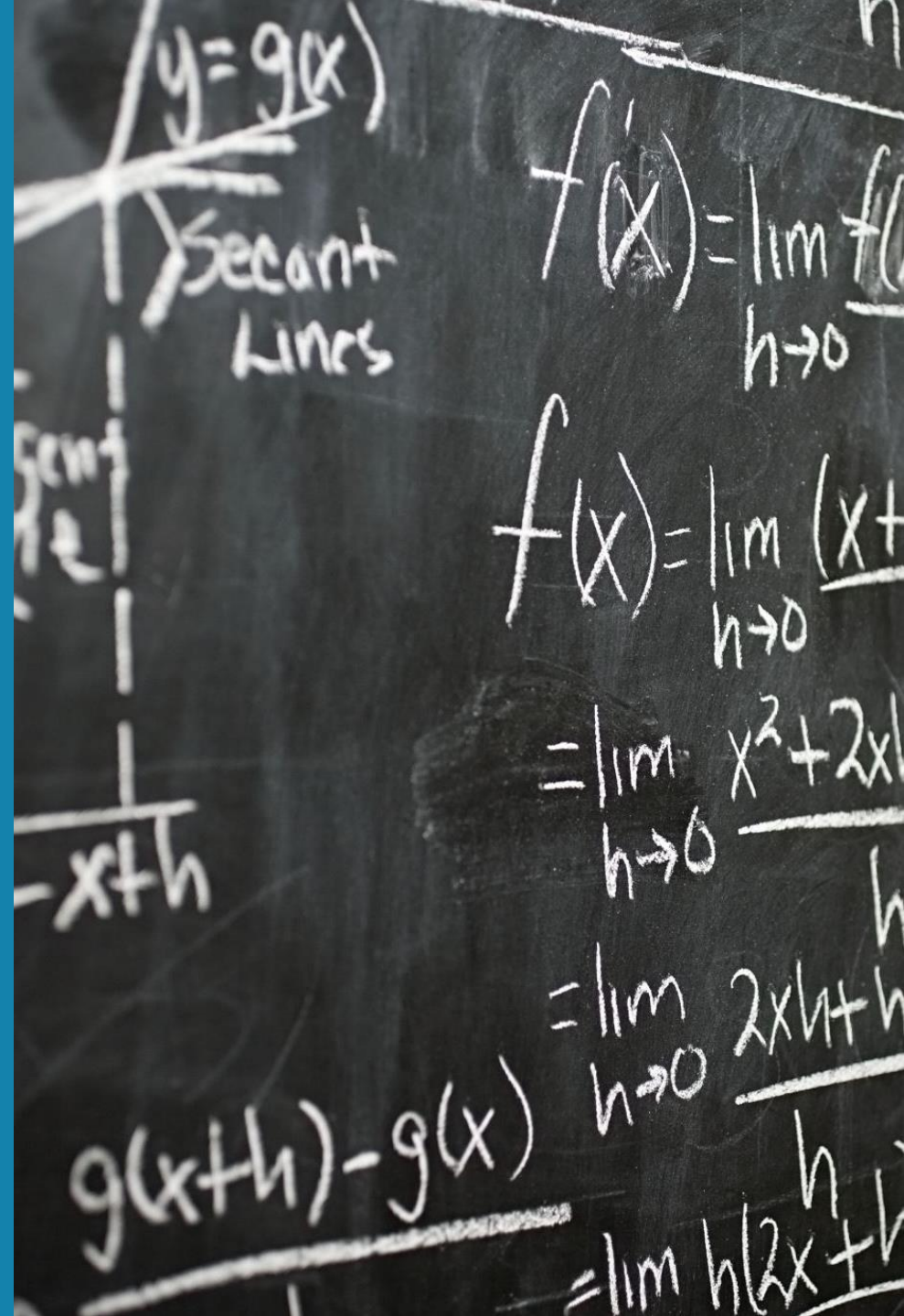


ANNOTAZIONE DEI CORPORA

Livelli di Annotazione

Metodologia



AGGIUNTA DI METADATI

- **Descrittivi** (informazioni di classificazione): includono i dati bibliografici classici, come il nome dell'autore, il titolo dell'opera, l'anno di composizione e la lingua utilizzata.
- **Amministrativi** (informazioni documentarie): servono alla gestione tecnica e legale del file. Comprendono la descrizione dei formati dei file, i tagset (l'elenco delle etichette usate), lo stato delle revisioni, la disponibilità dei dati (licenze d'uso) e chi ne sia il proprietario o il curatore.
- **Editoriali** (relazione tra il corpus e la fonte): riguardano gli interventi effettuati durante la digitalizzazione, come l'indicazione di inserimenti o omissioni, correzioni di refusi del testo originale o normalizzazioni ortografiche (fondamentali, ad esempio, per gestire le varianti del latino medievale).
- **Analitici** (interpretazione e analisi dei componenti del corpus): riguardano l'analisi profonda del testo.
 - **Proprietà strutturali**: divisione in capitoli, paragrafi, titoli.
 - **Contenuto specifico**: citazioni, parole straniere, riconoscimento di entità nominate (NER), ovvero l'identificazione automatica di nomi propri, date e luoghi.
 - **Analisi linguistica**: quella che chiamiamo propriamente "Annotazione del corpus".

LIVELLI DI ANNOTAZIONE

- **Tokenizzazione** (suddivisione del testo in unità minime come parole e punteggiatura)
- (Corpora di parlato) **Annotazione fonetica e prosodica** (accenti, intonazione...)
- **Analisi morfologica e lemmatizzazione**
- **PoS tagging** (etichettatura delle parti del discorso: nome, verbo, aggettivo...)
- **Analisi sintattica** (costruzione della struttura della frase)
- **Risoluzione dell'anafora e dell'ellissi** (identificare a cosa si riferiscono i pronomi o gli elementi omessi)
- **Annotazione pragmatico-retorica** (RST: Rhetorical Structure Theory; forza illocutoria, ovvero l'intenzione comunicativa)
- **Annotazione semantica**: SRL (Semantic Role Labeling - etichettatura dei ruoli semantici come agente, paziente), WSD (Word Sense Disambiguation - disambiguazione del significato delle parole)

LIVELLI DI ANALISI COMPUTAZIONALE

Questi punti rappresentano i "livelli di analisi" che un computer deve affrontare per comprendere un testo. Scala di complessità:

1. Si parte dalla base (la Tokenizzazione, cioè distinguere le singole parole).
2. Si sale verso la grammatica (PoS tagging e Sintassi).
3. Si arriva ai livelli più alti e difficili, dove serve il contesto (Semantica e Pragmatica).

Per le lingue classiche, i primi livelli (Morfologia e Sintassi) sono ormai molto avanzati, mentre la sfida attuale della ricerca si sta spostando proprio sull'annotazione semantica e pragmatica (capire, ad esempio, l'ironia o le metafore in un testo antico attraverso gli algoritmi).

PERCHÉ SI ANNOTA

Scopi Teorici:

Confronto tra Teoria e Dati: L'annotazione permette di verificare se le nostre teorie linguistiche (le "regole" che leggiamo nelle grammatiche) trovano riscontro nell'uso reale della lingua contenuto nei testi.

Caratteristiche Latenti (Nascoste): Aiuta a far emergere schemi e relazioni che non sono visibili a occhio nudo, come i complessi **modelli di dipendenza** tra diverse caratteristiche del testo (ad esempio, come la scelta di un certo tempo verbale influenzi la struttura della subordinata).

PERCHÉ SI ANNOTA

Scopi Pratici:

Addestramento di strumenti di NLP: L'annotazione umana fornisce il "materiale didattico" (il cosiddetto *Gold Standard*) necessario per insegnare agli algoritmi di Intelligenza Artificiale come analizzare correttamente nuovi testi.

Gestione di grandi collezioni di dati: Grazie a un **set limitato di termini** (etichette o *tag*), possiamo interrogare migliaia di testi contemporaneamente. Invece di cercare una singola parola, possiamo chiedere al sistema di trovare "tutti i verbi al congiuntivo perfetto", cosa impossibile senza un'annotazione preventiva.

I REQUISITI IDEALI DELL'ANNOTAZIONE (DESIDERATA)

Perché un'annotazione linguistica sia di alta qualità, deve bilanciare tre fattori spesso in conflitto tra loro:

Velocità: Il processo deve essere efficiente per poter gestire grandi moli di testi in tempi ragionevoli.

Coerenza (Consistency): Annotatori diversi (o lo stesso annotatore in momenti diversi) devono etichettare lo stesso fenomeno nello stesso modo. Senza coerenza, i dati sono "sporchi" e inutilizzabili per la macchina.

Profondità (Significatività): L'annotazione non deve essere superficiale; deve catturare la complessità e il significato reale del testo.

COSA SERVE PER OTTENERLI?

Un flusso di lavoro (workflow) chiaro e interfacce intuitive: Gli strumenti software (come *Inception*) devono essere facili da usare per ridurre gli errori tecnici e la fatica dell'utente.

Un gruppo di annotatori umani (per il controllo incrociato): Non basta una sola persona. Serve un team che lavori sugli stessi testi per effettuare il **cross-checking**. Questo permette di misurare l'*Inter-Annotator Agreement* (il grado di accordo tra esperti), che garantisce l'oggettività scientifica del lavoro.

Una teoria di riferimento (Source theory): Non si annota nel vuoto. Serve una solida base teorica (ad esempio, una specifica grammatica delle dipendenze per il latino) che stabilisca le regole da seguire quando si presentano casi ambigui.

ACCORDO TRA ANNOTATORI (INTER-ANNOTATOR AGREEMENT, IAA)

Il **Kappa di Cohen** e il **Kappa di Fleiss** sono misure statistiche utilizzate per valutare il livello di accordo tra due o più valutatori (o annotatori) durante la categorizzazione di elementi o soggetti.

Il **Kappa di Cohen** è progettato specificamente per il confronto tra **due** valutatori.

Il **Kappa di Fleiss** estende questo calcolo a **più** valutatori.

Entrambi i coefficienti tengono conto dell'accordo che potrebbe verificarsi per **puro caso**, fornendo una misura dell'accordo reale molto più accurata rispetto alla semplice percentuale di accordo.

Punteggio:

≤ 0 : Nessun accordo

0,01–0,20: Accordo da nullo a scarso

0,21–0,40: Accordo discreto

0,41–0,60: Accordo moderato

0,61–0,80: Accordo sostanziale

0,81–1,00: Accordo quasi perfetto

CICLO DI ANNOTAZIONE MATTER

L'annotazione non è un processo lineare, ma un ciclo virtuoso di apprendimento:

M - Model (Modellazione): Definiamo la teoria e lo schema di annotazione (es. quali etichette usare per i casi latini?).

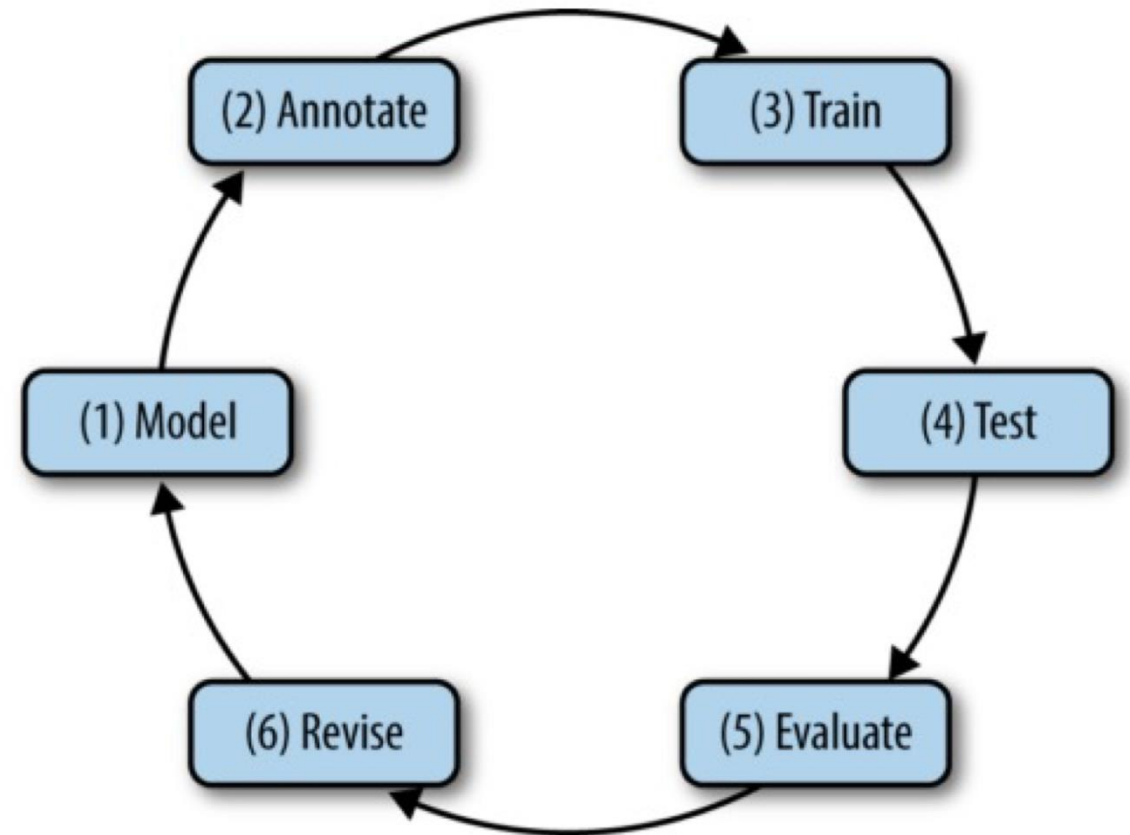
A - Annotate (Annotazione): Gli esperti umani annotano un campione di testi seguendo il modello.

T - Train (Addestramento): Forniamo questi dati annotati all'algoritmo perché impari a riconoscere gli schemi.

T - Test (Test): Facciamo analizzare alla macchina testi nuovi, mai visti prima.

E - Evaluate (Valutazione): Confrontiamo l'analisi della macchina con quella umana (Gold Standard) per misurare l'errore.

R - Revise (Revisione): Sulla base degli errori, modifichiamo il modello o aggiungiamo dati, e ricominciamo.



FORMATI DI ANNOTAZIONE

Esistono diversi modi per "consegnare" l'analisi linguistica a un computer. La scelta del formato dipende dal tipo di analisi (sintattica, semantica, NER) e dallo strumento software utilizzato.

Annotazione In-line: Le etichette linguistiche vengono inserite direttamente nel flusso del testo originale. Il formato principe di questa modalità è l'**XML** (nello standard TEI).

Annotazione Stand-off: Il testo sorgente rimane intatto in un file "puro", mentre le annotazioni vengono salvate in file separati che puntano al testo tramite coordinate o ID. È la soluzione necessaria per gestire analisi massive, collaborazioni globali e il problema delle gerarchie sovrapposte.

FORMATI DIVERSI PER OBIETTIVI DIVERSI

Oltre alla modalità (dove salvo i dati), il **formato** cambia radicalmente a seconda di cosa si vuole annotare e di *quale strumento* si intende addestrare:

- 1) Per la **Morfologia** e la **Sintassi delle Dipendenze**: formati tabulari, come il CoNLL-U. Sistema "pulito" e compatto, ideale per alimentare gli algoritmi di apprendimento automatico (Machine Learning) che devono imparare le regole grammaticali.
- 2) Per la **Sintassi a Costituenti: Bracketing**, che permette di visualizzare rapidamente i sintagmi come scatole cinesi annidate.
- 3) Per il **riconoscimento di entità (NER)**: formato **IOB**, una codifica sequenziale semplicissima che permette alla macchina di capire dove inizia e dove finisce un nome proprio o un luogo.
- 4) Per l'**edizione critica e la filologia**: Lo standard è l'**XML/TEI**, l'unico formato abbastanza flessibile da poter descrivere contemporaneamente la lingua, le varianti dei manoscritti e la struttura fisica del supporto (es. righe di un papiro o colonne di un codice).

ANNOTAZIONE IN-LINE: BRACKETING

È il formato storico dei primi grandi progetti (come il *Penn Treebank*). Rappresenta la struttura sintattica nidificando i costituenti tra parentesi tonde.

Esempio:

```
(S (NP-OBJ (NP (NOUN Arma)) (NP (NP (NOUN virum)) (CONJ que))) (VP (VERB cano)) (PUNCT ,) (S-REL (NP-GEN (PROPN Troiae)) (NP-SUBJ (PRON qui)) (ADJP-PRED (ADJ primus)) (PP (ADP ab) (NP-ABL (NOUN oris)))))
```

Vantaggio: Molto compatto.

Svantaggio: Difficile da leggere per l'occhio umano quando le frasi sono lunghe e complesse.

ANNOTAZIONE IN-LINE: XML-TEI

```
<s n="1">
  <w lemma="arma" pos="N" msd="Case=Acc|Numb=Plur|Gend=Neut">Arma</w>
  <w lemma="vir" pos="N" msd="Case=Acc|Numb=Sing|Gend=Masc">virum</w>
  <w lemma="que" pos="C">que</w>
  <w lemma="cano" pos="V" msd="Mood=Ind|Tense=Pres|Pers=1|Numb=Sing">cano</w>
  <pc>,</pc>
  <w lemma="Troia" pos="PropN" msd="Case=Gen|Numb=Sing|Gend=Fem">Troiae</w>
  <w lemma="qui" pos="Pron" msd="Case=Nom|Numb=Sing|Gend=Masc">qui</w>
  <w lemma="primus" pos="Adj" msd="Case=Nom|Numb=Sing|Gend=Masc">primus</w>
  <w lemma="ab" pos="Prep">ab</w>
  <w lemma="ora" pos="N" msd="Case=Ab1|Numb=Plur|Gend=Fem">oris</w>
</s>
```

Vantaggi:

Leggibilità: Il legame tra la parola e la sua analisi è immediato e inseparabile.

Gerarchia: Eccellente per rappresentare strutture annidate (parola > frase > paragrafo > capitolo).

SVANTAGGI DELL'IN-LINE

- **Gerarchie Sovrapposte (Overlapping):** Il limite più critico. Il formato XML non permette a due etichette di incrociarsi. Esempio: Non puoi chiudere il tag di una <pagina> se non hai prima chiuso quello della <frase>. Se la frase continua nella pagina successiva, il sistema va in errore.
- **"Tag Soup"** (Zuppa di etichette): All'aumentare dei livelli di analisi (metrica, sintassi, varianti), il testo originale "scompare" sotto una montagna di codice, diventando illeggibile per l'occhio umano.
- **Difficoltà di Collaborazione:** È quasi impossibile far lavorare due persone diverse sullo stesso file in-line senza creare conflitti tecnici o errori di formattazione.
- **Incompatibilità tra analisi:** Se vuoi aggiungere un nuovo livello di ricerca anni dopo (es. analisi del sentimento su un testo già annotato sintatticamente), rischi di dover riscrivere l'intero file.
- **Fragilità del dato:** Basta un solo errore di battitura in un tag (es. mancare una / o una >) per rendere l'intero documento inutilizzabile dai software.

ANNOTAZIONE STAND-OFF

Il testo sorgente e le informazioni linguistiche sono memorizzati in **file separati**. Il testo originale rimane "intatto" (sola lettura), mentre le annotazioni puntano a porzioni specifiche di esso tramite coordinate.

I vantaggi della separazione:

Gestione di gerarchie sovrapposte: Poiché le annotazioni sono in file diversi, non c'è più il limite dei tag XML che si incrociano. Posso annotare metrica e sintassi contemporaneamente senza conflitti.

Multilivello e Scalabilità: Posso aggiungere infiniti strati di analisi (semantica, pragmatica, varianti testuali) creando semplicemente nuovi file di annotazione, senza mai appesantire il testo originale.

Integrità del dato: Il testo sorgente non rischia di essere corrotto da errori di battitura nei tag.

Collaborazione: Diversi team possono lavorare in parallelo sullo stesso testo, ognuno producendo il proprio file di annotazione stand-off.

Come funziona? (Il meccanismo dei puntatori)

Testo (File A): "Arma virumque cano..."

Annotazione (File B): "Dal carattere 0 al 4: Lemma=arma, POS=noun..."

CALCOLO DEGLI OFFSET

Il calcolo degli **offset** è l'operazione tecnica che permette all'annotazione **Stand-off** di funzionare: è il "sistema di coordinate" che collega l'analisi al testo.

1. L'indice parte da Zero (0-based indexing): Nella maggior parte dei sistemi informatici e dei linguaggi di programmazione, il conteggio non parte da 1, ma da **0**. Il primo carattere del file è sempre in posizione 0.

2. Gli spazi sono caratteri a tutti gli effetti: Ogni spazio vuoto occupa una posizione numerica. Ignorare gli spazi è l'errore più comune nel calcolo manuale degli offset.

3. Logica dell'intervallo [Inizio, Fine): L'offset viene solitamente espresso come una coppia di numeri:

Start (Incluso): La posizione del primo carattere della parola.

End (Escluso): La posizione del carattere *immediatamente successivo* alla fine della parola (o la "staccionata" dopo l'ultima lettera).

Carattere	A	r	m	a	_	v	i	r	u	m	q	u	e	_	c	a	n	o
Indice	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

ESEMPIO STAND-OFF XML

```
<?xml version="1.0" encoding="UTF-8"?>
<standoff>
  <!-- I target puntano agli offset dei caratteri nel file di testo -->
  <entry target="eneide.txt#char=0,4" lemma="arma" pos="NOUN" case="ACC" gender="NEUT" number="PLUR"/>
  <entry target="eneide.txt#char=5,10" lemma="vir" pos="NOUN" case="ACC" gender="MASC" number="SING"/>
  <entry target="eneide.txt#char=10,13" lemma="que" pos="CCONJ"/>
  <entry target="eneide.txt#char=15,19" lemma="cano" pos="VERB" mood="IND" tense="PRES" person="1" number="SING"/>
  <entry target="eneide.txt#char=21,27" lemma="Troia" pos="PROPN" case="GEN" gender="FEM" number="SING"/>
  <entry target="eneide.txt#char=28,31" lemma="qui" pos="PRON" case="NOM" gender="MASC" number="SING"/>
  <entry target="eneide.txt#char=32,38" lemma="primus" pos="ADJ" case="NOM" gender="MASC" number="SING"/>
  <entry target="eneide.txt#char=39,41" lemma="ab" pos="ADP"/>
  <entry target="eneide.txt#char=42,46" lemma="ora" pos="NOUN" case="ABL" gender="FEM" number="PLUR"/>
</standoff>
```

ESEMPIO CONLL-U

Questo è lo standard internazionale per le *Treebanks* (Universal Dependencies). Ogni riga rappresenta un "token" e le colonne definiscono morfologia e relazioni sintattiche.

ORDINE COLONNE CONLL-U:

ID: Indice della parola.

FORM: Parola così come appare nel testo.

LEMMA: Forma base (vocabolario).

UPOS: Parte del discorso universale (NOUN, VERB, ecc.).

XPOS: Tag specifico per la lingua (opzionale).

FEATS: Tratti morfologici (Caso, Genere, Numero, ecc.).

HEAD: ID della parola "reggente" (0 se è il nodo principale/Root).

DEPREL: Relazione sintattica (obj = oggetto, nmod = modificatore nominale, ecc.).

```
# sent_id = 1
# text = Arma virumque cano, Troiae qui primus ab oris.
1> Arma> arma> NOUN> n-s---fn-> Case=Nom|Gender=Fem|Number=Sing> 0> root> _> TokenRange=0:4
2> virumque> virumque> NOUN> d-----> _> 1> amod> _> TokenRange=5:13
3> cano> canus> NOUN> n-s---mb-> Case=Abl|Gender=Neut|Number=Sing> 1> nmod> _> SpaceAfter=No|TokenRange=14:18
4> ,> ,> PUNCT> u-----> _> 5> punct> _> TokenRange=18:19
5> Troiae> Troia> PROPN> n-s---fg-> Case=Gen|Gender=Fem|Number=Sing> 3> nmod> _> TokenRange=20:26
6> qui> qui> PRON> p-s---mn-> Case=Nom|Gender=Masc|Number=Sing|PronType=Rel> 7> nsubj> _> TokenRange=27:30
7> primus> primus> ADJ> a-s---mn-> Case=Nom|Gender=Masc|Number=Sing> 1> conj> _> TokenRange=31:37
8> ab> ab> ADP> r-----> _> 9> case> _> TokenRange=38:40
9> oris> os> NOUN> n-s---ng-> Case=Abl|Gender=Neut|Number=Plur> 7> nmod> _> SpaceAfter=No|TokenRange=41:45
10> .> .> PUNCT> u-----> _> 1> punct> _> TokenRange=45:46
```

ESEMPIO IOB/BIO

Il formato **BIO** (noto anche come **IOB**, *Inside-Outside-Beginning*) è lo standard di riferimento per il **Named Entity Recognition (NER)**. A differenza dei formati sintattici, il suo unico scopo è identificare e classificare le "entità nominate" (nomi di persone, luoghi, organizzazioni, date) all'interno di una sequenza di testo.

Il formato BIO risolve il problema delle entità composte da più parole (es. *Caius Iulius Caesar*).

B (Beginning): Indica che il token è la prima parola di un'entità.

I (Inside): Indica che il token è una parte successiva di un'entità iniziata con un tag B.

O (Outside): Indica che il token non appartiene a nessuna entità.

Esempio con entità

multipla:

Caius -> **B-PER**

Iulius -> **I-PER**

Caesar -> **I-PER**

vicit -> **O**



**CORPUS LINGUISTICS E
ANALYTICS PER IL LICEO
CLASSICO**

L'EVIDENZA EMPIRICA

L'introduzione della Corpus Linguistics al liceo sposta l'accento dall'intuizione soggettiva, o la dichiarazione del manuale all'**evidenza dei dati**.

Intuizione: "Mi sembra che Cicerone usi molto questa parola."

Corpus Analytics: "Cicerone usa questo termine 45 volte ogni 10.000 parole, con una frequenza del 300% superiore rispetto a Cesare.»

Quali sono i pilastri dell'analisi dei corpora?

LE LISTE DI FREQUENZA (FREQUENCY LISTS)

Definizione: Elenco di tutte le parole (token) o lemmi presenti in un testo, ordinati per numero di occorrenze.

Obiettivo Didattico: Passare da uno studio del lessico "astratto" (ordine alfabetico del dizionario) a uno studio "strategico" (priorità statistica dell'autore).

Applicazione in Classe:

- **Pre-lettura:** Analizzare i primi 20 lemmi più frequenti di un'opera per ipotizzarne i temi portanti prima di tradurre.
- **Profilo dell'Autore:** Confrontare i lemmi dominanti in Cesare (es. *castra, flumen, milites*) vs Cicerone (es. *res publica, consul, senatus*).

Esempio Celebre: Scoprire che nel IV libro dell'Eneide la parola *fatum* ha una frequenza anomala rispetto agli altri libri; questo dato permette di avviare una discussione sul tema del destino e della tragedia di Didone basandosi su un dato oggettivo.

LE CONCORDANZE (KWIC- KEY WORD IN CONTEXT)

Visualizzazione di una parola (nodo) posta al centro di una lista di righe, con una porzione di testo a sinistra e a destra (*Key Word In Context*).

Obiettivo Didattico: Analizzare il "comportamento" di una parola nel suo habitat naturale per indurre la regola grammaticale o l'uso semantico.

Applicazione in Classe:

- **Sintassi induttiva:** Invece di imparare a memoria che *utor* regge l'ablativo, lo studente osserva 20 righe di concordanze e deduce la reggenza.
- **Analisi dei sinonimi:** Cercare *metus* e *timor* e osservare se compaiono in contesti sintattici o narrativi differenti.
- **Oltre il dizionario:** Superare l'ansia della "traduzione unica" vedendo come la stessa parola richieda sfumature diverse a seconda dei vicini di destra e sinistra.

LE COLLOCAZIONI (COLLOCATIONS)

La tendenza di due o più parole a comparire vicine più spesso di quanto ci si aspetterebbe per puro caso (parole che "si cercano").

Obiettivo Didattico: Comprendere la **prosodia semantica** e l'ideologia dell'autore (come l'autore "colora" i concetti).

Esempio Pratico:

- **Virgilio:** Quali aggettivi "collocano" con *fatum*? Spesso termini legati alla durezza o all'ineluttabilità (*acerbum, immite*).
- **Omero:** Lo studio degli epiteti fissi (es. *Achille piè veloce*) è, tecnicamente, uno studio di collocazioni cristallizzate.

Applicazione in Classe: Indagare il pregiudizio dell'autore. Esempio: quali parole compaiono vicino a *mulier* in un autore misogino rispetto a uno elegiaco? La statistica rivela il giudizio culturale implicito.

ANALISI DEL LATINO CON VOYANT TOOLS

Preparazione del Testo

Prima di iniziare, serve un testo "pulito" (formato .txt).

Dove trovarlo: Vai su <https://latin.packhum.org/> o <https://www.thelatinlibrary.com/>.

Cosa fare: Copia il testo (es. il primo libro delle *Catilinarie* di Cicerone o il IV dell'*Eneide*) e incollalo in un file Note, rimuovendo introduzioni moderne o note a piè di pagina. La pulizia del testo è importante.

Caricamento su Voyant

Vai su voyant-tools.org.

Incolla il testo nella grande casella bianca oppure clicca su **Upload** per caricare il tuo file .txt.

Clicca su **Reveal**.

CONFIGURAZIONE

Configurazione per il Latino (Fondamentale!)

Di default, Voyant potrebbe includere nelle statistiche parole comuni come *et, in, est*, che oscurano i termini significativi.

Gestione Stopwords: 1. In alto a destra in ogni widget (es. Cirrus), passa il mouse e clicca sull'icona delle opzioni (il cursore). 2. Sotto "Stopwords", seleziona **Italian** (spesso funziona meglio del 'None') o meglio ancora, crea una "**New List**" e incolla le parole latine più frequenti che vuoi escludere (es. *et, non, in, ad, ut, est, que*). 3. Clicca su **Save**.

ESPLORAZIONE DEL TESTO CON I WIDGET

A. Cirrus (La Nuvola di Parole) Mostra i termini più frequenti. Più grande è la parola, più alta è la frequenza.

Uso didattico: Chiedi ai ragazzi: "Sulla base di queste parole giganti, di cosa parla questo libro?". È un ottimo esercizio di *pre-reading*.

B. Trends (Andamento Temporale) Mostra come la frequenza di una parola cambia dall'inizio alla fine del testo.

Esempio: Caricando tutta l'Eneide, cerca *Dido* e *Aeneas*. Vedrai dei picchi in corrispondenza del IV libro e un calo drastico di *Didone* successivamente. È la "mappa del calore" della narrazione.

C. Links (Rete di Relazioni) Mostra quali parole compaiono spesso vicine.

Esempio: Cliccando su *fatum*, vedrai a quali verbi o aggettivi è collegato. È un modo visivo per spiegare le **collocazioni** e l'ideologia dell'autore.

D. Contexts (Concordanze / KWIC) In basso a destra, trovi il testo con la parola cercata al centro.

Uso didattico: Invece di cercare sul dizionario, guarda 20 esempi d'uso. Come cambia il caso? Che complementi regge quel verbo?

CENNI DI STATISTICA PER L'ANALISI DEL TESTO

La statistica solitamente studia fenomeni collettivi, sociali e di massa attraverso l'osservazione del dato.

Statistica Descrittiva: È una statistica di sintesi che descrive o riassume quantitativamente le caratteristiche di una raccolta di informazioni. Il suo obiettivo è riassumere un campione (es. quante volte compare "pietas" nell'Eneide?).

Statistica Inferenziale (o Induttiva): È il processo che utilizza l'analisi dei dati per dedurre le proprietà di una popolazione più ampia o di una distribuzione di probabilità sottostante, ad esempio testando ipotesi e derivando stime (es. partendo da un campione di 10 orazioni, cosa possiamo dire dello stile generale di Cicerone?).

Non possiamo analizzare *tutte* le manifestazioni di una lingua, quindi lavoriamo su **Campioni** rappresentativi.

LE VARIABILI

Nel testo antico, non tutto è un numero. Dobbiamo distinguere:

Variabili Qualitative (Nominali): Categorie non numeriche (es. Parti del discorso: Nome, Verbo, Aggettivo).

Variabili Ordinali: Categorie con un ordine intrinseco (es. Livello di salute in un questionario: scarso, buono, eccellente).

Variabili Quantitative: Numeri veri e propri (es. numero di versi per Canto, numero di sillabe).

PROBABILITÀ E FREQUENZA

Per analizzare un testo scientificamente, dobbiamo definire come misuriamo la presenza dei fenomeni linguistici:

Definizione Statistica di Probabilità: Nel lungo periodo, è il rapporto m/n , dove m è il numero di esiti in cui si verifica l'evento A , e n è il numero totale di esiti dell'esperimento.

In Linguistica: È il rapporto tra il numero di occorrenze di un'unità (parola, lemma, PoS, ecc.) e il numero totale di osservazioni nel corpus.

Frequenza Relativa (f_r): È il valore del rapporto tra occorrenze osservate e totale delle osservazioni.

Probabilità Condizionata $p(B | A)$: Esprime la probabilità che si verifichi l'evento B (es. un NOUN), dato che l'evento A (es. un ART) è già avvenuto.

Applicazione pratica: È lo strumento fondamentale per calcolare la probabilità di una parola data la precedente (modelli n-grammi). Rappresenta la base logica dei moderni suggeritori testuali e dei modelli di linguaggio (LLM).

LA MEDIA ARITMETICA

La media è la somma dei valori divisa per il numero di osservazioni. Ma è sempre utile?

Esempio Racine: L'aggettivo *heureux* compare 143 volte in 9 tragedie. La media è 15,89.

Il problema: Se le opere hanno dimensioni molto diverse, la media "astratta" non è significativa.

Confronto:

- **Dante (Commedia):** La media dei versi per Canto è molto stabile (~142). Dato **significativo**.
- **Ariosto (Orlando Furioso):** La lunghezza dei canti varia enormemente (da 500 a 1500 versi). La media qui **non è significativa** perché non rappresenta un "Canto tipico".

MEDIANA E MODA

Quando la media "inganna", usiamo altri indici:

Mediana: Il valore che divide a metà il campione. Non è influenzata da valori estremi (outliers).

- *Esempio:* In una classe dove tutti prendono 6 e uno prende 10, la mediana resta 6, mentre la media si alza artificialmente.

Moda: Il valore che appare più spesso. In linguistica, è la parola più frequente del testo.

LA LEGGE DI ZIPF

Il linguaggio naturale segue una distribuzione particolare, diversa da quella degli esseri umani (altezza, peso).

La Regola: La frequenza di una parola è inversamente proporzionale al suo rango (posizione in classifica).

=> Poche parole (le "parole vuote" come *et, in, non*) compaiono tantissimo; tantissime parole compaiono una volta sola (*Hapax*).

Principio del minimo sforzo: Zipf ipotizzò che né il parlante né l'ascoltatore vogliano faticare più del necessario. Questa legge vale per il Latino, il Greco, l'Inglese e persino per i post su Instagram. È una costante del cervello umano.