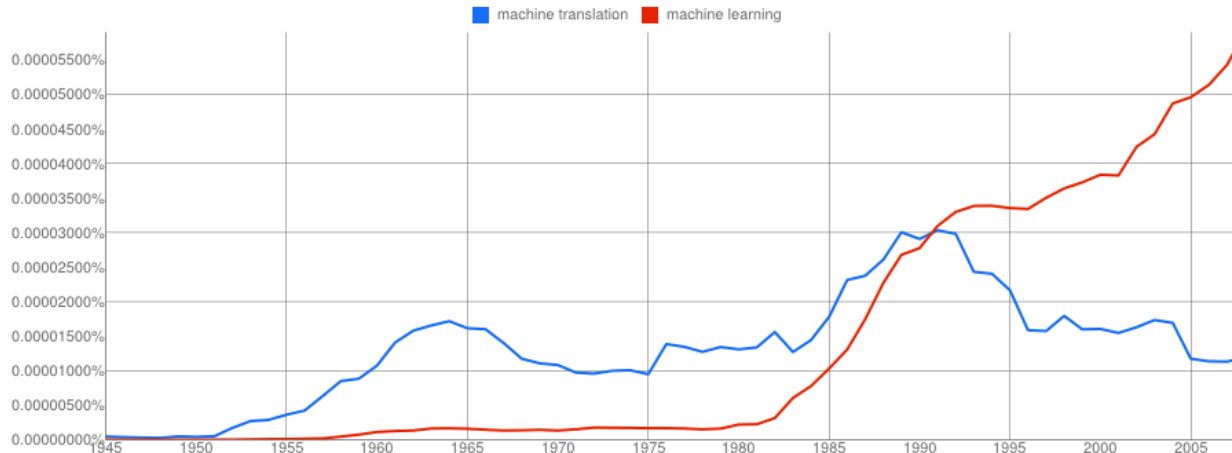
Resta però ancora molto da fare per risolvere l'ambiguità tra parole che appartengono alla stessa parte del discorso ma hanno significato diverso; per esempio "lancia" può significare "imbarcazione" o "arma", "mozzo" può essere una parte della ruota o un marinaio, ecc. Nella frase "il mozzo dubita del suo capitano" è possibile scegliere la seconda interpretazione, perchè nel "programma" memorizzato nel calcolatore è contenuta l'istruzione che verbi quali "dubitare", "ammirare", "temere" richiedono un soggetto animato.

Esistono però innumerevoli casi nei quali il lettore umano deve ricorrere alla sua esperienza e conoscenza del mondo reale, delle relazioni e dei rapporti che intercorrono tra gli oggetti, le persone ecc. Nel caso del calcolatore, si pensi al grande numero di istruzioni e nozioni necessarie a distinguere automaticamente tra la frase "il selvaggio scagliò la lancia" dove si tratta, con ogni verosimiglianza, di una arma, e quella "la lancia si staccò dall'imbarcadero" dove il discorso si riferisce ad un'imbarcazione. Consideriamo un altro esempio: "il cacciatore insegue la fiera" dove "fiera" sta per "belva", mentre nella frase "il cacciatore acquista un fucile alla fiera", "fiera" sta per "esposizione"; oppure pure i diversi rapporti indicati dalla preposizione "nel" nelle due frasi "mangio le fragole nel piatto", "mangio le fragole nel bosco".

La soluzione di questi problemi, in senso lato semantico, è il compito che molti assegnano alla linguistica computazionale per i prossimi trent'anni.

FROM RATIONALISM TO EMPIRICISM

- Da **rule-based** a **data-driven**
- Da “thinking” machines a “learning” machines ...e ora stiamo tornando a “thinking” (AI)?



BEST PRACTICES

- *Aut omnia aut nihil*: not samples!
- *Vel...vel* vs. *aut...aut*: nella maggior parte dei casi DEVI prendere una decisione
- Garbage in ... Garbage out

BEST PRACTICES

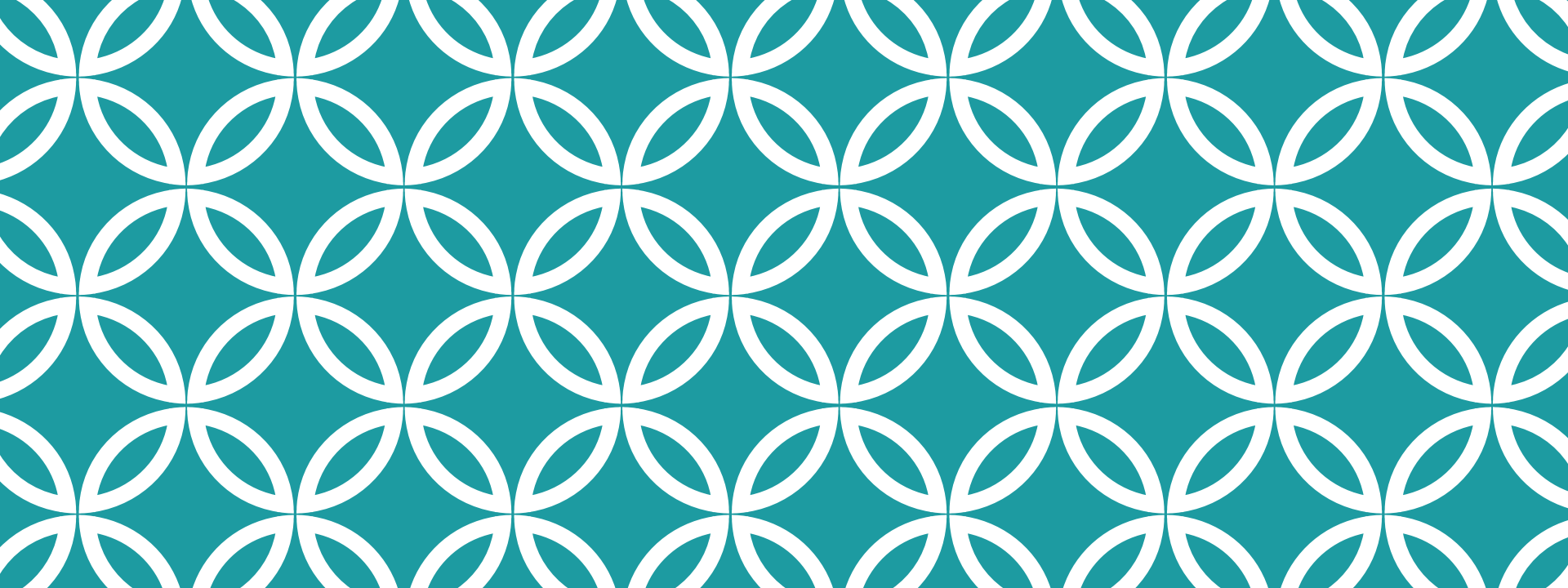
Il ruolo centrale dei dati

- **Conosci i tuoi dati:** pratica il *close reading* (lettura ravvicinata) sia prima che dopo il *distant reading* (l'analisi computazionale su larga scala).
- **Rispetta (e temi un po') i tuoi dati:** non puoi mai sapere con certezza cosa potresti trovarci dentro.
- **Ama i tuoi dati:** in fin dei conti, sono tutto ciò che hai a disposizione.
- **Valida i tuoi dati:** effettua sempre un controllo di qualità (non fidarti mai ciecamente dei risultati prodotti dai software di NLP!).
- **Spiega i tuoi dati:** cura la documentazione (ricordando però l'importanza dell'approccio prospettivista, che tiene conto dei diversi punti di vista nell'interpretazione del dato).
- **Condividi i tuoi dati:** siamo nell'era della Open Data Science, dove la condivisione è essenziale.

BEST PRACTICES

Re- Words

- **Ripetizione** (Repetition): consiste nell'eseguire esattamente la stessa soluzione o lo stesso approccio, nelle medesime condizioni, per giungere al medesimo risultato.
- **Replicazione** (Replication): consiste nell'ottenere la stessa risposta a una determinata domanda di ricerca utilizzando mezzi diversi (ad esempio, valutando un algoritmo su un nuovo set di dati).
- **Riproduzione** (Reproduction): consiste nell'eseguire lo stesso esperimento ma con una configurazione (set up) differente.
- **Riutilizzo** (Reuse): consiste nell'arrivare non solo a risultati finali (output) diversi, ma anche a risultati scientifici differenti. In questo caso l'esperimento è diverso, poiché solo alcune parti rimangono invariate (come gli strumenti, gli script, i dati o i flussi di lavoro).



TERMINOLOGIA

Il lessico della Linguistica
Computazionale

MODELLO

Nelle linguaggio scientifico, un modello è una rappresentazione concettuale astratta e semplificata di un fenomeno.

Serve per

- 1) **Spiegare** un fenomeno che è avvenuto
- 2) Fare **previsioni** riguardo ad un fenomeno

Generalizzare

Un modello deve saper **generalizzare**: estrarre tendenze dai casi particolari a partire dai quali è stato costruito che possano **spiegare/predire i casi ancora ignoti**

Semplificare

Ogni modello comporta una **riduzione** e, perciò, una **semplificazione** del fenomeno che vuole rappresentare.

MODELLO INFORMALE VS FORMALE

Informale

- Espresso in **linguaggio naturale** o attraverso segni interpretabili
- Il lettore umano ha bisogno di un **contesto** per interpretarlo
- Spesso servono anche dei **commenti**, o spiegazioni esplicite

Formale

- Espresso in **Linguaggio formale** (numero finito di simboli e di regole di combinazione)
- Interpretabile indipendentemente dal contesto
- Può essere usato da una macchina!

GRAMMATICA

Una grammatica è un **modello**! Infatti:

1. è una «versione in scala» di un fenomeno: schematizza la lingua con un insieme di enti, categorie e regole combinatorie
2. È esplicitativa e predittiva (prevede la creazione di enunciati ben formati)
3. Può essere informale (libro di grammatica), ma esistono anche grammatiche formali

● Particolarità dei verbi della terza coniugazione

- Il verbo *cucire* mantiene sempre il suono palatale /tʃ/ e per questo inserisce una *i* grafica fra la radice e le desinenze che cominciano per *a* o per *o*: *cuc-i-o*, *cuc-i-ano*; il verbo *fuggire* modifica il suono palatale /dʒ/ in velare /g/) davanti alle desinenze che cominciano per *a* o per *o*: *fugg-o*, *fugg-ano*.
- Alcuni verbi della terza coniugazione hanno due forme di participio presente, una regolare in *-ente* e una in *-iente*: *dormente* e *dormiente*. Altri presentano solo la forma in *-iente*: *obbediente*. Vediamo i casi più comuni:

<i>capire</i>	<i>capiente</i>	<i>nutrire</i>	<i>nutriente</i>
<i>consentire</i>	<i>consenziente</i>	<i>partorire</i>	<i>partoriente</i>
<i>convenire</i>	<i>conveniente</i>	<i>patire</i>	<i>paziente</i>
<i>dissentire</i>	<i>dissenziente</i>	<i>sconvenire</i>	<i>sconveniente</i>
<i>disubbidire</i>	<i>disubbidiente</i> (o <i>disobbediente</i>)	<i>sentire</i>	<i>senziente</i>
<i>dormire</i>	<i>dormiente</i>	<i>ubbidire</i>	<i>ubbidiente</i> (o <i>obbediente</i>)
<i>esaurire</i>	<i>esauriente</i>	<i>venire</i>	<i>veniente</i>
<i>esordire</i>	<i>esordiente</i>		

- Molti verbi della terza coniugazione inseriscono fra la radice e la desinenza l'infisso **-isc-**; ciò avviene nella prima, seconda, terza persona singolare e terza plurale del presente indicativo e congiuntivo, e nella seconda, terza singolare e terza plurale del presente imperativo (tutte le altre forme sono identiche a quelle di *servire*):

INDICATIVO PRESENTE	CONGIUNTIVO PRESENTE	IMPERATIVO PRESENTE
io cap-isc-o	che io cap-isc-a	–
tu cap-isc-i	che tu cap-isc-a	cap-isc-i tu
lui cap-isc-e	che lui cap-isc-a	cap-isc-a lui
noi cap-iamo	che noi cap-iamo	cap-iamo noi
voi cap-ite	che voi cap-iate	cap-ite voi
loro cap-isc-ono	che loro cap-isc-ano	cap-isc-ano loro

Seguono questo tipo di coniugazione i verbi *agire*, *ammonire*, *capire*, *finire*, *obbedire*, *percepire*, *punire*, *scoprire*, *sparire* ecc. Alcuni verbi ammettono ambedue le forme, con e senza infisso: *applaudire* (*applaudo* e *applaudisco*), *assorbire* (*assorbo* e *assorbi-sco*), *comparire* (*compaio* e *comparisco*), *eseguire* (*eseguo* ed *eseguisco*), *mentire* (*mento* e *mentisco*), *nutrire* (*nutro* e *nutrisco*), *partire* (“dividere”, *parto* e *partisco*), *inghiottire* (*inghiotto* e *inghiottisco*) ecc. Per maggiori dettagli v. **Per approfondire** a fine capitolo.

COMPUTER E “COMPUTAZIONE”

Cosa vuol dire «**computazionale**»? Cosa sono i «**modelli computazionali**»?

‘Computare’ significa **calcolare**, e ‘computazionale’ è il termine usato per tutte quelle discipline che usano **la capacità di calcolo dei computer** [...] per svolgere ricerche, eseguire compiti ed effettuare previsioni.

—Jezek & Sprugnoli, p. 18

Cosa vuol dire «calcolare» con gli elementi che costituiscono una lingua? E in che modo sfruttiamo la «capacità di calcolo dei computer» per conoscere la lingua e fare previsioni?

COSA VUOL DIRE CALCOLARE?

- Scriviamo dei simboli
- Ci concentriamo l'attenzione su un simbolo alla volta
- Eseguiamo un'operazione (ad es. moltiplichiamo 7 e 5)
- Scriviamo un risultato parziale (5) e registriamo uno «stato» (il riporto è 3)
- Spostiamo l'attenzione sul simbolo successivo

$$\begin{array}{r} 135 \times \\ 7 = \\ \hline \textcircled{3}5 \end{array}$$

Moltiplichiamo 7 e 5: $7 \times 5 = 35$
Scriviamo 5 e riportiamo 3

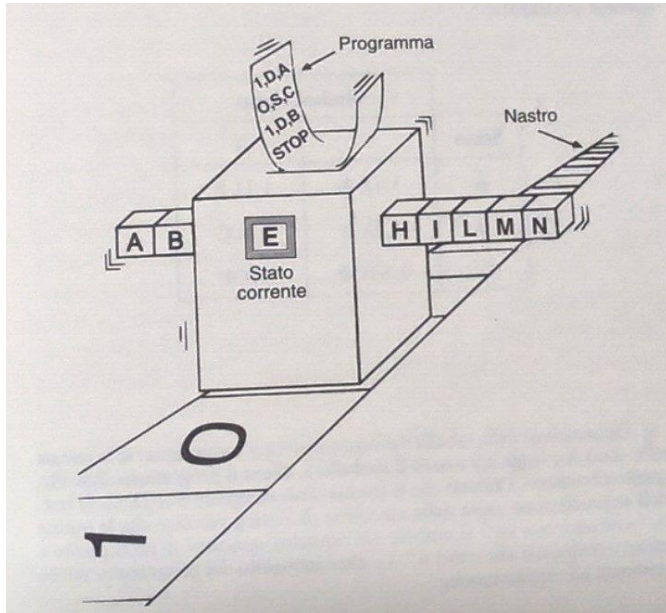
Questi passaggi sono indipendenti dalla persona che li esegue, sono sempre uguali e sono **meccanizzabili** (li può eseguire una macchina)!

ALGORITMO

Una procedura che consiste in una sequenza di istruzioni ben definite, e che consente di eseguire un'operazione in un numero finito di passi.

- Prima fai A e poi fai B
- Se/quando trovi A allora scrivi B
- La procedura per eseguire una moltiplicazione o la ricerca del percorso più breve tra 2 punti sono un esempio di algoritmo.

LA MACCHINA DI TURING



Da J. L. Casti, *[I cinque di Cambridge](#)*, p. 41

La **Macchina di Turing** è un modello astratto di calcolatore per eseguire algoritmi. Si compone di:

- 1. un nastro infinito**, diviso in celle discrete, su cui si leggono e scrivono simboli.
- 2. Una testina di lettura/scrittura**, che si sposta tra le celle e modifica il contenuto secondo regole definite.
- 3. Un insieme finito di stati interni**, che rappresentano la memoria della macchina.
- 4. Un insieme di istruzioni**, che determina il comportamento della macchina in base allo stato e al simbolo letto.

CHE COS'È UN TESTO PER UN COMPUTER?

Per un computer, un testo è semplicemente **un tipo di dato**, **rappresentato come una sequenza di numeri**, secondo un «codice» (che chiamiamo **codifica**). Questa rappresentazione numerica può essere memorizzata nella memoria, o elaborata dalla ALU attraverso operazioni aritmetiche e logiche su questi dati (come: conversione di maiuscole in minuscole, ordinamento alfabetico, crittografia...).



TASK = COMPITI LINGUISTICI

In linguistica computazionale, un **task** (o compito) è un problema specifico e ben definito che un sistema informatico deve risolvere elaborando il linguaggio umano. Possiamo immaginarli come i mattoni fondamentali che compongono applicazioni più complesse: per arrivare a un traduttore automatico o a un assistente vocale, la macchina deve prima risolvere una serie di task intermedi.

Questi compiti vengono generalmente suddivisi in una gerarchia, dai più semplici (analisi della forma) ai più complessi (comprensione del significato e del contesto).

TASK DI PRE-ELABORAZIONE E MORFOLOGIA (BASSO LIVELLO)

Sono i compiti "di base" che preparano il testo per analisi più sofisticate.

Tokenizzazione: Suddividere una stringa di testo in unità minime chiamate *token* (solitamente parole e punteggiatura).

Lemmatizzazione: Ricondurre una parola flessa alla sua forma di citazione sul dizionario (il lemma). Ad esempio: "*amaverunt*" → "*amo*".

Part-of-Speech (PoS) Tagging: Assegnare a ogni parola la sua categoria grammaticale (nome, verbo, aggettivo, ecc.).

TASK SINTATTICI

Si occupano di capire come le parole si relazionano tra loro per formare frasi dotate di struttura.

Parsing (Analisi Sintattica): Costruire la struttura grammaticale di una frase. Può essere a costituenti (gerarchia di sintagmi) o a dipendenze (relazioni logiche tra parole, come soggetto o complemento).

Chunking: Raggruppare parole in unità sintattiche semplici (sintagmi nominali o verbali) senza analizzare l'intera struttura della frase.

TASK SEMANTICI E DI ESTRAZIONE DELL'INFORMAZIONE

Named Entity Recognition (NER): Identificare e classificare entità nominate nel testo, come nomi di persone, luoghi, organizzazioni o date. (Esempio: In "Cesare passò il Rubicone", il sistema identifica *Cesare* come PERSONA e *Rubicone* come LUOGO).

Word Sense Disambiguation (WSD): Determinare quale senso di una parola polisemica è usato in un contesto. (Esempio: distinguere se "miglio" indica la misura di distanza o il cereale).

Sentiment Analysis: Determinare l'opinione o l'emozione espressa in un testo (positivo, negativo, neutro).

TASK DI ALTO LIVELLO (APPLICATIVI)

Questi task combinano tutti i precedenti per generare un risultato finale utile all'utente.

Machine Translation (Traduzione Automatica): Tradurre un testo da una lingua all'altra.

Text Summarization: Creare un riassunto automatico di un documento lungo.

Question Answering (QA): Rispondere a domande poste in linguaggio naturale basandosi su un corpus di testi.

IL CONCETTO DI "PIPELINE"

Nella linguistica computazionale tradizionale, questi task vengono organizzati in una **pipeline** (catena di montaggio): l'output di un task (ad esempio la tokenizzazione) diventa l'input per quello successivo (il PoS tagging), e così via.

Perché è importante per i classicisti?

Per chi studia latino e greco, i task di **Lemmatizzazione** e **PoS Tagging** sono cruciali. Poiché le lingue classiche sono altamente flesse, un sistema deve essere impeccabile nel riconoscere che *“feceris”* e *“faciam”* appartengono allo stesso verbo (*facio*) per poter permettere ricerche lessicali efficaci o analisi stilistiche accurate su un autore.

Oggi, con l'avvento dei **Large Language Models** (come GPT), molti di questi task vengono svolti simultaneamente dal modello, ma la distinzione teorica rimane fondamentale per valutare quanto bene una macchina "capisce" davvero la struttura di una lingua antica.

LINGUAGGI DI QUERY (INTERROGAZIONE)

L'obiettivo è interrogare il corpus per ottenere risposte a quesiti scientifici, come ad esempio:

"Quale lingua disloca il soggetto a destra con maggiore frequenza?"

"Quale lingua presenta più strutture non proiettive (ovvero sintagmi discontinui)?"

Si tratta di porre domande linguistiche direttamente all'evidenza empirica, rappresentata dalla nostra raccolta di dati. Per farlo, queste domande devono essere tradotte in un formato "formale", espresso attraverso un linguaggio che permetta di effettuare interrogazioni: un **linguaggio di query**.

Condizione essenziale: *conditio sine qua non* per poter interrogare i dati e acquisire conoscenza è la padronanza delle **espressioni regolari** (Regex).

RISORSE LINGUISTICHE

Le risorse linguistiche sono collezioni di dati che documentano principalmente gli atti comunicativi degli esseri umani attraverso qualche forma di registrazione e/o descrizione. Esse possono presentarsi sia in forma diretta, come nel caso dei **corpora**, sia a livelli più alti di astrazione, come nei **lessici** e nelle **ontologie**." (*International Standard for Language Engineering, ISLE*).

Componenti cardine nel campo del NLP, oggi sono fondamentali per il *benchmarking*, ovvero per valutare e testare le prestazioni dei grandi modelli linguistici (LLM).

Questi componenti devono essere:

- In un **formato leggibile dalla macchina** (*machine-readable*).
- Composti da **dati** di varia natura: scritti, parlati, audio o video.
- Supportati da **strumenti** specifici per l'acquisizione, la creazione, la manutenzione, la condivisione e l'interrogazione (query) delle Risorse Linguistiche (LR).

TIPI DI RISORSE LINGUISTICHE: RISORSE TESTUALI

- Possono essere **rappresentativi** (di una lingua in generale) o **specifici di un dominio** (settoriali).
- Si distinguono in corpora **grezzi** (solo testo nudo e crudo) o **annotati** (arricchiti con informazioni linguistiche).
- Possono essere: **monolingui** o **multilingui** (in quest'ultimo caso spesso "paralleli", ovvero con testi allineati frase per frase); **scritti, parlati o misti**; **sincronici** (relativi a un preciso momento) o **diacronici** (che studiano l'evoluzione nel tempo); **integrali o parziali**; **statici o dinamici** (in continuo aggiornamento).

TIPI DI RISORSE LINGUISTICHE: RISORSE LESSICALI

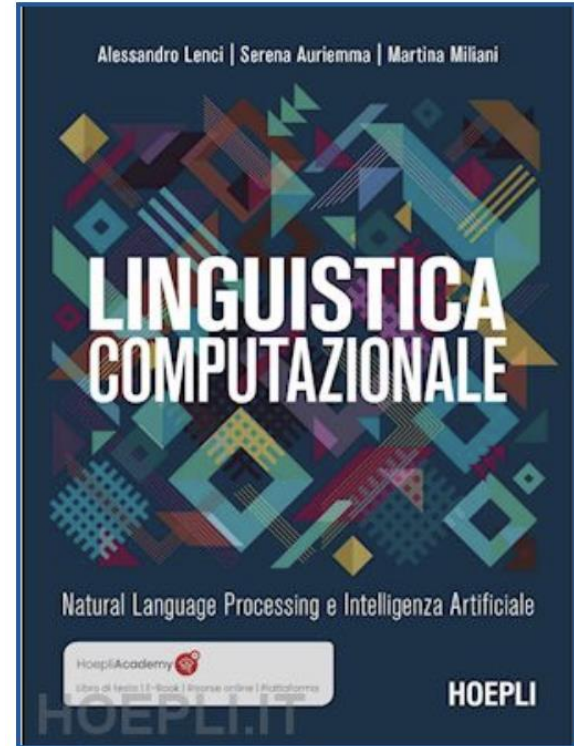
LESSICI E DIZIONARI

- Possono essere di uso **generale** o legati a **domini specifici** (in questo caso si parla di "terminologie").
- Spesso sono **annotati** con informazioni complesse (ad esempio basati sulla **valenza** verbale, che indica quanti e quali argomenti richiede un verbo) o arricchiti con dati sulla **frequenza** d'uso delle parole.

ONTOLOGIE

- Rappresentano in modo **formale ed esplicito** un concetto.
- Definiscono le **relazioni** logiche e gerarchiche che intercorrono tra i diversi concetti (ad esempio, relazioni di tipo "è un", "fa parte di", ecc.).

BIBLIOGRAFIA



DOMANDE?

litta.eleonora@liceomontaleroma.it

Cosa ti aspetti da questo corso:

<https://forms.gle/UxEQcUyT7dbgDBbr8>

