**Rivista Internazionale di Tecnica della Traduzione / International Journal of Translation 2021 (23), 219-233**
DOI: 10.13137/2421-6763/33469 ISSN 1722-5906 (print) ISSN 2421-6763 (online)
https://www.openstarts.units.it/dspace/handle/10077/2503

# Sulle tracce dell'espressione dell'interiorità: analisi diacronica di un corpus di narrativa italiana del XIX-XX secolo

Andrea Sciandra
Università di Modena e Reggio Emilia
andrea.sciandra@unimore.it

Matilde Trevisani
Università di Trieste
matilde.trevisani@deams.units.it

Arjuna Tuzzi
Università di Padova
arjuna.tuzzi@unipd.it

Abstract

This article concerns an example of statistical analysis of a large diachronic corpus, including 100 works of Italian fiction, with the aim of recognizing prototypical trends in words frequencies and identifying clusters of words that share similar temporal patterns. Starting from the insights of literary scholars, this study tries to test the hypothesis (often brought forward in literary history) that the inner life of the authors and/or their characters gained more importance during the Nineteenth century and represents one of the cornerstones of modernism. Although this hypothesis cannot be rejected, the adopted statistical methods do not highlight clear-cut temporal patterns: some slight trends emerge but a large share of words individually show an undifferentiated behaviour over time. The idea of studying the occurrence in the texts of the inner expression of the writers from a diachronic perspective is certainly very ambitious. This first experiment shows some strengths of the statistical tools but also some limits of the criteria adopted for both the constitution of the corpus and the selection of linguistic features to be observed.

Keywords

Diachronic corpora, chronological textual data, word embedding, functional data analysis, curve clustering

## 1. Introduzione

Il progetto europeo *Distant Reading for European Literary History* (COST Action CA16204, https://www.distant-reading.net/) è una rete internazionale di studiosi che aspira a sviluppare risorse per leggere la storia della letteratura europea attraverso le lenti dei metodi computazionali. Nell'ambito del progetto sono stati raccolti corpora di testi narrativi di diversi paesi e in diverse lingue con l'obiettivo di studiare diacronicamente fenomeni di interesse linguistico e letterario. In particolare, uno dei gruppi di lavoro del progetto si è concentrato sui problemi della periodizzazione della produzione letteraria con l'idea di verificare l'ipotesi, spesso avanzata nella storia della letteratura europea, secondo cui la rappresentazione della vita interiore dei personaggi abbia assunto maggiore importanza nel corso del XIX secolo e sia diventata uno dei capisaldi del modernismo.

Per poter verificare questa ipotesi attraverso metodi computazionali è necessario innanzitutto tradurre la domanda di ricerca in dati osservabili, cioè in fenomeni che possono essere identificati nei testi con strumenti informatici, e poi adottare strumenti statistici adeguati per la loro analisi e interpretazione.

Per quanto riguarda i fenomeni da isolare per seguirne lo sviluppo nel corso del tempo, il gruppo di lavoro del COST Action ha suggerito due percorsi differenti: (1) ricostruire la costellazione di significati che ruota attorno a sei verbi-seme: *feel*, *think*, *believe*, *know*, *hope*, *wish* e (2) selezionare i verbi più frequenti presenti nel corpus, con particolare riferimento a quelli che descrivono la vita interiore dell'autore e/o dei suoi personaggi.

Il presente lavoro illustra alcuni risultati ottenuti adottando i criteri del primo percorso suggerito con l'obiettivo di offrire un contributo allo sviluppo di metodi quantitativi per la storia della lingua che, come spiega Michele Cortelazzo in questo numero della rivista, rappresenta un campo ancora solo parzialmente esplorato ma sta assumendo una rilevanza specifica e un'identità come filone di ricerca autonomo. Sempre in questo numero della rivista, anche i lavori di Floriana Sciumbata, Paolo Nadalutti e Luca Tringali e di Stefano Ondelli si occupano, da prospettive diverse, del problema dello studio di corpora diacronici e presentano risultati che cercano di dare risposte a questioni emerse nell'ambito dello stesso progetto europeo e discusse durante il convegno *Esperimenti di Distant Reading. Estrazione, analisi e visualizzazione di dati linguistici da corpora letterari*, che si è tenuto all'Università di Trieste – IUSLIT/SSLMIT il 22 ottobre 2021.

Il corpus di testi italiani è stato progettato e raccolto come corpus propedeutico a quello previsto dal progetto COST Action e segue alcune linee di indirizzo del *European Literary Text Collection* (ELTeC): è costituito da 100 opere di narrativa pubblicate nel periodo 1825-1923 da 54 autori, 16 donne e 38 uomini. Con un'ampiezza di oltre sette milioni di occorrenze totali, la raccolta si può considerare un corpus di grandi dimensioni.

Una descrizione del corpus analizzato è disponibile in questo numero della rivista come introduzione al presente articolo e quello di Sciumbata, Nadalutti e Tringali (2021).

Il corpus di testi italiani è stato utilizzato nella versione di base, cioè sotto forma di semplici file di testo indipendenti, ripuliti dagli elementi paratestuali (intestazioni, numeri di pagina, ecc.) e privi di taggatura. Gli strumenti adottati per l'analisi statistica dei dati testuali sono stati sviluppati in ambiente *R* (R Core Team, 2021) sia attraverso procedure scritte ad hoc sia integrando pacchetti esistenti. Nel corso dell'elaborazione è stata creata in *R* una versione lemmatizzata attraverso il pacchetto `udpipe` (Wijffels 2021, Straka e Straková 2017) e la *Italian Stanford Dependency Treebank* (ISDT).

Dopo aver svolto un riconoscimento delle parole presenti nei testi come sequenze di lettere isolate nel testo per mezzo di spazi e punteggiatura (tokenizzazione), in vista dell'analisi statistica il corpus è stato rappresentato mediante matrici di dati del tipo "parole per testi" (*term-document matrix*) che riportano il numero di occorrenze di ciascuna parola in ciascun testo. Inoltre, dato che si tratta di un corpus diacronico, l'occorrenza di una parola è stata osservata con riferimento all'anno di pubblicazione dell'opera (figura 1) ed è diventata, quindi, una sequenza ordinata nel tempo, rappresentabile graficamente come una traiettoria. In questo caso le occorrenze di opere pubblicate nello stesso anno sono state aggregate.

Per rispondere alle domande di ricerca relative alla periodizzazione del corpus preso in analisi, sono state selezionate le traiettorie di un insieme limitato di parole e sono stati adottati strumenti statistici in grado di trovare regolarità utili a leggere l'andamento temporale della loro occorrenza e a costituire gruppi omogenei di parole che condividono un'evoluzione simile.

| | 1826 | 1834 | 1840 | 1849 | 1850 | 1853 | … | 1919 | 1920 | 1923 |
|---|---|---|---|---|---|---|---|---|---|---|
| adesso | 12 | 0 | 24 | 25 | 75 | 2 | … | 10 | 60 | 42 |
| almeno | 82 | 4 | 49 | 15 | 46 | 50 | … | 27 | 45 | 43 |
| amare | 20 | 13 | 163 | 21 | 59 | 48 | … | 86 | 56 | 144 |
| amico | 138 | 7 | 59 | 15 | 194 | 146 | … | 80 | 101 | 72 |
| amore | 43 | 27 | 196 | 23 | 62 | 32 | … | 89 | 59 | 124 |
| ancora | 293 | 24 | 157 | 95 | 228 | 64 | … | 177 | 176 | 245 |
| andare | 1056 | 16 | 234 | 98 | 315 | 332 | … | 267 | 385 | 300 |
| anima | 30 | 11 | 85 | 27 | 82 | 27 | … | 55 | 34 | 7 |
| animo | 137 | 27 | 143 | 21 | 102 | 84 | … | 28 | 28 | 61 |
| aspettare | 148 | 3 | 29 | 34 | 81 | 79 | … | 53 | 97 | 77 |
| assai | 9 | 6 | 70 | 3 | 5 | 80 | … | 6 | 12 | 7 |
| atto | 88 | 13 | 101 | 9 | 42 | 50 | … | 22 | 20 | 47 |
| avere | 3247 | 187 | 1650 | 594 | 1784 | 1696 | … | 1245 | 1371 | 3102 |
| avvenire | 49 | 5 | 36 | 4 | 43 | 38 | … | 24 | 18 | 63 |
| bastare | 111 | 3 | 38 | 6 | 61 | 31 | … | 22 | 41 | 80 |
| bello | 158 | 39 | 120 | 72 | 160 | 116 | … | 100 | 76 | 126 |
| bene | 557 | 15 | 349 | 103 | 340 | 193 | … | 119 | 174 | 245 |
| … | … | … | … | … | … | … | … | … | … | … |

Figura 1 – Esempio di matrice di occorrenze "parole per anno" (estratto)

## 3. Selezione delle parole oggetto di studio con *word embedding*

Il gruppo di ricerca del COST Action ha suggerito di ricostruire l'insieme di parole semanticamente riconducibili ai sei verbi-seme (*feel*, *think*, *believe*, *know*, *hope*, *wish*) tradotti in italiano (*sentire*, *pensare*, *credere*, *sapere*, *sperare*, *desiderare*) e di costruire una lista di parole per le quali rappresentare le occorrenze (normalizzate) su grafici allo scopo di trovare tendenze.

Il metodo suggerito per ricostruire l'insieme di parole collegate ai sei verbi-seme è il *word embedding* (Mikolov et al. 2013). Il *word embedding* è una rappresentazione delle parole di un corpus tramite vettori numerici tale per cui le parole con un significato simile sono rappresentate da vettori simili. Questi vettori identificano quindi le coordinate di ciascuna parola in uno spazio multidimensionale.

Nell'ambito dell'analisi statistica testuale, diverse tecniche si riferiscono a un approccio *bag-of-words* (BoW), che essenzialmente utilizza liste di parole pe-

sate con la loro frequenza. Tra gli svantaggi più importanti del BoW troviamo che l'ordine delle parole viene ignorato e che il BoW generalmente non riesce a cogliere la semantica delle parole. Le tecniche di *word embedding* rappresentano una delle soluzioni più comuni a questi problemi, in quanto generano un numero di dimensioni molto ridotto rispetto al vocabolario di un corpus, ma generalmente in grado di rilevare regolarità linguistiche. I vettori che definiscono le parole possono essere utilizzati anche per identificare analogie come "L'uomo sta alla donna, come il re sta alla regina". Questo risultato si ottiene definendo una finestra di contesto – cioè una finestra di parole che precedono o seguono una parola focale – che viene utilizzata per addestrare un modello di *word embedding*.

In questa applicazione è stata utilizzato il metodo *GloVe* (Pennington, Socher e Manning 2014), uno dei sistemi più comuni per il *word embedding*, che si applica alla matrice globale di co-occorrenze (parole-per-parole) ottenuta da un determinato corpus. In confronto ad altri metodi come *Word2Vec*, *GloVe* mostra solitamente una migliore capacità nell'identificare le relazioni non lineari nello spazio vettoriale e conferisce un peso inferiore alle parole altamente frequenti, evitando che dominino il processo di formazione dei vettori.

Dopo aver svolto alcune operazioni preliminari sui testi che servono a ottenere una lista di frequenza dei lemmi presenti nel corpus (riduzione delle parole in tutto minuscolo, lemmatizzazione con *R* `udpipe`, rimozione delle parole grammaticali e dei numeri) si è scelto di mantenere nell'analisi solo le parole (lemmi) con almeno 5 occorrenze nel corpus. Per ogni verbo-seme è stata utilizzata l'implementazione del metodo *GloVe* del pacchetto *R* `text2vec` (Selivanov, Bickel e Wang 2020), definendo una finestra di 5 parole a destra e a sinistra di ogni termine. Successivamente, sono state selezionate le prime 100 parole, ordinate per similarità decrescente, per ciascuno dei sei verbi-seme. Unificando le sei liste di 100 parole ciascuna ed eliminando i duplicati, si ottiene una lista[1] di 134 parole di interesse per l'analisi, in quanto dovrebbe delineare l'insieme di parole che meglio rappresentano l'area semantica dei sei verbi-seme di partenza.

1 adesso, almeno, amare, amico, amore, ancora, andare, anima, animo, aspettare, assai, atto, avere, avvenire, bastare, bello, bene, bisognare, bisogno, buono, capire, caro, caso, cercare, certo, chiamare, chiedere, cominciare, comprendere, conoscere, continuare, cosa, cuore, dare, dimenticare, dio, dire, dolore, domandare, donna, dovere, ecco, essere, famiglia, fanciulla, fare, fatto, figlio, figliuolo, finire, forse, forte, forza, fratello, giovane, grande, idea, insieme, intanto, intendere, invece, lasciare, luogo, madre, mai, male, marito, meglio, metro, mettere, modo, moglie, momento, mondo, morire, morte, mostrare, nome, non, nuovo, ogni, padre, parere, parlare, parola, parte, pensiero, perdere, persona, piacere, pieno, poi, potere, povero, prendere, presto, provare, punto, qualche, qui, quinto, ragione, rendere, resto, riconoscere, ricordare, rimanere, rispondere, riuscire, sembrare, sempre, servire, sicuro, signore, soltanto, sorella, stare, subito, temere, tempo, tenere, toccare, trattare, trovare, uomo, vecchio, vedere, venire, veramente, vero, vita, vivere, volere, volta

L'idea di cercare regolarità nelle traiettorie tracciate da osservazioni temporali di dati discreti è molto studiata in statistica e, con riferimento specifico allo studio dell'andamento temporale delle occorrenze di parole in corpora diacronici, risulta utile fare riferimento ai risultati già ottenuti in precedenza su corpora di letteratura scientifica (Tuzzi 2018, Trevisani e Tuzzi 2018, Trevisani e Tuzzi 2015) e di discorsi istituzionali (Trevisani e Tuzzi 2013).

La procedura adottata per il riconoscimento della dinamica delle traiettorie tracciate dai lemmi identificati attraverso il *word embedding* lavora in un'ottica di apprendimento (*statistical learning*) in tre passi: (1) normalizzazione delle occorrenze, (2) lisciamento della traiettoria (*smoothing*) e (3) raggruppamento delle parole con andamento simile (*curve clustering*).

Le traiettorie analizzate sono 140 (una per ogni verbo-seme e per ogni lemma selezionato con il *word embedding*) e i punti-anno di osservazione sono in tutto 56, cioè tutti gli anni compresi tra il 1825 e 1923 per i quali nel corpus è disponibile almeno un'opera.

L'occorrenza (assoluta) di una parola in un testo non rappresenta una buona misura del suo tasso di presenza in quanto il valore dipende dall'ampiezza del testo. In questo senso una prima trasformazione che tenga conto delle dimensioni in gioco, cioè il calcolo di una frequenza relativa, è un passaggio obbligato. Più in generale il tipo di normalizzazione viene scelto in base alle caratteristiche delle traiettorie e, in questo caso, si è optato per una normalizzazione non lineare che tiene conto sia dell'ampiezza dei testi (la frequenza relativa di una parola *i* in un anno *j* viene calcolata dividendo l'occorrenza osservata *nij* per il numero di occorrenze totali *Nj* nei testi dello stesso anno *j*) sia della popolarità della parola (la frequenza relativa viene ulteriormente normalizzata operando una trasformazione non lineare che, oltre a standardizzare tra 0 e 1 il *range* di frequenza per parola, trasforma ogni valore in una media pesata delle differenze dal minimo e rispettivamente dal massimo, per tenere conto della presenza di asimmetria positiva o negativa caratterizzante la parola; Grilli, Russo e Gismondi 2012). Questo tipo di normalizzazione riesce a tenere conto della diversa ampiezza dei subcorpora associati a ciascun anno di osservazione, a ridurre l'effetto delle parole ad alta frequenza, per evitare che mettano in ombra tutte le altre, e a smussare in qualche misura i bruschi saliscendi (picchi e precipizi) delle traiettorie, dovuti all'asimmetria dello spettro di frequenza specifico di ogni parola.

In particolare, se $x_{ij} = n_{ij}/N_j$ indica la frequenza relativa (rispetto alla dimensione $N_j$ dell'anno $_j$) della parola *i*, la trasformazione non lineare $y=f(x)$ è tale per cui

$$\frac{y_{ij} - m_i^y}{M_i^y - y_{ij}} = \frac{p_i\left(x_{ij} - m_i^x\right)}{\left(1 - p_i\right)\left(M_i^x - x_{ij}\right)}$$

dove $p_i$ è un parametro di lisciamento, tra 0 e 1, per la differenza tra $x_{ij}$ e il minimo $m_i^x$. Se $p_i = 0.5$, il rapporto delle distanze del valore dal minimo e rispettivamente dal massimo è uguale sia nella scala del valore trasformato *y* che nella scala del valore originario *x*, altrimenti il peso assegnato alla distanza dal minimo è maggiore o minore di quello assegnato alla distanza dal massimo a seconda del tipo di asimmetria (maggiore se asimmetria positiva, minore se negativa). Inoltre, per uniformare il livello delle traiettorie delle parole (l'effetto *popolarità*) si è imposto $m_i^y = 1$ e $M_i^x = 1$ per ogni parola.

La ricerca di regolarità nelle traiettorie osservate muove dall'idea che la frequenza normalizzata di una parola in un punto del tempo (anno) sia in grado di rappresentarne la vitalità, come un singolo fotogramma coglie un'immagine in un preciso istante, e che la sequenza delle occorrenze nel corso di diversi punti del tempo ne possa cogliere il ciclo di vita, come una serie di fotogrammi messi in successione ricostruisce un'immagine in movimento.

Dal punto di vista statistico è opportuno in questo contesto adottare un approccio per dati funzionali (*functional data analysis*) e leggere ogni osservazione, cioè la frequenza normalizzata *yij* di una parola *i* in un istante di tempo *j*, come la realizzazione di una sottostante funzione continua, che ne determina lo sviluppo nel tempo ma non è direttamente osservabile. A partire da una serie di osservazioni, si vuole ricostruire la dinamica generale, cioè la forma della funzione *x* che ha generato la traiettoria osservata:

$$y_{ij} = x_i(t_j) + \varepsilon$$

tenendo conto che l'osservazione è affetta da un rumore di fondo (una componente stocastica indicata come un errore ε) così come i fotogrammi risultano un po' sfocati perché l'immagine non è statica.

La figura 2 rappresenta l'insieme delle 140 traiettorie normalizzate oggetto di analisi (sei verbi-seme e 134 parole estratte attraverso il *word embedding*).

La sfida dei passaggi successivi consiste nel riuscire a estrarre da un ammasso indifferenziato di traiettorie gruppi di parole che condividono uno sviluppo temporale simile.

Attraverso un'operazione di filtraggio (*smoothing*) le traiettorie originali vengono lisciate in maniera da eliminare le asperità che rendono illeggibile l'andamento di fondo. Il lisciamento delle traiettorie viene svolto grazie a una scomposizione della funzione in una combinazione lineare di funzioni più semplici che, in questo esempio, è costituita da *B-spline* (Ramsay e Silverman 2005).

Attraverso un approccio di stima che penalizza l'irregolarità delle curve (*roughness penalty*) si ottiene un lisciamento ottimale regolando opportunamente i parametri che definiscono le curve. Come esempio, la figura 3 riporta l'andamento delle frequenze normalizzate dei sei verbi-seme e, in rosso, la versione lisciata della traiettoria che ne evidenzia la tendenza di fondo.

Le traiettorie lisciate vengono quindi raggruppate in insiemi (*clusters*) di parole aventi un andamento temporale simile (*curve clustering*) tramite un algoritmo (*k-means*) che utilizza una distanza euclidea tra curve come misura di similarità. L'algoritmo viene reiterato 20 volte (variandone l'inizializzazione) per ogni potenziale numero di gruppi (da un minimo di 2 a un massimo di 26 gruppi), generando così 20 partizioni possibili per ogni numero.

Per determinare il numero di gruppi ottimale si confrontano circa 50 diversi criteri di qualità del *clustering* al fine di trovare il miglior compromesso tra tutti i suggerimenti ottenuti (figura 4). Stabilito il numero ottimale, si individuano tra le 20 partizioni generate quelle scelte dai criteri di qualità e, tra quest'ultime, quella che massimizza il grado di sovrapposizione con le altre tramite un indice (*Rand index*) che rappresenta la stabilità e coerenza dei gruppi.

In questa fase, per scegliere la configurazione di *clustering* definitiva, è opportuno inserire tra i criteri di valutazione anche la lettura qualitativa da parte di esperti, per dare senso ai gruppi di parole identificati dalla procedura e avviare da qui l'interpretazione.

I calcoli necessari sono stati operati con il supporto di librerie e pacchetti *R*, `fda` (Ramsay et al. 2021), `clusterCrit` (Desgraupes 2018), `cclust` (Dimitriadou 2021), `clusterSim` (Walesiak e Dudek 2020), `kml` (Genolini et al. 2005), integrati da codice ad hoc.

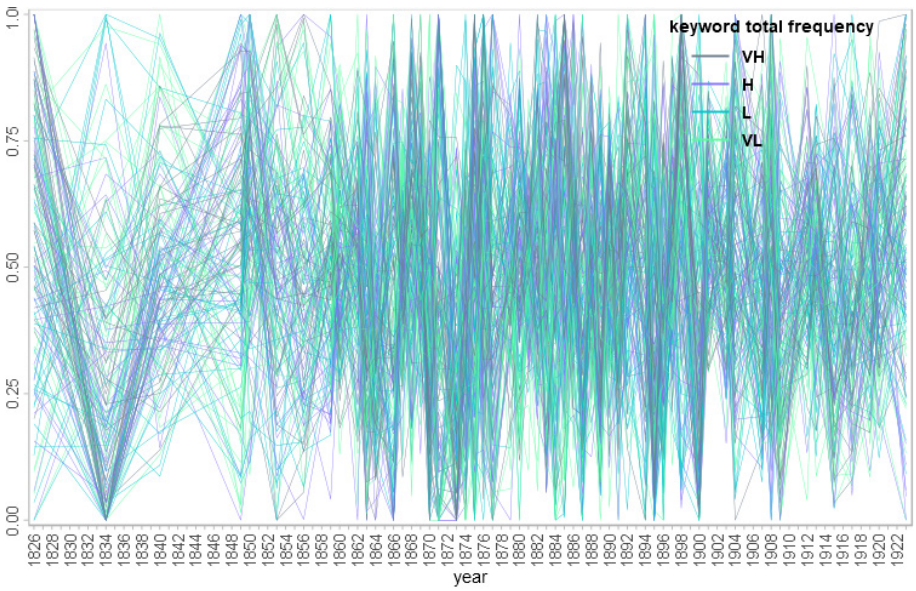


Figura 2 – Rappresentazione delle traiettorie normalizzate delle 140 parole oggetto di analisi. Parole con frequenza molto elevata (VH), elevata (H), bassa (L), molto bassa (VL)

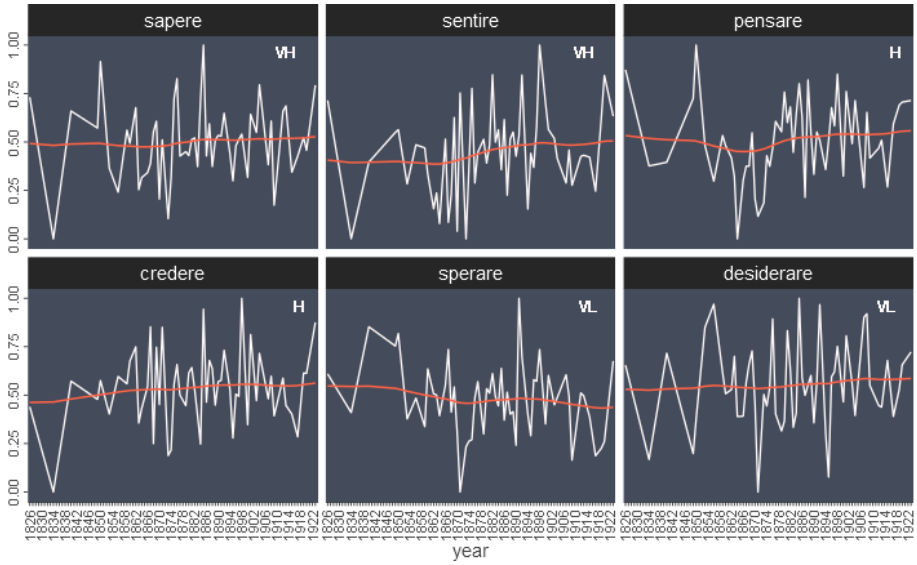


Figura 3 – Rappresentazione delle traiettorie normalizzate dei sei verbi-seme

## 5. Risultati

Il numero ottimale di gruppi suggerito dall'analisi dei criteri di qualità si può considerare quello pari a 6 (fig. 4) in quanto è fra i valori più scelti assieme a 3, 4 e 5, ma ha il vantaggio di creare partizioni più fini e di essere stato spesso indicato come prima o seconda soluzione. I metodi statistici impiegati non evidenziano andamenti temporali molto netti: alcune timide tendenze emergono, ma molte parole mostrano individualmente un comportamento indifferenziato nel corso del tempo. Sebbene il numero di gruppi e il numero di parole siano limitati, si può cercare di leggere congiuntamente alcuni andamenti interessanti a livello di *cluster* (figura 5).

Per quanto riguarda la collocazione dei sei verbi-seme, si può osservare come la frequenza di *sentire* (cluster B) e *pensare* (cluster C) abbia sperimentato una lieve crescita nel corso del tempo; *sapere* e *sperare* (cluster A), *credere* e *desiderare* (cluster F) si trovano invece in due cluster dove le parole mostrano una tendenza costante (molto numeroso e leggermente decrescente A, leggermente crescente F). Anche se è molto difficile trarre conclusioni sui temi trattati a partire da un numero così limitato di parole, i legami familiari e personali sembrano avere andamenti interessanti: *moglie* (B), *sorella* e *marito* (C) sono crescenti, restano presenze costanti *madre* e *amico* (F) *fratello*, *figlio* e *figliolo* (A) e risultano, invece, in diminuzione i riferimenti a *famiglia* (D) e *padre* (E).

Cresce nel tempo la frequenza di *donna* (B) mentre cala quella di *uomo* (D).

Il *dolore* e la *morte* mostrano nel periodo di tempo osservato una presenza pressoché costante (A) così come *morire* (F); mentre *vivere*, *piacere*, *continuare*

(B), *volere* e *vita* (C) sono crescenti. Sono crescenti anche *amare* (C), *anima* e *amore* (B), mentre i riferimenti al *cuore* e all'*animo* (D) sembrano in declino. Anche *Dio*, *male*, *buono* (E) e *bene*, *bello*, *ragione* (A, assieme a *sperare* e *sapere*) mostrano una tendenza in diminuzione o costante e rappresentano spunti interessanti per futuri approfondimenti specifici. Alcuni verbi appartenenti al cluster C, come *parlare*, *rispondere*, *chiedere*, *domandare*, *ricordare*, *capire* e *comprendere* risultano avere una tendenza di fondo in crescita e potrebbero indicare una maggiore attenzione all'interazione e al dialogo sia con altre persone sia a livello interiore.

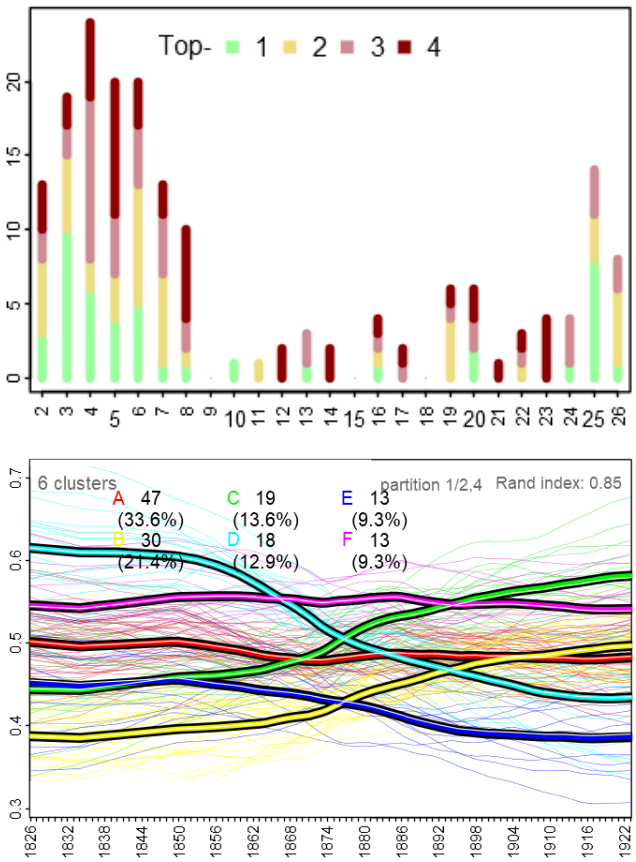


Figura 4 – Sopra, confronto tra le scelte del numero ottimale di gruppi suggeriti dal *basket* dei criteri: numero di criteri (in ordinata) che assegnano la preferenza al numero di gruppi (in ascissa) come prima (1 verde), seconda (2 giallo), terza (3 rosa) o quarta (4 vinaccia) scelta. Sotto, soluzione con 6 gruppi dove viene evidenziata la curva che rappresenta l'andamento medio e vengono riportate numerosità e percentuale di parole raccolte in ciascun gruppo

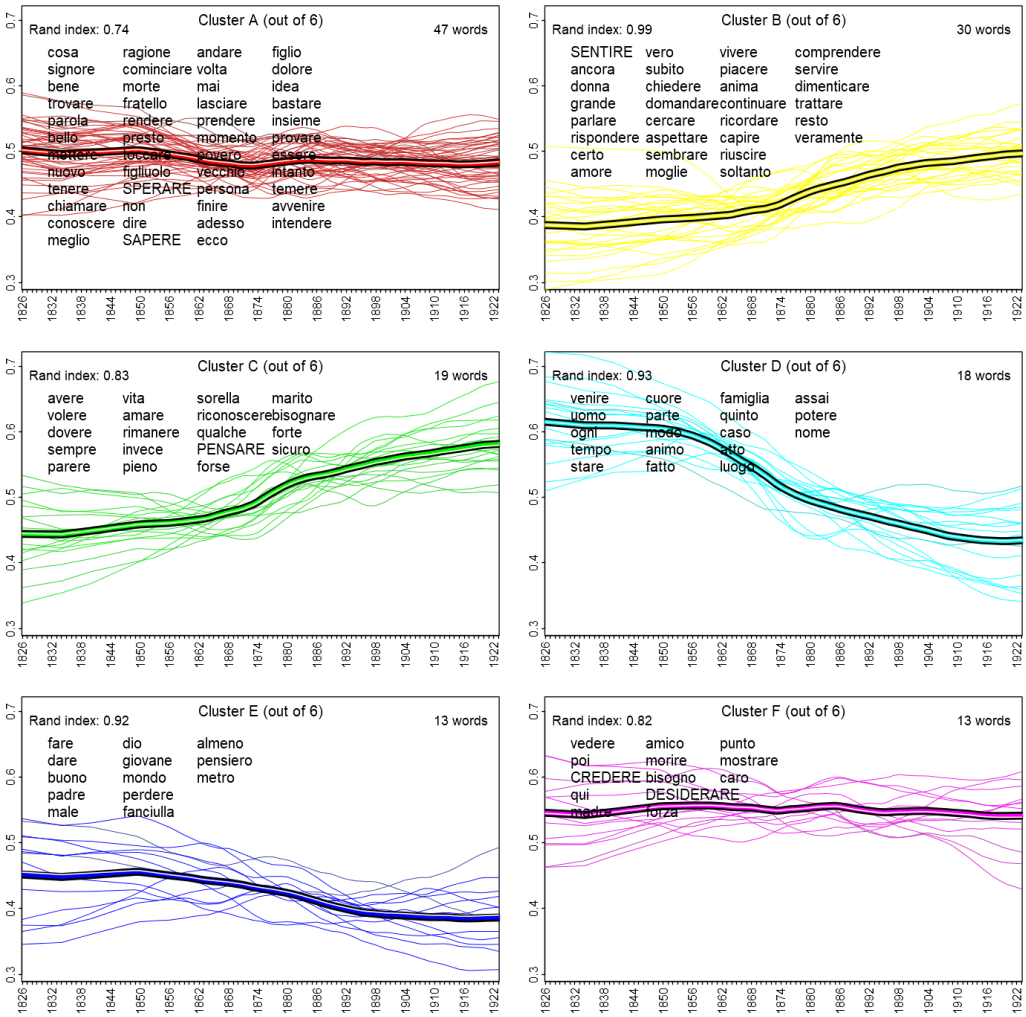


Figura 5 – Rappresentazione delle traiettorie delle parole inserite dalla procedura nei 6 gruppi. La curva evidenziata rappresenta la tendenza media del gruppo

## 6. Considerazioni conclusive

L'idea di studiare con prospettiva diacronica la presenza nelle opere dell'espressione interiore degli autori e dei loro personaggi è sicuramente molto ambiziosa e questo primo esperimento mostra alcune potenzialità degli strumenti statistici impiegati, ma anche alcuni limiti dei criteri adottati, sia per la costituzione del corpus sia per la scelta dei fenomeni linguistici da osservare.

Sebbene non si possa rigettare l'ipotesi che l'espressione interiore abbia assunto maggiore importanza nel corso dell'Ottocento, le analisi statistiche dei dati testuali non evidenziano con chiarezza le tendenze prospettate. I metodi statistici impiegati per l'analisi e il raggruppamento delle traiettorie disegnate nel tempo dalle occorrenze normalizzate evidenziano tendenze temporali deboli e la gran parte dei verbi seme e delle parole selezionate individualmente mostra, dopo il lisciamento, un andamento poco differenziato nel corso del tempo e solo dall'analisi dei clusters ci vengono alcune suggestioni riguardo a possibili tendenze in aggregato. In questo senso lo strumento statistico riesce solo parzialmente a essere d'aiuto e la riflessione deve essere rivolta più alla fase di selezione del corpus e di costruzione dei dati testuali che a quella di elaborazione. Nonostante le problematiche emerse, alcune timide tendenze, specie a livello di *cluster* e nel contesto specifico della costellazione di espressioni del sentire, dell'interiorità, dei modelli e dei ruoli sociali, delle relazioni e dei legami familiari, permettono di cogliere segnali di crescita e diminuzione, evidenziando globalmente le potenzialità dei metodi utilizzati.

Probabilmente la scelta di partire da sei verbi-seme, che forse nella lingua italiana sono troppo polisemici per essere delle buone *proxy* dell'espressione interiore, e da un insieme limitato di parole semanticamente legate a questi sei verbi-seme non porta a cogliere la complessità del fenomeno che si desidera indagare. La presenza nei testi dell'espressione interiore potrebbe essere effettivamente aumentata come suggerisce l'ipotesi, ma questa tendenza non può essere visibile nella frequenza delle parole se, parallelamente, sono cambiate nel corso del tempo le strategie discorsive per rappresentarla nel testo. Per esempio, nel testo narrativo moderno potrebbero essere frequenti monologhi interiori senza che siano presenti veri e propri marcatori (*credo che*, *lei pensò che*) che, viceversa, erano più presenti in passato.

Il metodo adottato per selezionare l'insieme di parole (*word embedding*) ha sicuramente grandi potenzialità ma si scontra con alcuni problemi tecnici, che riguardano sia la dimensione limitata del corpus (100 opere) sia l'addestramento dell'algoritmo. Con riferimento alla dimensione del corpus, si specifica che, nonostante il numero di *token* appaia elevato, generalmente l'addestramento dei *word embedding* utilizza corpora molto più estesi (solitamente almeno 5-10 volte la dimensione del corpus analizzato in termini di occorrenze) e questo può aver limitato la capacità di cogliere la semantica presente nei testi. In mancanza di un corpus molto esteso, sono disponibili anche vettori pre-addestrati su grandi corpora, come i contenuti di Wikipedia e di Twitter. Esistono in questo senso *word embedding* specifici per la lingua italiana, ma si è deciso di usare lo stesso corpus che abbiamo analizzato per l'addestramento poiché il contesto (pagine web e social media) in cui sono stati creati questi vettori pre-addestrati non appare coerente per analizzare testi narrativi dell'Ottocento. Anche per il processo di lemmatizzazione valgono considerazioni simili, in quanto le *treebank* disponibili non sono state create sulla base di testi di narrativa, ma prin-

cipalmente a partire da articoli di quotidiani e pagine di Wikipedia come nel caso di ISDT.

L'ultima considerazione riguarda i criteri adottati per la costituzione del corpus utilizzato rispetto agli obiettivi perseguiti dal progetto europeo. Se messo a confronto con il caso dell'analisi di corpora di testi di genere molto diverso (raccolte di messaggi pubblicati nei social network, articoli di giornale, articoli scientifici, ecc.) sembra più difficile identificare tendenze di fondo nelle frequenze sperimentate dalle parole nelle opere di narrativa. Si tratta, infatti, di testi più dispersivi, che includono contesti molto differenti e trattano temi in maniera più diluita, incidendo in tal modo sulla probabilità di associazione tra le parole su cui si basa il metodo di selezione con *word-embedding*. Avere a disposizione 100 opere significa sicuramente lavorare con un corpus di dimensioni di tutto rispetto. Tuttavia, queste dimensioni risultano insufficienti se le opere devono diventare la base di studi cronologici su un arco temporale molto ampio (un secolo). Nel caso del corpus esaminato, le 100 opere si collocano su 56 punti-anno diversi e, di conseguenza, ogni anno di osservazione è rappresentato da una sola opera o, comunque, da un numero molto ridotto di opere. Per poter rappresentare ogni punto-anno in maniera adeguata è necessario disporre di corpora di dimensioni molto più grandi e, soprattutto, costituiti da testi più numerosi e meglio distribuiti nel periodo di osservazione.

Cortelazzo M. A. (2021) "Corpora e storia della lingua", *RITT, Rivista Internazionale di Tecnica della Traduzione*, 23, pp. 179-186.

Desgraupes B. (2018) *clusterCrit: Clustering Indices*, R package version 1.2.8 https://CRAN.R-project.org/package=clusterCrit.

Dimitriadou E. (2021) *cclust: Convex Clustering Methods and Clustering Indexes*. R package version 0.6-23. https://CRAN.R-project.org/package=cclust.

Durum Wheat Varieties, *Quaderno DSEMS*, 23/2007.

Genolini C., Alacoque X., Sentenac M. e Arnaud C. (2015) "kml and kml3d: R Packages to Cluster Longitudinal Data", *Journal of Statistical Software*, 65(4), pp. 1-34. http://www.jstatsoft.org/v65/i04/.

Grilli, L., Russo, M. A. e Gismondi, R. (2012) "Methodological Proposals for a Qualitative Evaluation of Italian Durum Wheat Varieties", *Journal of Applied Economic Sciences*, 7, Issue 2 (20), summer 2012, pp. 103-122.

Mikolov, T., Chen, K., Corrado, G. S. & Dean, J. (2013) "Efficient Estimation of Word Representations in Vector Space", in *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. arXiv:1301.3781.

Ondelli S. (2021) "Tempi verbali e perifrasi gerundivali in un corpus di italiano letterario (1800-2000)", *RITT, Rivista Internazionale di Tecnica della Traduzione*, 23, pp. 187-212.

Pennington J., Socher R. & Manning C.D. (2014) "GloVe: Global Vectors for Word Representation", in *Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, pp. 1532-1543 http://www.aclweb.org/anthology/D14-1162.

R Core Team (2021) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramsay J. & Silverman B. W. (2005) *Functional data analysis*, Springer series in Statistics, New York, Springer.

Ramsay J.O., Graves S., Hooker G. (2021). *fda: Functional Data Analysis*, R package version 5.5.1. https://CRAN.R-project.org/package=fda.

Sciumbata F.C., Nadalutti P., Tringali L. (2021) "'Trovare lavoro' in un corpus di narrativa del XIX-XX secolo. Procedure, aspetti e problemi di creazione, estrazione e rappresentazione dei dati", *RITT, Rivista Internazionale di Tecnica della Traduzione*, 23, pp. 235-268.

Selivanov D., Bickel M., Wang Q. (2020). *text2vec: Modern Text Mining Framework for R*. R package version 0.6. https://CRAN.R-project.org/package=text2vec.

Straka M. & Straková J. (2017) "Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe", in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, pp. 88-99.

Trevisani M. e Tuzzi A. (2013) "Shaping the history of words", in *Methods and Applications of Quantitative Linguistics: Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO)*. Ed. by I. Obradović, E. Kelih & R. Köhler, Belgrade, Akademska Misao, pp. 84-95.

Trevisani M. e Tuzzi A. (2015) "A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal", *Quality and Quantity*, 49, pp. 1287-1304.

Trevisani M. e Tuzzi A. (2018) "Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories", *Knowledge-based systems*, 146, pp. 129-141.

Tuzzi A. (2018) (ed.) *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, Cham, Springer.

Walesiak M. & Dudek A. (2020) "The Choice of Variable Normalization Method in Cluster Analysis", in *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges. Proceedings of the 35th International Business Information Management Association Conference (IBIMA), 1-2 April 2020 Seville, Spain*. Ed. by K. S. Soliman, IBIMA, Seville, pp. 325-340.

Wijffels J. (2021) *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit*. R package version 0.8.8. https://CRAN.R-project.org/package=udpipe.