



Esercitazione con Voyant

Rachele Sprugnoli (Università di Parma)

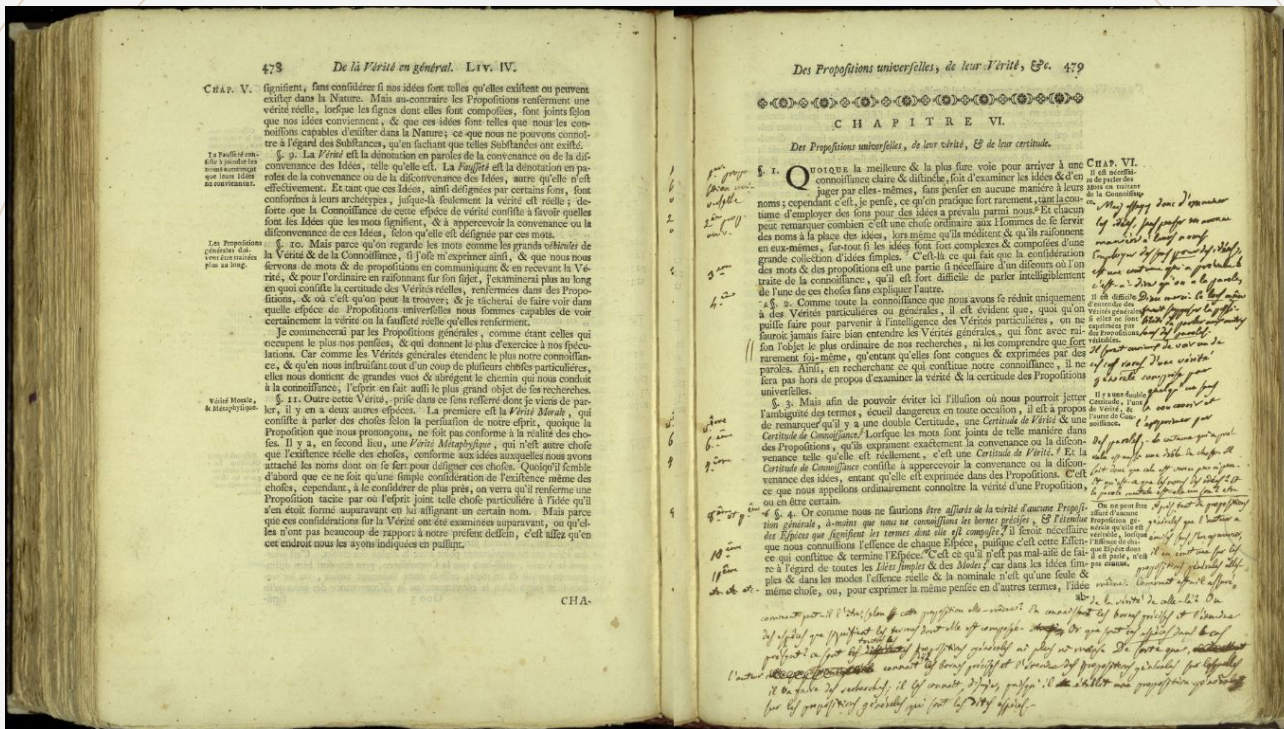
Trieste-online, 10 marzo 2022



1

I Concetti Fondamentali





Close Reading

<http://www.alessandromanzoni.org/biblioteca/esemplari/3883/>

Scalable Reading



Non sto suggerendo di archiviare la lettura da vicino e "letture" di letteratura altamente interpretative. Al contrario, sto suggerendo un **approccio misto**. [...] È esattamente questo tipo di unificazione, della scala macro e micro, che promette una nuova, migliorata e migliore comprensione della letteratura.

Le due scale di analisi **lavorano in tandem e comunicano tra loro**. L'interpretazione umana dei "dati", sia che siano estratti su scala macro o micro, rimane essenziale. Sebbene i metodi di indagine, di raccolta delle prove, siano diversi, non sono antitetici e condividono lo stesso obiettivo finale di **rafforzare la comprensione della letteratura**, sia essa scritta in grande o in piccolo.

Jockers, *Macroanalysis. Digital Methods and Literary History*, 2013



2

Panoramica su Voyant

Cosa è Voyant

- Voyant Tools è un ambiente web per la lettura e l'analisi di testi
 - vari formati di input: txt, pdf, html, xml
 - può essere integrato su altri siti
 - interattivo
 - permette una lettura scalabile
 - indipendente dalla lingua
 - analisi lessicale





Dove trovare Voyant

- Sito ufficiale: <https://voyant-tools.org/>
- Mirror:
 - <https://voyant-tools.huma-num.fr>
 - <https://voyant.lincsproject.ca>
 - <https://service.sadilar.org/voyant/>
- Server installabile:
<https://github.com/voyanttools/VoyantServer>



Che tipo di file caricare

- Formati accettati: TXT, HTML, XML, TEI, PDF, RTF, MS Word, JSON, tabelle in fogli di calcolo
- PDF: contiene un OCR, il risultato può variare
- **ATTENZIONE:** il modo in cui vengono caricati i file influenza il tipo di analisi successiva
 - eg.: testo unico *versus* divisione in capitoli

Struttura della skin



Funzioni principali (1)

- Cambiare skin



- READER: lettore del testo, permette il close reading
- CIRRUS: visualizzatore frequenza dei termini
- BUBBLES: visualizzatore frequenza dei termini
- TERMS: analisi della frequenza dei termini
- TRENDS: andamento delle frequenza dei termini
- BUBBLELINES: frequenza e distribuzione dei termini
- MICROSEARCH: frequenza e distribuzione dei termini
- CONTEXT: contesti di occorrenza dei termini

Funzioni principali (2)

- Cambiare skin




- PHRASES: sequenze di parole che co-occorrono
- COLLOCATES: termini che appaiono vicino ad altri termini
- CORRELATIONS: termini la cui frequenza varia in sintonia
- MANDALA: relazioni tra termini e documenti
- SUMMARY: informazioni sul corpus
- DOCUMENTS: informazioni sui singoli documenti
- TOPICS: topic modeling



<https://voyant-tools.org/docs/#!/guide/tools>



Come e cosa esportare

- L'esportazione  si può applicare all'intero progetto Voyant o a una singola skin
 - Puoi esportare una **URL**, uno strumento incorporabile (**interattivo**) o un riferimento **bibliografico**
 - Puoi anche esportare un file **.png** statico nel caso delle visualizzazioni (uno screenshot potrebbe avere una migliore qualità dell'immagine)
 - Puoi esportare i dati dagli skin a forma di **tabella** in vari formati



Come cercare

- Sintassi delle ricerche lessicali
 - pestilenza: trova il termine esatto
 - pestilen*: trova termini che iniziano con "pestilen"
 - "marito e moglie": cerca l'intera espressione
 - "opera misericordia"~5: "opera" e "misericordia" co-occorrono entro 5 termini
 - @Personaggi: ricerca raggruppata di tutti i termini inclusi in una *categoria*
 - ^@Personaggi: ricerca dei singoli termini inclusi in una *categoria*




3

Esercitazione Pratica



Come caricare file

- Andare su Voyant (sito ufficiale, mirror o versione in locale)

- cliccare su opzioni 

- sotto **Processing** scegliere **Simple Word Boundaries**

The following table summarizes tokenization for the string **What's voyant-tools.org?**:


Tokenization	Count	Tokens	Notes
Automatic	3	what's, voyant, tools.org	the hyphen is split but the tools.org is considered a URL token; tokens are lowercase
Word Boundaries	5	what, s, voyant, tools, org	any non-word character is a delimiter, tokens are lowercase
Whitespace Only	2	What's, voyant-tools.org?	punctuation is kept in tokens and case is unchanged

- cliccare su **Upload**, aprire la cartella **capitoli**, selezionare (ctrl+a) tutti i file e cliccare su **Apri**: sono i capitoli de I Promessi Sposi

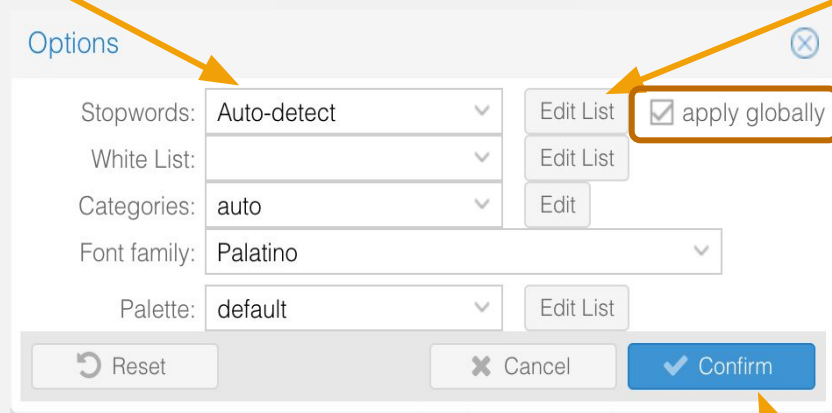
Aggiungere le stopwords (1)

- Parole funzione (*versus* parole contenuto) da ignorare: congiunzioni, preposizioni, articoli...
- Dove trovare le liste di stopwords?
 - Lingue moderne: <https://www.ranks.nl/stopwords>
 - Latino e greco antico:
<https://github.com/aurelberra/stopwords/tree/master/ancientstopwords/data>
 - Lista di stopwords per l'italiano nella cartella Voyant: aprire il file **stopword-it.txt** con un editor di testo, selezionare e copiare la lista

Aggiungere le stopwords (2)

- Su Voyant, cliccare sulle opzioni della skin Cirrus 

1) Selezionare “Italian”




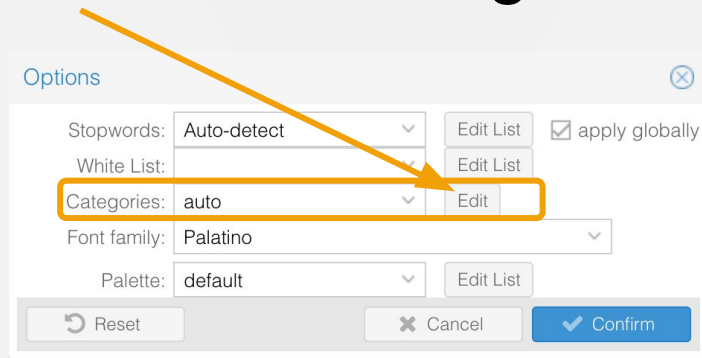
2) Cliccare su **Edit List**,
incollare la lista
del file
stopwords-it.txt

3) Cliccare su **Confirm**

Aggiungere le categorie (1)

- CATEGORIE: gruppi di parole semanticamente connesse, ad esempio lista di personaggi, lista di luoghi, lista di emozioni da usare per ricerche mirate

- Cliccare sulle opzioni 
- Cliccare su **Edit** vicino a **Categories**



Aggiungere le categorie (2)

Categories Builder

Categories Features

Terms

Term	Count
renzo	561
disse	560
don	442
lucia	391
gran	330
padre	281
parole	253
mano	238
abbondio	227
buon	220
agnese	216
dio	211
voce	203
signor	197
tosto	196
gente	191
porta	190
giorno	187
signore	174

Categories

positive

negative

freedom
bad
advantage
concern
excellent
fail
superior
despair
confidence
desperate
enjoy
disadvantage
wonderful
depression
amazing
disaster
enthusiasm
criticize
bliss
suffering
optimistic
suffer
good
sad
hope
inferior
happy
horror
happiness
hesitation
praise
terrible
safe
forbidden
success
failure

Add Category Remove Selected Terms

Cancel Save

- 1) Rimuovere le categorie esistenti
- 2) Cliccare su **Add Category**
- 3) Scrivere un nome per la categoria, e.g. **Personaggi** e cliccare su **Add**

Add Category

Category Name:

Cancel Add

Aggiungere le categorie (3)

The screenshot shows the 'Categories Builder' interface. On the left, under 'Terms', there is a table with two columns: 'Term' and 'Count'. The table lists various terms and their counts. On the right, under 'Categories', there is a list of terms that have been added to a category named 'Personaggi'. An orange arrow points from the 'Terms' table to the 'Personaggi' list. At the bottom of the interface, there are buttons for 'Add Category', 'Remove Selected Terms', 'Cancel', and 'Save'. Another orange arrow points from the 'Save' button to the text '5) Salvare cliccando su Save'.

Term	Count
passi	63
pensare	62
sicuro	62
braccia	61
giù	61
signora	61
attorno	60
teneva	60
trovò	60
bravi	59
entrò	59
famiglia	59
persona	59
risposta	59
domani	58
principe	58
trovava	58
cominciò	57
amici	56

Categories Builder

Categories

Personaggi

- renzo
- lucia
- abbondio
- agnese
- rodrigo
- crisoforo
- gertrude
- perpetua
- federigo
- griso
- ferrer

Cancel Save

4) Trascinare i nomi dei personaggi dalla lista **Terms** alla lista **Personaggi**

5) Salvare cliccando su **Save**

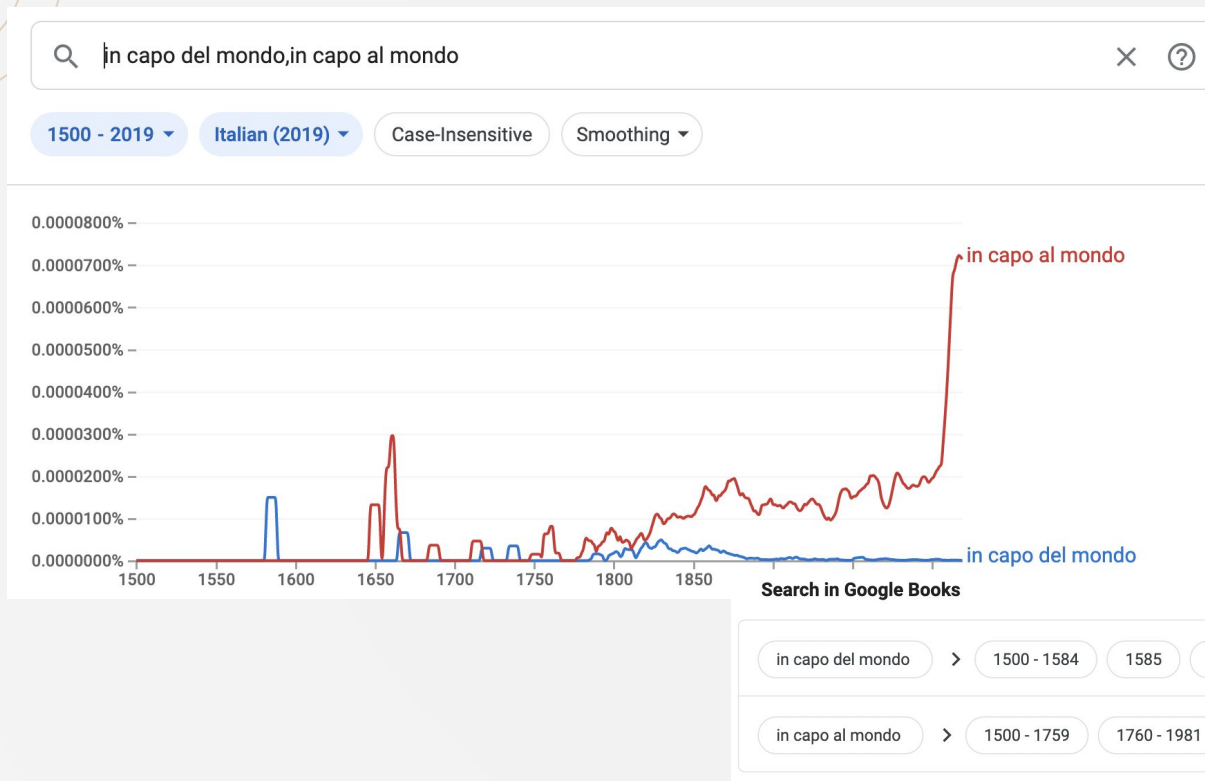
Esempi di utilizzo

1. Come individuare gli hapax legomena? → TERMS
2. Quali sono le parole più frequenti del cap 34? → CIRRUS + SCALE, DISTINCTIVE WORDS (nella skin DOCUMENTS)
3. Quali sono i capitoli più connessi alla pestilenza? → MANDALA
4. Quale personaggio viene menzionato di più e in che parti? → usare categoria su TRENDS, BUBBLELINES, MICROSEARCH
5. Qual è il capitolo con più densità lessicale? → DOCUMENTS
6. Si possono intuire delle caratteristiche psicologiche di Renzo e Lucia? → COLLOCATES
7. Come individuare gli usi metaforici di una parola? CONTEXTS
8. Come individuare forme desuete? → CONTEXTS (es. is*)

Un'analisi più dettagliata (1)

- “In capo del mondo” oppure “in capo al mondo”?
 - Cercare in CONTEXTS "capo mondo"~1: cosa osserviamo?
 - TLIO: <http://tlio.ovc.cnr.it/TLIO/> (cercare capo)
 - Vocabolario della Crusca del 1826 appartenuto a Manzoni: <https://www.alessandromanzoni.org/biblioteca/esemplari/4155/reader#page/335/mode/1up>
 - Google Books Ngram Viewer: <https://books.google.com/ngrams> (vedi slide successiva)

Un'analisi più dettagliata (2)



Analisi di una locuzione (1)

- “Porre le mani addosso” oppure “mettere le mani addosso”?
 - Cercare in CONTEXTS "mani addosso": cosa osserviamo?
 - Confrontiamo con due corpora diacronici:
 - <https://www.corpusmidia.unito.it/index.php>
 - <https://corpora.ficlit.unibo.it/DiaCORIS/>
 - Confrontiamo con un corpus d'italiano contemporaneo:
 - https://www.corpusitaliano.it/it/access/simple_interface.php

Analisi di una locuzione (2)

■ MIDIA

- 1) Ricerca forma
- 2) Opzioni avanzate
- 3) Ricerca forma precedente

The screenshot shows the MIDIA search interface. At the top, there is a search bar labeled "Cerca" containing the text "addosso" and a dropdown menu set to "Uguale a". Below the search bar are two buttons: "Opzioni avanzate..." and "Storico ricerche". A green bar highlights the "Posizione" section. Below this, there is a text box with the instruction "Scegli le caratteristiche appartenenti alla stringa precedente o successiva a quella cercata". Underneath, the word "Precedente" is displayed. A text input field labeled "Forma" contains the text "mani". To the right of the main interface, there is a separate section with the text "4) Cerca" and two buttons: "Reset" and "Cerca". Orange arrows point from the numbered list on the left to the corresponding elements in the interface: from "1) Ricerca forma" to the search bar, from "2) Opzioni avanzate" to the "Opzioni avanzate..." button, from "3) Ricerca forma precedente" to the "Forma" input field, and from "4) Cerca" to the "Cerca" button in the bottom right.

Analisi di una locuzione (3)

■ DiaCORIS

- 1) Ricerca
- 2) Mostra tutti i risultati
- 3) Esegui

The screenshot shows the DiaCORIS search interface with three orange arrows pointing from the list on the left to specific parts of the interface:

- The first arrow points from "1) Ricerca" to the "Query" input field containing the search string: "mani" [0,0] "addosso".
- The second arrow points from "2) Mostra tutti i risultati" to the "Concordance Options" section, specifically to the "Show" button and the radio button for "all" lines.
- The third arrow points from "3) Esegui" to the "Esegui" button at the bottom of the interface.

The interface is divided into several sections:

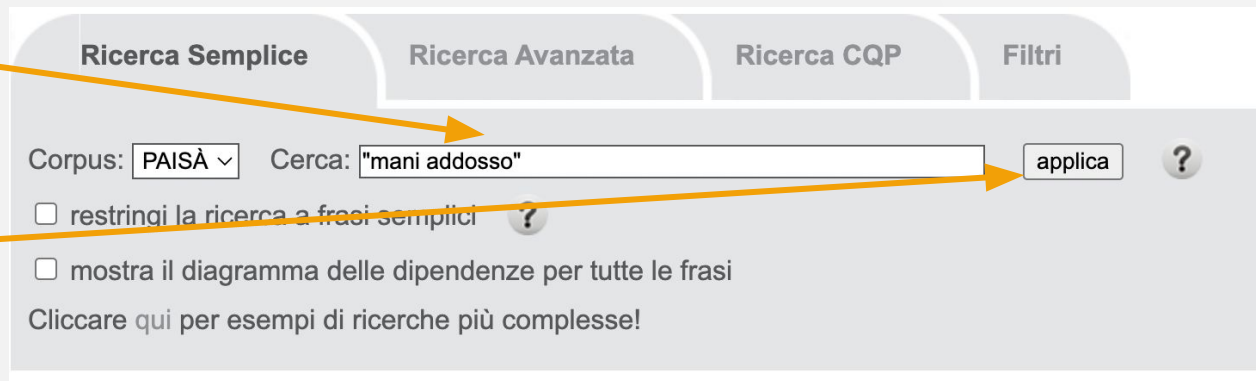
- User Authentication:** Contains a message: "DiaCORIS access is now free for research purposes (Please, read the footnote carefully)." and a link to "(Query Language Help)".
- Query:** Contains the search input field and dropdowns for "Section" (All) and "SubCorpus" (All).
- Concordance Options:** Includes a "Show" button and radio buttons for "30", "100", "300", and "all" lines.
- Collocations:** Includes a "Get Collocates?" section with radio buttons for "NO" (selected) and "Yes".
- Sort position:** A dropdown menu set to "Unsorted".
- Sort using:** Radio buttons for "Log-Likelihood Ratio." (selected), "Mutual Information.", "T-score.", and "Raw frequency.".
- Buttons:** "Esegui" and "Cancella" buttons at the bottom.

Analisi di una locuzione (4)

- Paisà (ricerca semplice)

1) Ricerca stringa

2) Applica



The screenshot shows a search interface with four tabs: "Ricerca Semplice", "Ricerca Avanzata", "Ricerca CQP", and "Filtri". The "Ricerca Semplice" tab is active. Below the tabs, there is a search form with the following elements:

- Corpus: PAISÀ (dropdown menu)
- Cerca: "mani addosso" (text input field)
- applica (button)
- ? (help icon)
- restringi la ricerca a frasi semplici ?
- mostra il diagramma delle dipendenze per tutte le frasi
- Cliccare qui per esempi di ricerche più complesse!

Two orange arrows point from the text on the left to the interface: one from "1) Ricerca stringa" to the search input field, and another from "2) Applica" to the "applica" button.

Analisi di una locuzione (5)

- Paisà (ricerca avanzata)

1) Ricerca con lemma
“porre”

Ricerca Semplice Ricerca Avanzata Ricerca CQP Filtri

Corpus: PAISÀ

Parola 1
Forma =
Lemma = porre
POS =
da 1 a 1 occorrenze
 ignora maiuscole/minuscole ?
 ignora diacritici (come è, è) ?

parole
da 1 a 1
a 1

Parola 2
Forma = mani
Lemma =
POS =
da 1 a 1 occorrenze
 ignora maiuscole/minuscole ?
 ignora diacritici (come è, è) ?

parole
da 0 a 0

Parola 3
Forma = addosso
Lemma =
POS =
da 1 a 1 occorrenze
 ignora maiuscole/minuscole ?
 ignora diacritici (come è, è) ?

2) Ricerca con lemma
“mettere”

Ricerca Semplice Ricerca Avanzata Ricerca CQP Filtri

Corpus: PAISÀ

Parola 1
Forma =
Lemma = mettere
POS =
da 1 a 1 occorrenze
 ignora maiuscole/minuscole ?
 ignora diacritici (come è, è) ?

parole
da 1 a 1
a 1

Parola 2
Forma = mani
Lemma =
POS =
da 1 a 1 occorrenze
 ignora maiuscole/minuscole ?
 ignora diacritici (come è, è) ?

parole
da 0 a 0

Parola 3
Forma = addosso
Lemma =
POS =
da 1 a 1 occorrenze
 ignora maiuscole/minuscole ?
 ignora diacritici (come è, è) ?

Provate voi!

- Caricate i 3 file della cartella **versioni**, caricate la lista di stopwords e provate a rispondere alle seguenti domande:
 1. Ci sono differenze evidenti tra i 3 testi dal punto di vista quantitativo?
 2. Ci sono sintagmi ricorrenti? Sintagmi con piccole variazioni?
 3. La presenza di Lucia è simile nelle 3 versioni?
 4. Il personaggio di Geltrude appare in “Fermo e Lucia” similmente a come appare Gertrude ne “I Promessi Sposi”?
 5. Quante parole contengono la lettera “j” nelle varie versioni?
 6. Cosa notate nella frequenza di “egli”? !!ATTENZIONE!!



Grazie!

Domande?

Mi trovate a rachele.sprugnoli@unipr.it.

Su Twitter: @RSprugnoli

Per scoprire di più sulle mie ricerche

<https://personale.unipr.it/it/ugovdocenti/person/236480>.