



**Bembus**

Promosso da



**Il Liutaio  
nel Bazaar**

Iniziativa realizzata con i fondi  
per le attività studentesche  
dell'Università Ca' Foscari Venezia



Università  
Ca' Foscari  
Venezia

# Analisi dei *corpora* testuali con Voyant Tools

**Rachele Sprugnoli** (Università di Parma)

Venezia-online, 8 marzo 2022

Il workshop è parte del ciclo “Metodi e strumenti per gli umanisti digitali” organizzato da Bembus, promosso da Il Liutaio nel Bazaar e finanziato con i fondi per le attività studentesche dell'Università Ca' Foscari Venezia.

A cura di: Marco Sartor (Università di Parma). Organizzazione: Mara Caron.



**Bembus**

# Piacere!

Mi presento! Sono **Rachele Sprugnoli**,  
ricercatrice all'Università di Parma.

Mi occupo di Linguistica Computazionale  
e di Informatica Umanistica.

Sono nel Direttivo dell'AIUCD.

La mia vita accademica e professionale ha  
visto varie tappe: Pisa, Trento, Milano,  
Parma.



# Cosa faremo



Bembus

1

## DEFINIZIONE CONCETTI

- Corpus
- Lettura da vicino/da lontano/scalabile
- Visualizzazione dei dati

2

## PANORAMICA SU VOYANT

- Cosa è
- Come caricare file
- Quali sono le funzioni principali

3

## ESERCITAZIONE PRATICA

- Lavoriamo su “I Promessi Sposi”



Bembus

**1**

# **Introduzione Teorica**

## **I Concetti Fondamentali**



**Bembus**

# **Cos'è un corpus?**

# Corpus



Bembus



A corpus is a collection of pieces of language **text** in **electronic** form, **selected** according to external criteria to **represent**, as far as possible, a language or language variety as a source of data for linguistic research.

Sinclair, *Corpus and text – Basic principles*, 2005



Bembus

# Rappresentatività



A corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what is meant to represent. **Representativeness** of the corpus, in turn, determines the **kind of research questions** that can be addressed and the **generalizability** of the results of the research.

Biber et al., *Corpus Linguistics Investigating Language Structure and Use*, 1998

# Caratteristiche



Bembus

- Testuale
  - anche il parlato solitamente è trascritto
- Elettronico
  - per permettere l'interrogazione e l'analisi
- Selezionato
  - criteri scientifici (linguistica dei corpora)
- Rappresentativo
  - misura il grado in cui è possibile indurre generalizzazioni sull'intera lingua o varietà



# Criteri (1)



Bembus

1. Modalità
  - scritto, parlato, misto, audio, multimediale
2. Generalità
  - generalista (orizzontale) o specializzato (verticale)
3. Cronologia
  - sincronico (statico) o diacronico (dinamico)
4. Distribuzione geografica
  - varietà e dialetti

# Criteri (2)



Bembus

5. Lingua
  - mono/bi/plurilingue (parallelo, allineato, comparabile)
6. Integrità
  - testi interi o porzioni
7. Annotazione linguistica
  - fonologica, morfologica, sintattica, semantica...
8. Competenza linguistica
  - madrelingua (L1) o apprendenti (L2)

# Criticità



Bembus

- I corpora sono sempre frammenti parziali e incompleti del linguaggio
  - ogni corpus è un campione (finito) con cui si vuole rappresentare una popolazione infinita
- Ogni corpus è necessariamente sbilanciato in qualche modo a causa della sua finitezza
  - possono mancare costruzioni importanti anche se rare
  - altre costruzioni possono essere presenti in eccesso

# Chomsky



Bembus



**Corpus linguistics doesn't mean anything.** It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this.

Chomsky, Intervista, 2004



Bembus

# Ma...



*Corpus linguistics is not about experimentation;  
it is about observation. Observation is an  
essential component of the hard sciences.*

*Desagulier, Noam Chomsky's colorless green idea: "corpus linguistics  
doesn't mean anything", 2017*

*Knowing that your corpus is unbalanced is what  
counts.*

*Atkins et al., Corpus Design Criteria, 1992*



# Esempi di corpora per l'italiano

- MIDIA: <https://www.corpusmidia.unito.it>
  - 7,8 milioni di parole: scritto, generalista, diacronico (1200-1947), monolingue, annotato
- DiaCORIS: <https://corpora.ficlit.unibo.it/DiaCORIS/>
  - 15 milioni di parole: scritto, generalista, diacronico (1861-2001), monolingue, non annotato
- Paisà: <https://www.corpusitaliano.it>
  - 250 milioni di parole: scritto (web), generalista, sincronico (2010), monolingua, annotato



Bembus

# **Tipi di lettura: da vicino, da lontano o scalabile?**



Bembus

# Close Reading



The principal object of close reading is to **unpack the text**. Close readers linger over words, verbal images, elements of style, sentences, argument patterns, and entire paragraphs and larger discursive units within the text to explore their significance on multiple levels.

Jasinski, Sourcebook on Rhetoric, 2001





Bembus

# Distant Reading



*A **condition of knowledge**: it allows you to focus on units that are much smaller or much larger than the text.*

Moretti, *Conjectures of World Literature*, 2000

*The construction of **abstract models***

Jasinski, *Sourcebook on Rhetoric*, 2001



Bembus

# Distant Reading

*A macroanalytic approach*

Jockers, *On Distant Reading and Macroanalysis*, 2011

*The idea of processing content in or information about a large number of textual items **without** engaging in the reading of the actual text.*

Drucker, *Distant Reading and Cultural Analytics*, 2013

**Bembus**

# 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



Bembus

# **Cos'è la visualizzazione dei dati?**



# Visualizzazione dei dati

- Rappresentazione dei dati attraverso un linguaggio visivo
- Perché?
  - esplorare i dati
  - trovare schemi ricorrenti
  - comprendere il contenuto
  - supervisionare la procedura di analisi
  - comunicare



# Visualizzazione dei dati

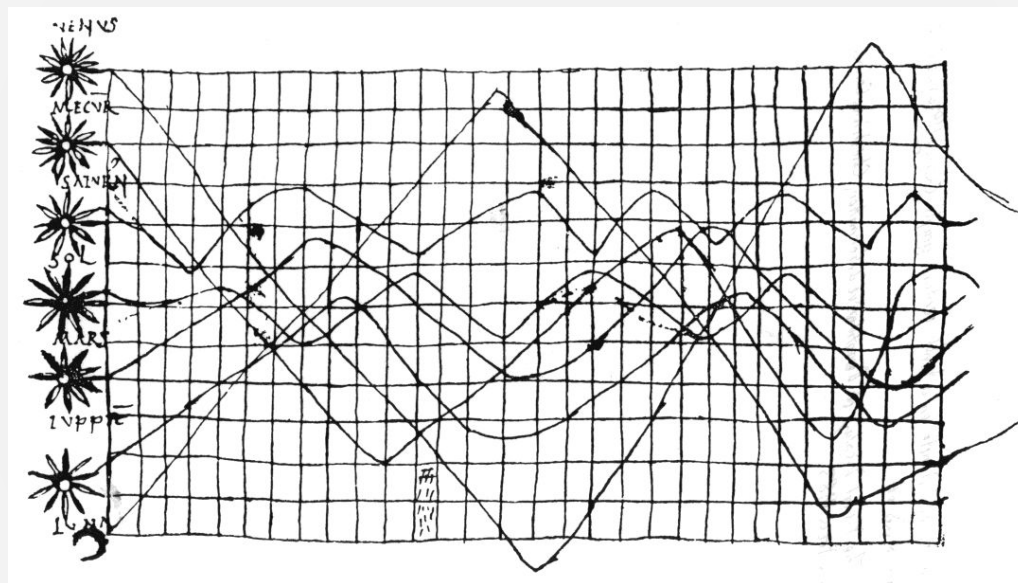
- Secondo Edward Tufte, una corretta visualizzazione dei dati dovrebbe:
  - mostrare i **dati**, indurre a pensare alla **sostanza**, evitare **distorsioni**, presentare molti dati in uno **spazio piccolo**, rendere **coerenti** anche grandi collezioni di dati, incoraggiare **confronti**, avere vari livelli di **dettaglio**, avere uno **scopo** chiaro, essere strettamente **integrato** con la parte statistica e descrittiva

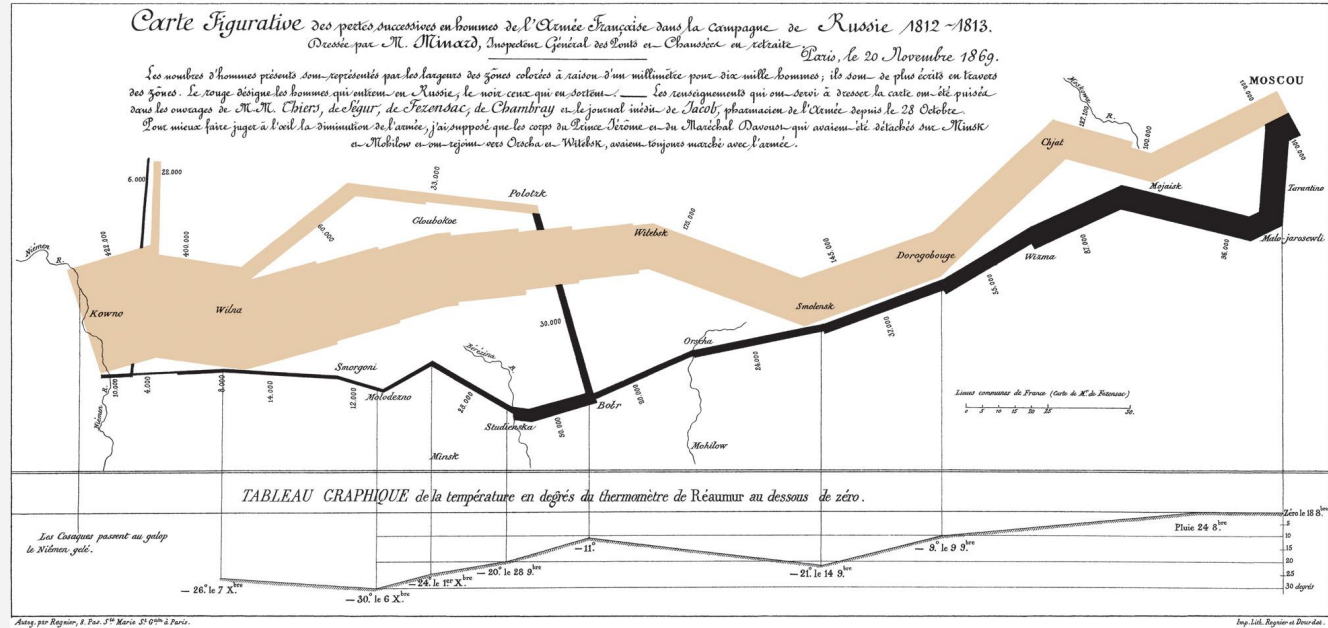
Tufte, *The Visual Display of Quantitative Information*, 2001

# Molto prima di Tufte...



Bembus









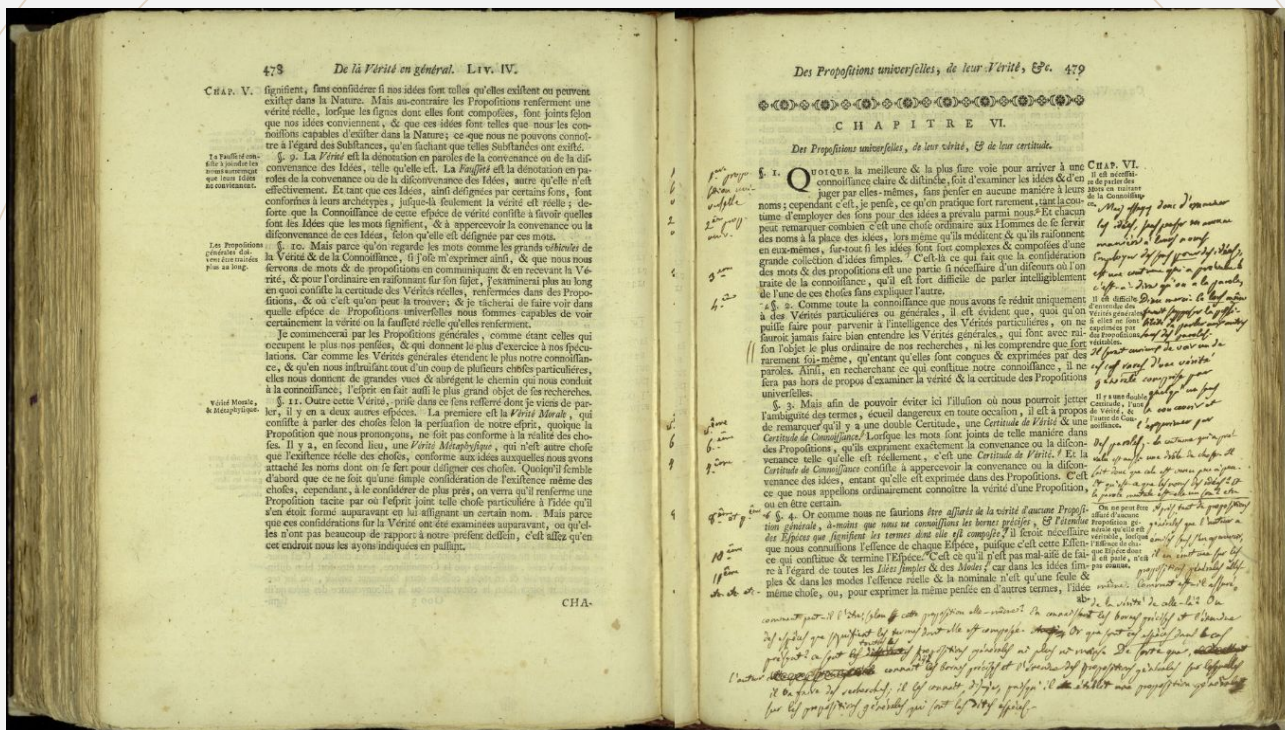
Bembus

# **Qual è la relazione tra distant reading e visualizzazione dei dati?**

**Bembus**

Le immagini **vengono prima di tutto**, nei nostri pamphlet, perché – visualizzando i dati empirici – costituiscono l’oggetto specifico di studio della critica computazionale: esse sono il nostro “testo”, il corrispettivo di quanto un preciso segmento narrativo sta al close reading.

Moretti, *La letteratura in laboratorio*, 2019



# Ce Reading

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





<http://www.fedoabooks.unina.it/index.php/fedoapress/catalog/book/104>

# Dai testi alle visualizzazioni



Bembus

- Il testo subisce un processo di deliberata riduzione e astrazione prendendo a prestito modelli da 3 discipline:
  1. Grafici → storia quantitativa
  2. Mappe → geografia
  3. Alberi → teoria dell'evoluzione



Bembus



**Graphs, maps, and trees** place humanities disciplines literally in front of our eyes-and show us how little we still know about it.

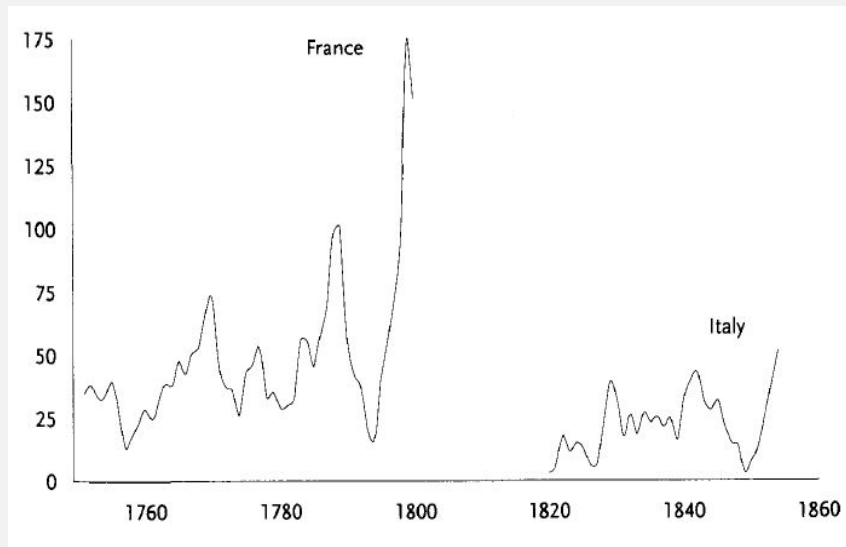
Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*, 2007

# Grafici



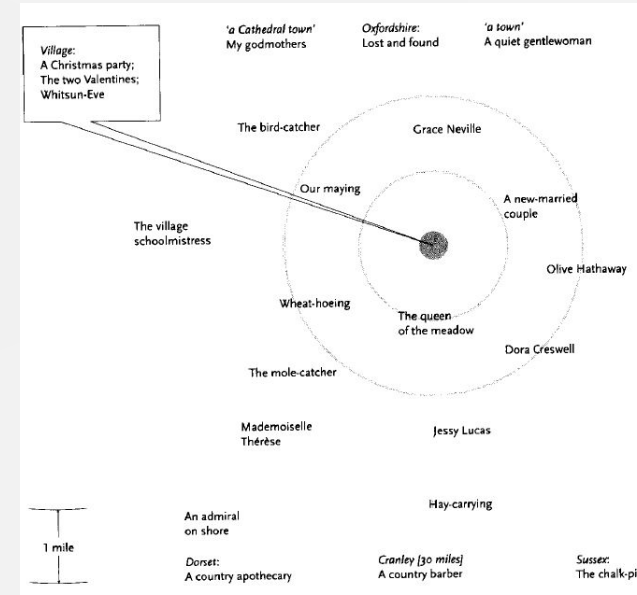
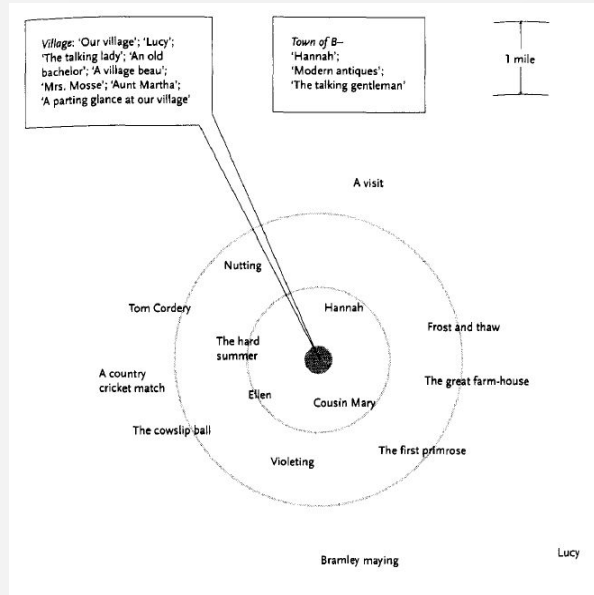
Bembus

- L'ascesa e la caduta del romanzo



# Mappe

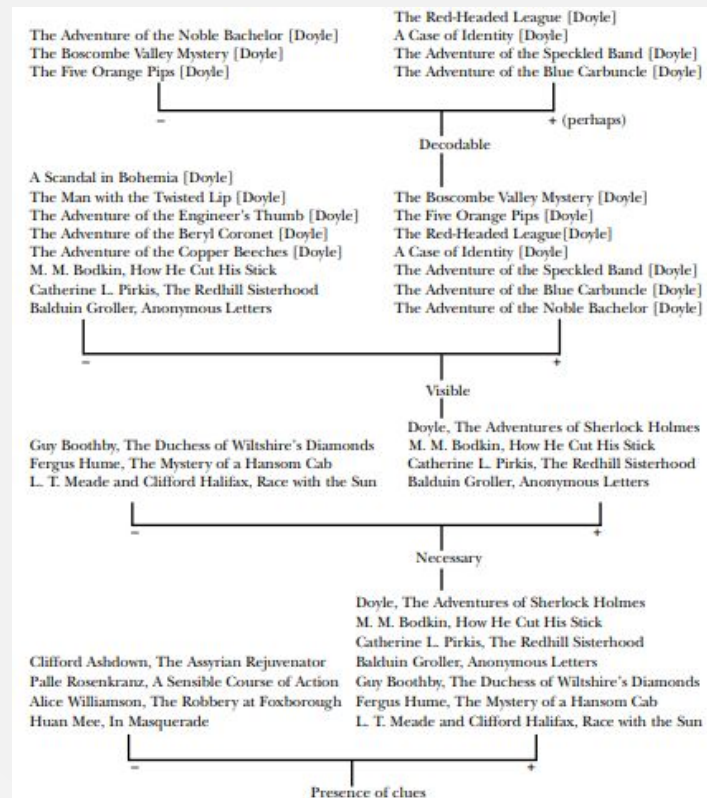
- Mary Mitford, “Our Village”: 1824 versus 1828





# Alberi

- Il successo di Conan Doyle



Bembus



Bembus

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



Bembus



# Dove trovare Voyant



Bembus

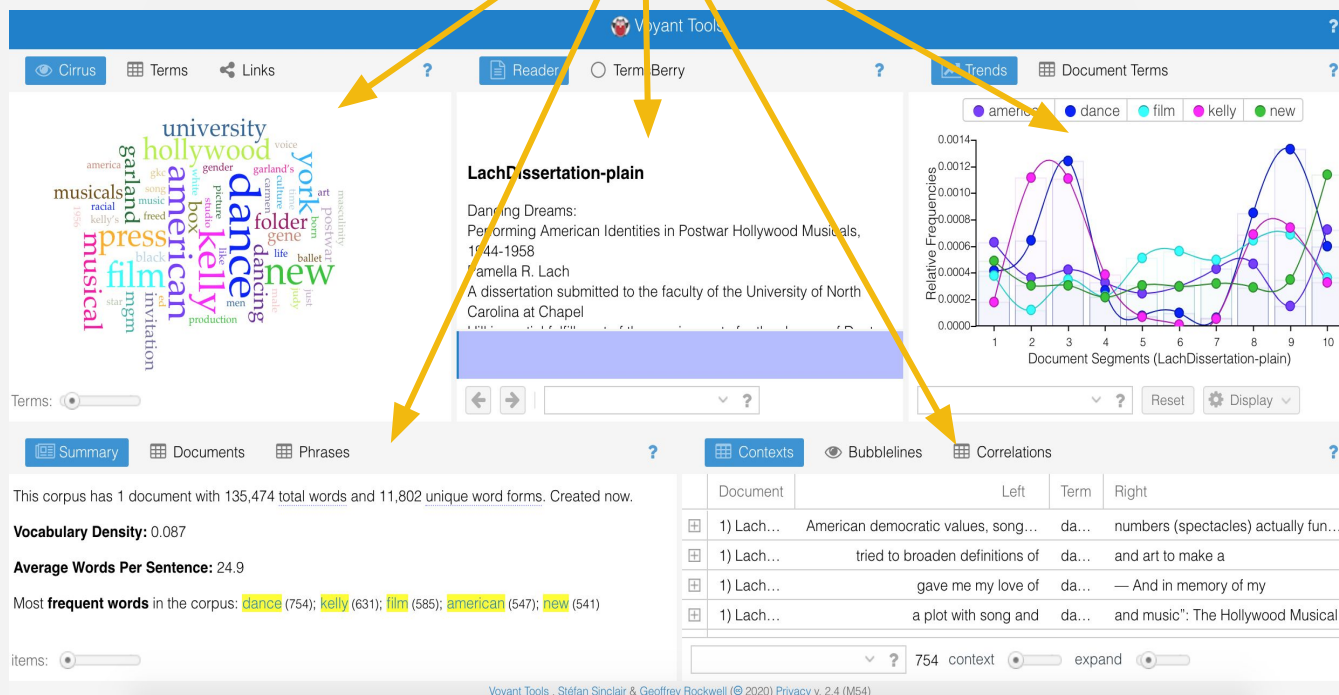
- Sito ufficiale: <https://voyant-tools.org/>
- Mirror:
  - <https://voyant-tools.huma-num.fr>
  - <https://voyant.lincsproject.ca>
  - <https://service.sadilar.org/voyant/>
- Server installabile:  
<https://github.com/voyanttools/VoyantServer>



# 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

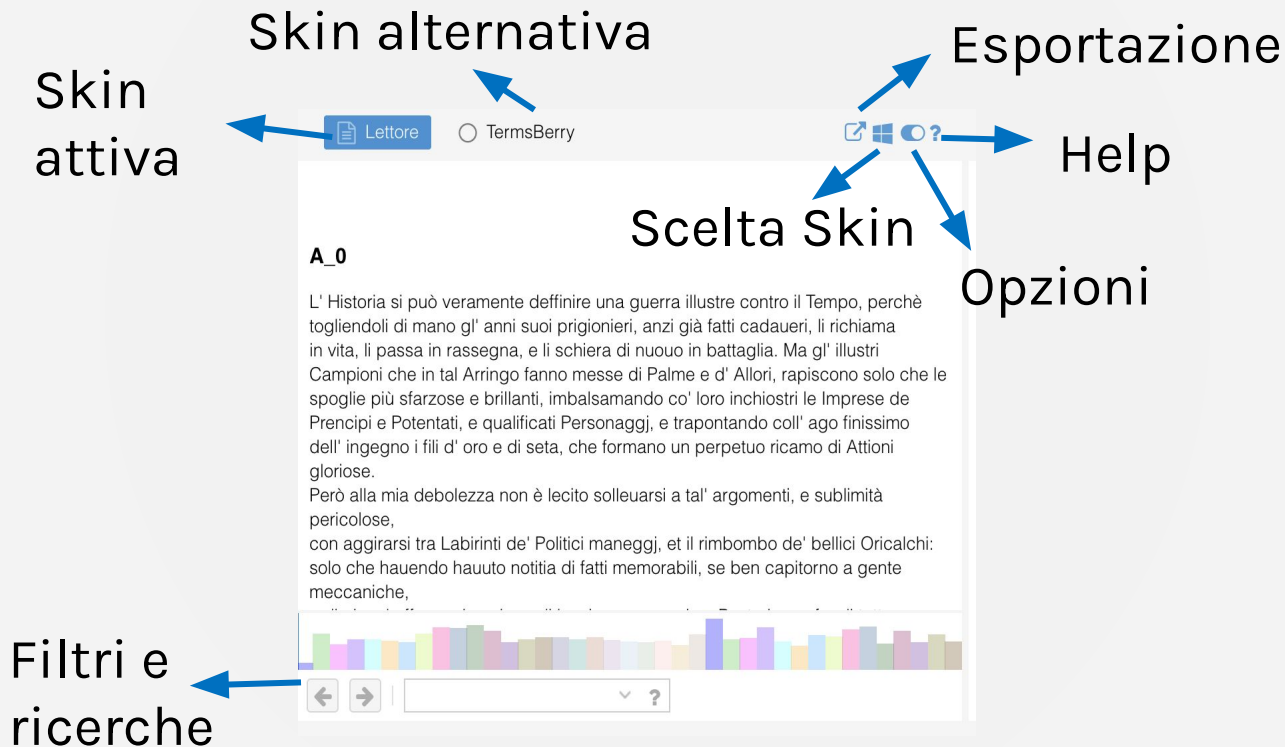
# SKIN



# Struttura della skin




Bembus



# Funzioni principali (1)



Bembus


- 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)



Bembus

- 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



Bembus

- 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



Bembus

- 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*



Bembus

3


# **Esercitazione Pratica**



Bembus

# 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

# ATTENZIONE!

- È corretto considerare i capitoli de “I Promessi Sposi” un corpus?
  - Rappresentatività?
  - Criteri di selezione?



Bembus



# Aggiungere le stopwords (1)




Bembus

- 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

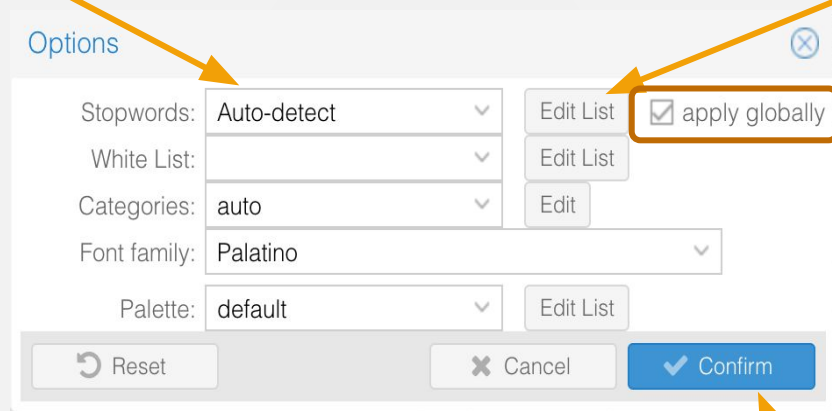
# Aggiungere le stopwords (2)



Bembus

- 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 stopwords (3)



Bembus

- L'effetto su Cirrus

PRIMA




DOPO

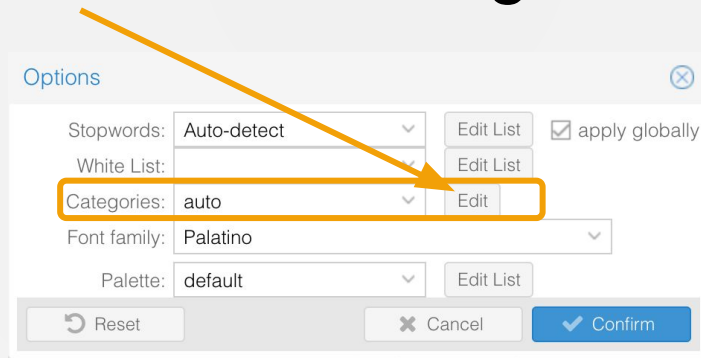




Bembus

# 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**



Options

Stopwords: Auto-detect Edit List ☒ apply globally

White List: Edit List

Categories: auto Edit

Font family: Palatino

Palette: default Edit List

Reset Cancel Confirm

# Aggiungere le categorie (2)



Bembus

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  
advantage  
excellent  
superior  
confidence  
enjoy  
wonderful  
amazing  
enthusiasm  
bliss  
optimistic  
good  
hope  
happy  
happiness  
praise  
safe  
success

bad  
concern  
fail  
despair  
desperate  
disadvantage  
depression  
disaster  
criticize  
suffering  
suffer  
sad  
inferior  
horror  
hesitation  
terrible  
forbidden  
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)



Bembus

Categories Builder

Categories Features

Terms

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

Personaggi

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

Add Category Remove Selected Terms

Cancel Save

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

5) Salvare cliccando su **Save**

# Esempi di utilizzo



Bembus

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\*)

# Provate voi!



Bembus

- 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!!

# Un'analisi più dettagliata (1)



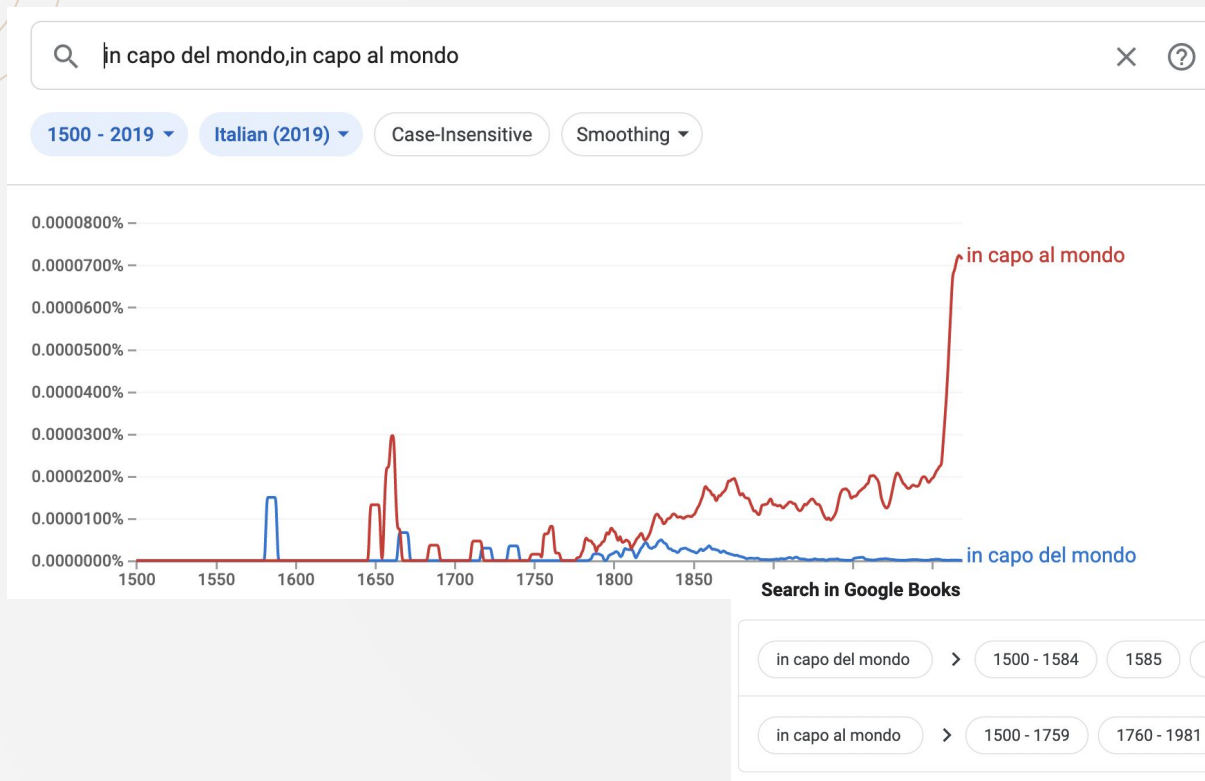
Bembus

- “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)



Bembus





# Analisi di una locuzione (1)



Bembus

- “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](https://www.corpusitaliano.it/it/access/simple_interface.php)



Bembus

# Analisi di una locuzione (2)

## ■ MIDIA

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

Cerca

Uguale a  ×

☐ Espressione regolare

Opzioni avanzate... Storico ricerche

Posizione

Scegli le caratteristiche appartenenti alla stringa precedente o successiva a quella cercata

**Precedente**

Forma

4) Cerca

Reset Cerca



Bembus

# Analisi di una locuzione (3)

## ■ DiaCORIS

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

<b>User Authentication</b> DiaCORIS access is now free for research purposes. (Please, read the footnote carefully).	<b>Query</b> <a href="#">(Query Language Help).</a> <input "addosso"="" [{"0",0}="" type="text" value="man"/> <b>Section</b> <input type="text" value="All"/> <b>SubCorpus</b> <input type="text" value="All"/>
<b>Concordance Options</b> Show <input type="radio"/> 30 <input type="radio"/> 100 <input type="radio"/> 300 <input checked="" type="radio"/> all lines.	<b>Sort position:</b> <input type="text" value="Unsorted"/>
<b>Collocations</b> Get Collocates? <input checked="" type="radio"/> NO <input type="radio"/> Yes.	<b>Sort using</b> <input checked="" type="radio"/> Log-Likelihood Ratio. <input type="radio"/> Mutual Information. <input type="radio"/> T-score. <input type="radio"/> Raw frequency.
<input type="button" value="Esegui"/> <input type="button" value="Cancella"/>	

# Analisi di una locuzione (4)



Bembus

- Paisà (ricerca semplice)

1) Ricerca stringa

2) Applica

Ricerca Semplice   Ricerca Avanzata   Ricerca CQP   Filtri

Corpus: PAISÀ ▾   Cerca: "mani addosso"   applica   ?

☐ restringi la ricerca a frasi semplici   ?

☐ mostra il diagramma delle dipendenze per tutte le frasi

Cliccare qui per esempi di ricerche più complesse!



# Analisi di una locuzione (5)

## ■ Paisà (ricerca avanzata)

1) Ricerca con lemma

“porre”

2) Ricerca con lemma

“mettere”

# Altre risorse utili



Bembus

- Interrogazione/analisi di corpora:
  - Sketch Engine (web, a pagamento): <https://www.sketchengine.eu/>
  - NoSketch Engine (web, gratuito): <https://nlp.fi.muni.cz/trac/noske>
  - AntConc (desktop, gratuito):  
<https://www.laurenceanthony.net/software/antconc/>
- Topic modeling:
  - Online demo: <https://mimno.infosci.cornell.edu/jsLDA/>
  - Mallet: <http://mallet.cs.umass.edu/>



Bembus



# Grazie!

**Domande?**

Mi trovate a [rachele.sprugnoli@unipr.it](mailto:rachele.sprugnoli@unipr.it).

Su Twitter: @RSprugnoli

Per scoprire di più sulle mie ricerche

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