

# Digital Humanities

## 3. LA CODIFICA DEL TESTO, XML E LA TEI

### 3.1 INTRODUZIONE

I testi e i documenti nelle varie forme materiali e immateriali sono tanto principali oggetti quanto i principali strumenti di studio di buona parte delle discipline umanistiche.

La complessità e plurivocità del concetto di testo tipiche dell'ambito umanistico hanno richiesto un supplemento di analisi teorica e progettazione di linguaggi e modelli rappresentazionali.

Il testo è un oggetto dall'antologia complessa e in quanto tale veicola o costruisce significato su più livelli, attraverso l'instaurazione di più relazioni tra tali. Quando parliamo di rappresentazione digitale ci riferiamo a un insieme di strutture dati e linguaggi formali organizzati in una gerarchia di livelli *isomorfi*.

### 3.2 ASPETTI TEORICI E METODOLOGICI DELLA CODIFICA TESTUALE

Tre dimensioni concettuali: dimensione *semiotica*, *ontologica* ed *epistemologica*. La prima deriva dal fatto che la codifica mette in gioco due sistemi rappresentazionali eterogenei. La natura dei processi computazionali pone dei vincoli sulla forma della rappresentazione e dei linguaggi con cui può essere condotta.

Linguaggi formali costituiti da:

1. un alfabeto finito e definito di simboli atomici;
2. un insieme finito di regole sintattiche che consentono di concatenare i simboli atomici per generare espressioni complesse.

Un automa accetta ed elabora solo espressioni ben formate.

La struttura, complessità e l'espressività dei formalismi informatici può variare lungo una scala che ha come unità di misura il grado di astrazione, o virtualizzazione, rispetto ai processi fisici che avvengono a livello hardware.

La codifica del testo si caratterizza come un processo di "formalizzazione"; richiede la scelta o costruzione del linguaggio formale che meglio risponde alle esigenze dell'elaborazione testuale.

È possibile individuare un insieme finito di caratteristiche testuali per affermare che un oggetto digitale sia la rappresentazione corretta dell'oggetto testuale ma → il testo è ciò che permane in ogni operazione di riproduzione materiale della sequenza di simboli grafici.

Questa caratterizzazione fornisce un criterio di individuazione → per affermare che il testo rinvenuto in un esemplare è esattamente quel testo.

Ogni discorso teorico relativo ai testi usa nelle sue spiegazioni termini come es. paragrafo, capitolo ecc. → questi oggetti testuali debbono essere indipendentemente dalle teorie stesse. Questi elementi sono gerarchici poiché ogni sottoinsieme di tali oggetti ha una relazione parte-intero con un oggetto testuale che non ne fa parte, e ordinati in quanto insieme per ogni sottoinsieme di oggetti testuali. (teoria OHCO)

Nella teoria OHCO la definizione della nozione di genere testuale si rivela sfuggente.

In realtà non esiste alcun accordo tra specialisti su cosa sia un genere testuale e quanti ne esistano: è una teoria dipendente dal punto di vista che si assume sulla nozione di genere.

L'introduzione del concetto di dipendenza ha delle conseguenze teoriche → è determinante nella definizione dei generi di testo e delle caratteristiche; non esistono ragioni teoriche per non applicarlo alla determinazione dell'insieme di caratteristiche di un singolo tipo di testo.

La pratica comune è quella di definire l'oggetto a partire dal punto di vista interna della disciplina stessa → anche se ammettiamo che il concetto di testo come invariante allografica o OHCO fornisca una descrizione corretta, dobbiamo osservare come essa sia applicabile solo dopo aver operato una scelta tra i livelli descrittivi del testo.

Alcuni studiosi hanno sostenuto che la codifica del testo dovrebbe prescindere da qualsiasi assunto teorico o atteggiamento interpretativo; tuttavia, è una posizione inattuabile.

Un discorso simile potrebbe essere effettuato per ognuno dei componenti testuali della teoria OHCO → prevede la presenza di una struttura gerarchica di oggetti riconoscibili perché connessi in modo casuale con alcune caratteristiche fisiche. Tale identificazione può essere controversa es. spazi bianchi che definiscono i paragrafi = può non essere così. Si aggiungerebbero a prescindere degli elementi interpretativi.

La codifica di un testo si colloca interamente all'interno del processo analitico-interpretativo; il linguaggio di codifica può essere inteso come uno speciale linguaggio teorico usato per formalizzare teorie o modelli dei fenomeni testuali d'interesse. → il problema della codifica è individuare e sviluppare un linguaggio abbastanza potente da permettere a ogni studioso di rappresentare le caratteristiche testuali interessate e le loro interpretazioni.

La rappresentazione di un dato testo dipende dalle assunzioni teoriche e dalle interpretazioni che vengono fatte da uno studioso; a volte possono non essere condivise.

### 3.3 I REQUISITI DEI LINGUAGGI DI CODIFICA E LA PRESERVAZIONE DIGITALE

Problemi → requisiti con i quali valutare la minore o maggiore adeguatezza tecnica e sostenibilità di una tecnologia informatica per la produzione di risorse testuali digitali.

I requisiti generali consentono di valutare i rapporti costi/benefici di una certa soluzione in modo più appropriato.

Vanno considerate le caratteristiche logiche che forniscono a un linguaggio di codifica un'adeguata capacità rappresentazionale. Codificare un testo, la conoscenza e le interferenze che un interprete umano formula su quel testo, dev'essere in grado di rappresentare i fenomeni testuali esplicitati. Uno schema di codifica dovrebbe salvaguardare quanto più possibile due proprietà fondamentali di una risorsa informativa digitale:

1. accessibilità tecnica e riusabilità → un oggetto informativo deve opporre il minor numero di vincoli tecnici possibile per consentire l'accesso ai contenuti informativi che veicola e il riuso.
2. preservabilità a lungo termine → gli strumenti che consentono l'accesso agli oggetti informativi sono soggetti a un'obsolescenza rapida; le cause sono:
  - a. degrado fisico dei supporti
  - b. obsolescenza tecnica dell'hardware
  - c. obsolescenza tecnica del software

Se la mediazione strumentale diventa troppo complessa e soggetti a vincoli è accessibile solo agli utenti dotati della catena strumentale necessaria.

Due sono gli aspetti rilevanti al fine di massimizzare il grado di accessibilità e permanenza di una risorsa digitale:

1. standardizzazione → insieme di norme di progettazione e uso relative a una particolare tecnologia che vengono emesse da un ente istituzionale nazionale o internazionale.
2. portabilità tecnica → si suddivide in tre livelli:

- a. indipendenza dall'hardware
- b. indipendenza dal software
- c. indipendenza logica da particolari tipologie di trattamento

i metadati svolgono un ruolo chiave nell'identificazione, utilizzo e preservazione delle risorse digitali; vanno espressi mediante specifici linguaggi e possono essere messi in correlazione con i linguaggi di modifica mediante i quali viene rappresentato il documento primario.

Ogni progetto di archiviazione su supporto digitale del patrimonio testuale con finalità scientifiche o di conservazione deve misurarsi con queste esigenze.

### 3.4 I LINGUAGGI PER LA RAPPRESENTAZIONE DIGITALE DEL TESTO

I linguaggi per la codifica digitali dei testi derivano in gran parte da applicazioni di *text processing* sviluppate dall'industria informatica. Rappresentano il "grado zero" della codifica testuale.

Il documento elettronico è costituito da un flusso di caratteri: unità atomica per la rappresentazione, organizzazione e controllo di dati testuali sull'elaboratore.

La codifica dei caratteri non esaurisce i problemi di rappresentazione delle caratteristiche di un testo; sono stati sviluppati sistemi in grado di arricchire la rappresentazione digitale di un testo. Sono nati programmi che nascondono le informazioni all'utente ma non allo sviluppatore → poiché hanno lo scopo di ottenere una rappresentazione virtuale più possibile mimetica del testo a stampa, gli applicativi grafici interattivi vengono anche definiti sistemi WYSIWYG (*what you see is what you get*).

L'inaccessibilità del documento senza la mediazione del software con cui è stato creato impone limiti alla portabilità delle risorse informative.

#### 3.4.1 I SISTEMI DI CODIFICA DEI CARATTERI

Si stabilisce un'associazione biunivoca tra elementi di una collezione di simboli distinti e un insieme di codici numerici; l'insieme viene denominato "*coded character set*"; per ciascuno di essi si definisce una codifica dei caratteri.

Il numero di caratteri rappresentabili è determinato dal numero di cifre binarie utilizzate per codificare ciascun carattere.

La rappresentazione di un sistema di scrittura di questo genere è molto complessa: accanto alle lettere vanno codificati tutti i segni di punteggiatura.

A ogni standard corrisponde una famiglia di *coded character set*.

Il più antico standard è l'ISO 646 che adotta una codifica a 7 bit; nel 1968 venne registrato l'ASCII; poi l'ISO 8859 a 8 bit; dagli anni '90 l'ISO 10646/UCS. La più comune codifica per le lingue occidentali è l'UTF.

#### 3.2.2 I LINGUAGGI MARKUP

Si parla di "linguaggi di codifica testuale" solo per i formalismi che consentono la rappresentazione formale di caratteristiche grafiche, strutturali o semantiche di un testo o segmenti di testo.

L'espressione "*markup*" deriva dall'analogia che questi sistemi di codifica hanno con la simbologia e annotazioni inserite da autori. Questi linguaggi consistono in un insieme di istruzioni che vanno inserite all'interno della sequenza di caratteri. Un linguaggio markup è caratterizzato da:

1. un insieme di caratteristiche testuali;
2. un insieme di identificatori simbolici;

3. una correlazione tra identificatori e caratteristiche testuali;
4. una sintassi che regola il modo in cui gli identificatori devono essere inseriti nel testo.

Il linguaggio di markup prescrive solo il numero e la forma degli identificatori e il loro corretto utilizzo sotto forma di marcatori; deve avere anche una semantica associata agli identificatori.

La sintassi può essere più o meno restrittiva → diviene la proiezione sintattica delle relazioni strutturali che sussistono tra le sequenze di testo verbale cui ogni caratteristica è associata.

I linguaggi di markup si distinguono anche in base alle loro caratteristiche semantiche:

1. procedurali → specifica i processi computazionali;
2. dichiarativi → indica l'esistenza di una data caratteristica o qualità in una data porzione di testo.

La distinzione che riguarda la tipologia di fenomeni testuali che un dato linguaggio rappresenta:

1. presentazionali → le caratteristiche codificate sono strutture di grafiche o formattazione;
2. analitici o descrittivi → strutture astratte o logiche.

### 3.4.3 EXTENSIBLE MARKUP LANGUAGE (XML)

Deriva dal progenitore SGML ed è progettato per sostituire l'HTML per la creazione di documenti da pubblicare sul web.

È uno standard pubblico, progettato per essere "indipendente" dai sistemi informativi, dai supporti e dispositivi digitali; fornisce le maggiori garanzie di preservazione dei dati nei progetti di archiviazione digitale a lungo termine di dati sensibili.

Si basa su un paradigma di markup dichiarativo che punta a descrivere la struttura astratta di un documento piuttosto che il suo aspetto grafico e consente la dichiarazione dell'insieme di elementi che lo compongono e le loro relazioni. È un "metalinguaggio" che fornisce le regole per realizzare un numero indefinito di linguaggi di codifica.

Ogni elemento deve comparire in un punto preciso della struttura del documento → presuppone la descrizione di una "struttura" del documento e questa dev'essere rigorosamente gerarchica, ad albero. Ne consegue che ogni documento XML dovrà avere una sola struttura ad albero.

La struttura logica prevista dallo schema viene rappresentata nel documento usando una "notazione parentetica annidata".

La presenza dell'insieme di vincoli formali che sovrintendono alla creazione dei documenti XML permette di applicare loro un vero e proprio processo automatico di verifica della conformità sintattica → *parsing*.

### 3.5 LA TEI: UN LINGUAGGIO DI MARKUP PER LE SCIENZE UMANE

XML lascia la massima libertà nel costruire uno specifico linguaggio di codifica, ma questa si scontra con le difficoltà di realizzazione di una simile operazione.

TEI → iniziativa internazionale diventa un punto di riferimento per chiunque si occupi di trattamento informatico dei testi in area umanistica. 1988: avviato un progetto internazionale per sviluppare uno schema di codifica che mettesse ordine in questo settore (nato appunto TEI).

I principi che hanno orientato lo sviluppo dello schema di codifica TEI sono basati sui fondamentali teorici dei linguaggi di markup dichiarativo e lo schema prevede anche elementi di tipo presentazionale. Lo schema della TEI è un'ontologia astratta del testo e delle sue caratteristiche.

L'insieme di caratteristiche testuali che costituiscono l'ontologia della TEI si articola su tre livelli:

1. caratteristiche testuali considerate come universalmente valide per tutti i documenti;
2. strutture testuali proprie di un limitato numero di sottoclassi testuali individuate in base a macrogeneri;
3. proprietà e caratteristiche testuali prodotte da singole prospettive analitiche.

## 4. L'ANALISI DEL TESTO

### 4.1 INTRODUZIONE

I metodi e gli strumenti per l'analisi computazionale del testo hanno rappresentato uno dei pilastri fondamentali delle Digital Humanities.

L'analisi del testo ha una forte contiguità la linguistica computazionale e la linguistica dei *corpora*. Questi domini condividono fondamenti e strumenti operativi: ciò che cambia è il ruolo attribuito nell'indagine dei testi.

La rilevanza dell'analisi del testo non si limita agli studi letterali e linguistici, ma in prima istanza sembrerebbe che questi sarebbero il suo terreno di elezione.

Il fondamento dell'analisi computazionale del testo consiste nella considerazione quantitativa del linguaggio e dei testi. analizzare computazionalmente i testi significa in gran parte contare le parole in modi più o meno complicati. Ridurre i testi a dati da misurare e analizzare con metodi statistici o simili, appare quanto di più remoto dalla pratica ermeneutica del critico letterario.

### 4.2 CONTARE LE PAROLE

La sociologia della letteratura e la storia del libro da tempo applicano metodi quantitativi per studiare la vita sociale dei testi.

Se si vuole adottare un approccio quantitativo all'analisi dei testi da un punto di vista interno, l'unica componente testuale immediatamente accessibile sono le parole di cui è composto, il livello "superficiale" del discorso.

Poiché le parole si articolano in classi di vario genere, questi conteggi ci forniscono la distribuzione statistica di tali classi di ricorrenza.

Misurando le distribuzioni di frequenza delle classi di parole possiamo applicare vari metodi statistici e probabilistici di elaborazione dei dati per cercare di capire che cosa queste misurazioni ci possono dire sui testi.

Il più basilare modo per raggruppare le parole in classi è quello di considerare una parola nella forma in cui essa appare scritta nel testo, a un livello di descrizione che è perfino più basso di quello lessicale, dove una parola è intesa come unità lessematica.

Il prodotto dell'enumerazione è una distribuzione di frequenze.

La prima tipologia di elaborazione è la "media aritmetica". Un altro modo di calcolare la tendenza centrale è quello della mediana.

Una volta calcolate le misure della tendenza si calcola variabilità/ dispersione → deviazione standard.

I metodi descritti permettono di avventurarsi nell'analisi dello stile di un testo o un autore.

L'analisi dei dati e la sua evoluzione vero il *text mining* e *machine learning*, provvedono metodi di studio dei testi più potenti.

L'analisi quantitativa su vasta scala richiede tuttavia l'adozione di modalità di rappresentazione dei testi come dati numerici più sofisticate. La strategia più comune consiste nella "rappresentazione vettoriale dei testi".

Selezionando le parole uniche si costruisce una lista ordinata → dizionario. Ogni elemento dei vettori contiene il conteggio di frequenza della parola corrispondente nel dizionario.

Ogni testo sarà rappresentato da un vettore i cui valori sono le frequenze assolute delle parole nell'ordine in cui sono elencate nel dizionario.

### 4.3 DALLA STILOMETRIA CLASSICA AL DISTANT READING

L'uso di metodi di inferenza statistica è il fondamento matematico dell'analisi testuale, a partire dai primi esperimenti in quella che viene denominata "stilometria computazionale classica".

Se questi studi hanno avuto scarsa penetrazione nella critica e nella storiografia letteraria, a partire dall'inizio del millennio si è assistito a un rovesciamento della condizione di subalternità dei metodi computazionali nell'analisi del testo.

Queste hanno reso disponibili vasti archivi e *corpora* di testi letterari in molte lingue → depositi di *big data* culturali.

La spinta principale è dovuta a un cambio del paradigma teorico → dal *distant reading* di Moretti. È una concezione sviluppata in antitesi al *close reading* → esistono fatti e fenomeni letterari e culturali, sia sincronici che diacronici, che non sono accessibili ai tradizionali metodi di lettura profonda e di interpretazione di poche opere, ma che richiedono l'analisi massiva di centinaia di testi e documenti.

Queste tecniche, applicate a grandi insiemi di documenti digitali o di metadati, permettono di far emergere strutture, regolarità e pattern altrimenti non riconoscibili.

### 4.4 METODI COMPUTAZIONALI PER L'ANALISI DEL TESTO: UNA RASSEGNA

L'affermarsi del paradigma del *distant reading* è stato allo stesso tempo determinante e determinato dall'adozione di nuovi e più avanzati metodi di analisi dei testi, sviluppati nell'ambito della *data analytics* e di *machine learning*.

### 4.5 TOPIC MODELING

La tecnica del *topic modeling* → da un lato la logica di funzionamento è estremamente complessa, dall'altro i risultati appaiono altrettanto intuitivi.

Alla base del *topic modeling* si collocano algoritmi come LDA, che sfruttano le cooccorrenze di parole in uno o più documenti per valutare la probabilità che esse appartengano a uno o più *topic*.

I *topic* non sono altro che raggruppamenti astratti di probabilità di parole, tutte le parole che possano appartenere a tutti i *topic*, i quali a loro volta possono comparire in tutti i documenti.

Al termine di un processo iterativo che rassegna di volta in volta le probabilità, cercando di raggiungere uno stato di equilibrio in cui più nulla può cambiare, i *topic* emergono e se più di due volte le parole compaiono assieme in diversi documenti, più alta è la probabilità che appartengano allo stesso *topic*.

Trattandosi di un processo stocastico, i risultati cambieranno a ogni nuova ripetizione dell'analisi; in aggiunta diversi parametri devono essere definiti prima ancora di iniziare la procedura.

Sul piano dell'analisi permette di far emergere pattern di larga scala da ampie quantità di documenti.

Sul piano pratico non richiede un particolare lavoro di annotazione o preparazione dei documenti: può essere eseguito rapidamente su qualsiasi *corpus* testuale.

Nell'ambito degli studi letterari è stato usato sia per l'esplorazione di vasti archivi testuali che per la discussione di sofisticati problemi teorici.

I *topic modeling* hanno messo in evidenza la gradualità dei cambiamenti, che assai di rado mostrano variazioni drastiche nel tempo.

Si adattano idealmente allo studio di testi argomentativi.

Problema → può far emergere pattern anche dove non esistono. Sul piano teorico: in confronto serrato con la filosofia foucaultiana e con le tradizioni critiche di linguistica funzionalistica, critica tematica,

strutturalismo e semiologia si è dimostrato che un corrispettivo diretto del *topic modeling* non esista negli studi letterari.

Un lavoro necessario nelle Digital Humanities è la risoluzione di tali incompatibilità per lavorare alla creazione di un framework condiviso.

#### 4.6 WORD EMBEDDINGS

Così come i *topic* emergevano dalla presenza delle stesse parole in documenti diversi, così gli *embeddings* dipendono dalle diverse “compagnie” frequentate da una parola.

Mentre i *topic* sono riferibili a interi documenti gli *embeddings* rappresentano la semantica delle singole parole del vocabolario e si presentano come serie di numeri.

L'implementazione più celebre → *Word2vec* → era stato pensato in due varianti principali, entrambe con al centro una rete neurale e uno pseudo-processo di machine learning.

Nell'implementazione le reti neurali sono chiamate a svolgere due compiti all'apparenza impossibili:

1. data una parola in un testo, prevedere tutte le parole che la circondano;
2. dato quello stesso gruppo di parole, prevedere la parola che si trova al centro.

Ma questo non è l'obiettivo dell'algoritmo. *Word2vec* esegue questo compito per un certo numero di iterazioni e alla fine si disinteressa del risultato, selezionando l'ultimo stato della rete neurale allenata per generarlo.

Le parole semantiche vicine saranno anche rappresentate da vettori molto vicini tra loro; in aggiunta, vere e proprie operazioni matematiche potranno essere compiute sui significati delle parole.

Questi studi si soffermano più sulla valutazione dei metodi che sull'analisi dei risultati. → la metodologia è stata solo recentemente adottata dalle Digital Humanities.

Gli aspetti critici dei Word Embeddings sono molteplici e la novità del metodo suggerisce cautela nell'applicarlo indiscriminatamente. I risultati possono variare a ogni ripetizione dell'analisi e l'ordine delle parole, la struttura delle frasi non sono presi in considerazione.

Il lavoro sulle cooccorrenze di parole genera almeno due tipi di problemi: si riscontrano difficoltà nel cogliere differenze di significato valido sul piano della percezione: queste metodologie mancano della multimodalità che determina il significato delle parole nella mente umana.

L'utilizzo delle co-occorrenze per modellizzare la somiglianza tra parole può generare numerosi effetti di disturbo.

Nonostante le criticità, le potenzialità di questi metodi per l'analisi computazionale del testo sono innegabili.

#### 4.7 SENTIMENT ANALYSIS

La sentiment analysis è recentemente diventata uno degli argomenti più discussi delle Digital Humanities. Nell'introdurre l'argomento agli studi letterari, *Jockers* non ha nascosto l'ambizione di fornire un contributo diretto alla narratologia → utilizzando un *tool* il *plot arc* di un romanzo veniva prodotto automaticamente. Sul piano delle ascisse era rappresentato lo svolgersi dell'intreccio, in quello delle ordinate era il “sentiment”.

In questa proposta gli archetipi avrebbero dovuto rappresentare le forme base di tutte le narrazioni nella nostra tradizione letteraria.

Tramite il fenomeno noto come *social reading*, il web ha iniziato a fornire un'enorme quantità di dati sull'esperienza della letteratura → *Whattpad*: un'analisi comparata del “sentiment” di testo e



commenti, proponendo una forma di *scalable reading*, che sfrutta la sentiment analysis per individuare i passaggi di testo capaci di suscitare reazioni contrastanti o peculiariintonie emotive tra lettori.

I sistemi di sentiment analysis più semplici si limitano a considerare la sola valenza, intesa come generica positività/negatività delle emozioni espresse dal testo. Più complesse sono le rappresentazioni discrete, che cercano di individuare ed isolare gruppo di “emozioni bailari”.

Al cuore della sentiment analysis poi si collocano le “risorse emotive” che associano una serie di valori numerici a una selezione di parole o frasi.

Le risorse focalizzate su frasi o porzioni di testo richiedono complessi lavori di annotazione manuale e non sono facilmente reperibili.

#### 4.8 STILOMETRIA

Attraverso analisi statistiche degli usi linguistici, la stilometria tenta di “misurare” lo stile degli autori, discernendo così le loro “impronte autoriali” latenti.

L’affermazione definitiva di questo campo di studi risale alla fine del XX secolo, quando venne proposto un metodo per l’attribuzione dell’autorialità dei testi → distanza Delta.

Data una collezione di testi digitalizzati:

1. viene generata una lista contenente le parole più frequenti nell’intera collezione;
2. per ciascun testo viene misurata la frequenza di utilizzo delle parole che compongono la lista;
3. la “distanza” tra i testi è calcolata confrontando le diverse liste di frequenza tramite una formula *ad hoc*.

Burrows ha testato questo metodo su un *corpus* di poeti della restaurazione inglese → i testi più vicini tra loro erano quelli scritti dagli stessi autori.

È stato suggerito come l’efficacia di questo metodo sia determinata dalle scelte inconsce che ogni scrittore compie quando seleziona le parole più frequenti nel suo vocabolario.

La gamma delle metodologie applicate varia, includendo approcci di machine learning e focalizzandosi su aspetti come punteggiatura, parti del discorso e la struttura della frase.

*Keyness analysis* → si parte da testi già attribuiti per individuarne le caratteristiche ligustiche distintive. Focalizzandosi sulle “parole contenuto”, questi algoritmi possono essere utilizzati sia per un’esplorazione dei *corpora* che per la conferma di attribuzioni.

Approccio “analisi zeta”: dati due gruppi di documenti:

1. ogni testo è suddiviso in un numero di segmenti di uguale lunghezza;
2. per ogni parola contenuta nei due gruppi di documenti viene calcolata la proporzione dei segmenti in cui appare;
3. il valore zeta di una parola si ottiene sottraendo i due valori.

Trattandosi di proporzioni i risultati saranno sempre compresi tra valori -1 (sottorappresentazione di una parola nel primo gruppo di documenti) e +1 (l’esatto opposto).

I metodi stilometrici sono stati applicati con successo a problemi attributivi di alta rilevanza. es. Elena Ferrante → analizzando il testo si è notato come utilizzi un “linguaggio maschile”: si conferma l’ipotesi che possa essere Domenico Starnone.

Difetti → limitazione per la quantità di testo disponibile e analisi su elementi “alti” (costruzioni sintattiche) meno efficaci di quelle basate su elementi “bassi” (parole più frequenti).