

What's in a corpus?
Utilizing metadata in
Latin and Greek text
collections

Neven Jovanović

University of Zagreb

`neven.jovanovic@ffzg.hr`

Greek and Latin text collections

Greek and Latin

Perseus (internet, free access)

Greek

TLG (Thesaurus linguae Graecae; CD + internet); **PHI** (Greek inscriptions, documentary papyri; CD + internet, commercial)

Latin

Bibliotheca Teubneriana Latina (CD, commercial); *Library of Latin Texts* (CLCLT5; CD, commercial); *PHI Latin library* (CD + internet, commercial); *IntraText Digital Library* (internet, free access); *The Latin Library* (internet, free access); *Itinera electronica* (internet, free access); *Thesaurus Linguae Latinae* (a dictionary; CD, commercial)

What Greek and Latin text collections are not

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 2005)

Maximize number of users
Maximize number of uses

... a library?

But libraries have *catalogues*.
Catalogues *enhance* libraries.

Users of Greek and Latin text collections

Learners

Researchers

A learner's experience

The screenshot shows a web browser window with the URL <http://www.lib.uchicago.edu/efts/PERSEUS/perseuslatin.whizbang.form.html#>. The page header features the Perseus Digital Library logo and the name of the editor-in-chief, Gregory R. Crane, at Tufts University. Navigation links for "Perseus under PhiloLogic" and "User Manual" are present, along with a search bar for text sections. A dark red banner highlights "Perseus Latin Texts". The main content area is divided into "Search in Texts or Find Documents" and "Your query:". The search section includes a search input field, "Search" and "Clear" buttons, and options for display (Context, KWIC, Similarity Search) and search context (Word or Phrase, Phrase separated by 3 words). Bibliographic search fields for Title, Author, and Date are also visible. A navigation bar at the bottom contains links for "More Bibliographic Search Fields", "Refined Search Results", "Text Object Search Fields", and "Info & Help". The "Info & Help" section, titled "PhiloLogic Information and Help", provides a welcome message and search tips, such as using wildcard characters and "Caps Lock" for diacritics. It also includes a link to "Perseus under Philo:" and a paragraph of gratitude to the Perseus Project and its contributors.

PERSEUS DIGITAL LIBRARY
GREGORY R. CRANE, EDITOR-IN-CHIEF
TUFTS UNIVERSITY

[Perseus under PhiloLogic](#) [User Manual](#) Go to text section:

Perseus Latin Texts

Search in Texts or Find Documents **Your query:**

Search for:

Display: Context KWIC Similarity Search

Search Context:
 Word or Phrase Phrase separated by words

Proximity Search in: Sentence Paragraph

Bibliographic Search Fields:
Title: (e.g. 'Aulularia')
Author: (e.g. 'Vergilius')
Date: (e.g. '1899')

[More Bibliographic Search Fields](#) [Refined Search Results](#) [Text Object Search Fields](#) [Info & Help](#)

PhiloLogic Information and Help

Welcome to [PhiloLogic](#). Here are some tips to make your search more productive:

- For pattern matching one may employ wildcard characters (e.g., *widow** retrieves widow, widowe, widowhood, etc.).
- To search without considering diacritics turn on "Caps Lock" and type in all uppercase (e.g., *NAIVETE* finds both naivete and naiveté).
 - But but but: *Greek is accent sensitive!* Use Unicode only.
- Selecting similar word search applies to all words in the database.

[Perseus under Philo:](#)

We are grateful to the [Perseus Project](#) for making their texts available for this project, and specifically to Greg Crane and Adrian Packer for their help in initial troubleshooting. Please note that the conditions of use of Perseus materials fully apply to the texts on this site as well. For details, view the Header information in each of the texts.

A researcher's experience

The screenshot shows a web browser window with the address bar containing the URL `http://www.lib.uchicago.edu/efts/PERSEUS/perseuslatin.whizbang.form.html#`. The page header features the Perseus Digital Library logo, which includes an illustration of a runner, and the text "PERSEUS DIGITAL LIBRARY" and "GREGORY R. CRANE, EDITOR-IN-CHIEF TUFTS UNIVERSITY". Navigation links for "Perseus under PhiloLogic" and "User Manual" are present, along with a search box labeled "Go to text section:" with "Search" and "Clear" buttons. A dark red banner reads "Perseus Latin Texts". Below this, the "Search in Texts or Find Documents" section includes a search input field with "Search" and "Clear" buttons, and radio button options for "Context", "KWIC", and "Similarity Search". The "Search Context" section has radio buttons for "Word or Phrase" and "Phrase separated by" followed by a text input field containing "3" and the word "words". The "Proximity Search in:" section has radio buttons for "Sentence" and "Paragraph". The "Bibliographic Search Fields:" section includes input fields for "Title:", "Author:", and "Date:", each with a "Terms" button and a small example text in parentheses. A navigation bar at the bottom of the search section contains buttons for "More Bibliographic Search Fields", "Refined Search Results", "Text Object Search Fields", and "Info & Help". The "Philologic Information and Help" section is enclosed in a red border and contains a welcome message and a list of search tips. The "Perseus under Philo:" section is also enclosed in a red border and contains a paragraph of text.

PERSEUS DIGITAL LIBRARY
GREGORY R. CRANE, EDITOR-IN-CHIEF
TUFTS UNIVERSITY

[Perseus under PhiloLogic](#) [User Manual](#) Go to text section:

Perseus Latin Texts

Search in Texts or Find Documents **Your query:**

Search for:

Display: Context KWIC Similarity Search

Search Context:
 Word or Phrase Phrase separated by words

Proximity Search in: Sentence Paragraph

Bibliographic Search Fields:
Title: (e.g. 'Aulularia')
Author: (e.g. 'Vergilius')
Date: (e.g. '1899')

[More Bibliographic Search Fields](#) [Refined Search Results](#) [Text Object Search Fields](#) [Info & Help](#)

Philologic Information and Help

Welcome to [Philologic](#). Here are some tips to make your search more productive:

- For pattern matching one may employ wildcard characters (e.g., *widow** retrieves widow, widowe, widowhood, etc.).
- To search without considering diacritics turn on "Caps Lock" and type in all uppercase (e.g., *NAIVETE* finds both naivete and naiveté).
 - But but but: *Greek is accent sensitive!* Use Unicode only.
- Selecting similar word search applies to all words in the database.

[Perseus under Philo:](#)

We are grateful to the [Perseus Project](#) for making their texts available for this project, and specifically to Greg Crane and Adrian Packer for their help in initial troubleshooting. Please note that the conditions of use of Perseus materials fully apply to the texts on this site as well. For details, view the Header information in each of the texts.

A proposal

Design a collection of texts in such a way to:

- a) help learners orientate, and learn what is inside
- b) help researchers ask complex questions

Questions expected

- In which metre are those poems?
- How do I search just the poems in hendecasyllables?
- Which texts in the collection are letters?
- How do I search just the letters in the collection?
- Which texts in the collection were produced in first century b. C?
- How do I search just the texts produced in first century b. C?

Problems expected

- There are too many texts!
- How do we find metadata?
- How do we actually *do* it?
- Where do we find an army of coders?

What is already around?

(Old) scholarship as source of metadata

http://books.google.com/books?id=_lQAAAAIAAJ&printsec=frontcover#PPA199,M1

perseus philologic

Google Book Search Search Books

filologanoga@gmail.com | My library | My Account | Sign out

Bibliotheca Classica Latina sive Collectio Auctorum Classicorum Latinorum ... By Nicolas-Éloi Lemaire

Page 199 Full screen

EPISTOLÆ, XVI, 23. 199

EPISTOLA XXIII.

Hortatur, ut professionem conficiat, scribitque sibi in animo esse amicitiam Antonii conservare. (Puteolis, anno U. C. 709.)

TULLIUS TIRONIS S. P. D.

Tu vero confice professionem¹, si potes: etsi hæc pecunia ex eo genere est, ut professione non egeat. Verumtamen² Balbus ad me scripsit, tanta se epiphora oppressum, ut loqui non possit³, Antonius de lege quid egerit. Liceat modo rusticari⁴. Ad Bithynicum scripsi. De Servilio⁵ tu videris, qui senectutem non contemnis. Etsi Atticus noster, quia quondam me commoveri πανιχοις⁶ intellexit, idem semper putat, nec videt, quibus præsidiis philosophiæ septus sim; et hercle, quod timidus ipse est, θορυβοποιει⁷. Ego tamen Antonii inveteratam sine ulla offensione amicitiam retinere sane volo, scribamque ad eum, sed non ante,

1. *Professionem. Proferi*, est hic, publice denunciare, quidquid pecunie, agri, aliarumve rerum Cicero possideat. Et videtur, ut hæc per Tironem, sic altera superioris anni professio per Philotimum esse facta. Vide epist. 33 lib. XIII ad Att. MAN.

2. *Verumtamen. Subaud. confice*; et hanc ipsam pecuniam profitere.

banis rebus, quas Antonius consul e sua libidine administrat, antepono. IDEM.

5. *De Servilio. Isaurico*, quem hoc anno obiisse testatur Dio lib. XLV, quem felicem videtur scripsisse Tiro, qui effugerit Antonii minas et gladios. Id ad me, inquit Cicero, non pertinet, qui cædem non timeo, et vitæ

Download PDF - 29.7M
View plain text

About this book
Edit review
In my library [Remove]

Contents

Popular passages

Buy this book
Abebooks
Alibris
Amazon
Barnes&Noble.com
Google Product Search

Find this book in a library

Search in this book
Go

Other editions
Titi Lucretii Cari De rerum ...
1838
Titi Lucretii Cari De rerum ...
1838
show more »

Basic HTML mode

Chicago Homer

http://www.library.northwestern.edu - The Chicago Homer - Mozilla Firefox

SEARCH BROWSE HELP CORRECTIONS OPTIONS HOME

SEARCH FOR: WORDS PHRASES

WITH THE FOLLOWING CHARACTERISTICS:

- SEARCH TERMS
- FREQUENCY
- WORD TYPE
- INFLECTIONAL CATEGORIES
- LINE RANGES
- NARRATIVE SPEECH /SPEAKERS

REPORT RESULTS AT THE LEVEL OF: LEMMA WORDFORM

RESET SEARCH

QUERY FORM WORD LIST CONCORDANCE REPETITIONS

[Search criteria relating to speech and narrative](#)

Within Speech or Narrative:

Within speeches with the following speaker characteristics:

Mortality: Gender:

Within speeches by specific speakers:

Select desired speakers from the list below and then click the 'Add Speaker to Constraints' button to add that speaker to your search constraints.

Speaker List:

Achaian	▲
Achaian and Trojan	■
Achilleus	
Adrestos	
Agamemnon	
Agelaos	
Ageleos	
Agenor	
Aiante	
Aias	
Aias the Lesser	
Aigyptios	▼

You may remove any speakers from your search constraints by selecting that speaker in the list below and clicking the 'Delete Speaker from Constraints' button.

Done

[emacs22@furfifer.v...] [knjige-tif - Preglednik...] Welcome to the Chica... http://www.library.nort...

[Search criteria relating to speech and narrative](#)

Within Speech or Narrative:

Within speeches with the following speaker characteristics:

Mortality: Gender:

Within speeches by specific speakers:

Select desired speakers from the list below and then click the 'Add Speaker to Constraints' button to add that speaker to your search constraints.

Speaker List:

Achaian	▲
Achaian and Trojan	≡
Achilleus	
Adrestos	
Agamemnon	
Agelaos	
Ageleos	
Agenor	
Aiante	
Aias	
Aias the Lesser	
Aigyptios	▼

You may remove any speakers from your search constraints by selecting that speaker in the list below and clicking the 'Delete Speaker from Constraints' button.

TLG / PHI with Diogenes

Diogenes

Διογένης

TLG Classification

Here are the various criteria by which the texts contained in the TLG are classified.

You may select as many items as you like in each box. Try holding down the control key to select multiple items.

1. Author's genre:

-
- Anthologia
- Apocrypha
- Apologetica
- Astrologica
- Astronomica
- Biographa
- Bucolica

2. Text genre:

-
- Acta
- Alchemica
- Anthologia
- Apocalypsis
- Apocrypha
- Apologetica
- Astrologica

3. Location:

-
- Abdera
- Adramytteum
- Aegina
- Aegyptus
- Aetolia
- Agrigentum [vel Acragas]
- Alexandria

4. Gender:

-
- Femina

5. Name of Author(s):

6. Date Range:

After

Before

Include Varia and Incerta?

You may select multiple values for as many of the above criteria as you wish.

Then indicate below how many of the stipulated criteria a text must meet in order to be included in the search.

Number of criteria to match:

All

TLG / PHI with Diogenes

1. Author's genre:

--

- Anthologia
- Apocrypha
- Apologetica
- Astrologica
- Astronomica
- Biographa
- Bucolica

2. Text genre:

--

- Acta
- Alchemica
- Anthologia
- Apocalypsis
- Apocrypha
- Apologetica
- Astrologica

3. Location:

--

- Abdera
- Adramytteum
- Aegina
- Aegyptus
- Aetolia
- Agriantum [vel Acragas]
- Alexandria

4. Gender:

--

Femina

5. Name of Author(s):

6. Date Range:

After --

Before --

Include Varia and Incerta?

Perseus under PhiloLogic



Perseus Latin Texts

Search in Texts or Find Documents

Search for:

Display: Context KWIC Similarity Search

Search Context:

Word or Phrase Phrase separated by words

Proximity Search in: Sentence Paragraph

Bibliographic Search Fields:

Title: (e.g. 'Aulularia')

Author: (e.g. 'Vergilius')

Date: (e.g. '1899')

Your query:

Enter search criteria to form a new search.

[More Bibliographic Search Fields](#) [Refined Search Results](#) [Text Object Search Fields](#) [Info & Help](#)

Text Object Search Fields

Find documents or limit word searches (use OR only).

Div Objects

Head:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. '[commline]')
Type:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'commline')
Lang:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'en')
N:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. '9')
Id:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'p3111')
Ocauthor:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'GEBBIE CO.')
Ocdateline:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'Scr. Romae (?) m. Apr. 710 (44)')
Ocsalutation:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'TO C. MUNATIUS (IN A PROVINCE)')

SubDiv Objects:

Tag:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'note')
Type:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'Scazons')
N:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. '43')
Id:	<input type="text"/>	<input type="button" value="Terms"/>	(e.g. 'ref21014')

Vindolanda tablets online

you are here: [home](#) | [Tablets](#) | [Browse](#)

Search database

View tablet number
(118-573):

[View all Tablets](#)

[Search Tablets](#)

[Browse Tablets](#)

[Tab. Vindol. II
Introductory
chapters](#)

[Tab. Vindol. II
Category
introductions](#)

[Tab. Vindol. II
Abbreviations and
Bibliography](#)

[Digitising
Vindolanda](#)

[Tab. Vindol. II
Addenda and
Corrigenda](#)

[Tab. Vindol. I
Introductory
Chapters](#)

[The print
publication and
the online edition](#)

[Print-friendly](#)

Browse by:

[Highlights](#)

- All highlights
- Families, pleasures and ceremonies
- Necessities of life
- Letters to make, keep (or lose) friends
- Military matters

[Number](#)

Tablet number in the published sequence

[Subject](#)

General subjects, e.g. "Food", "Clothing", "Utensils"

[Category](#)

Chapter headings, e.g. "Correspondence of Cerialis"

[Type](#)

Document type, e.g. "letter" or "list"

[People](#)

People mentioned in tablets, e.g. "Genialis", "Lepidina"

[Places](#)

Places mentioned in the tablets, e.g. "Coria"

[Military terms](#)

Latin military terms e.g. "centuria"

[Archaeological context](#)

The archaeological context (level and room) where the tablet was found. e.g. "Period 3, room WVIA"

[TVI number](#)

TVI (first edition) tablet numbers in the published sequence

Croatiae auctores Latini (CAuLa)

- ca. 300.000 words pilot
- short texts, long texts, poetry, prose, literature, functional texts (e. g. notarial documents)
- until now: uncentralised, undigitised, sometimes unindexed, not easily (world-wide) accessible or searchable, not always reliably edited...

Croatiae auctores Latini (CAuLa)

Search and browse by:

- Auctores (A-Z)
- Tempora (e. g. 1400-1950)
- Loca (e. g. Dubrovnik, Split, Trogir)

Croatiae auctores Latini (CAuLa)

Search and browse by:

- Genera
 - Poesis
 - Prosa

Croatiae auctores Latini (CAuLa)

- Genera
 - Poesis
 - epica
 - elegiaca
 - epigrammata
 - eclogae
 - saturaе

Croatiae auctores Latini

(CAuLa)

- Themata
 - funeraria
 - amicitia
 - amores
 - antiturcica
 - ...

Croatiae auctores Latini (CAuLa)

- Damjan Beneša (Dubrovnik, around 1500),
 - *De morte Christi* (10 books, 8300+ verses)
 - Liber I
 - Opening scene (*vv. 1-30*). Before Easter: everywhere sorrow. The poet thinks about faraway places, about Christ's passion and death. Jerusalem: Christ is being taken to Pilates' palace. The poet sees a vision of Christ hanging on the cross, his Mother grieving
 - Invocation (*vv. 31-43*): one who sings about Christ will earn a place in heaven; why did the Virgin bear a son, etc.

CAuLa: sample queries

What did people write about when they wrote in Latin in Split between 1500 and 1600?

How did poetry about friendship look like in Dubrovnik between 1500 and 1600?

In what types of texts is word *arma* used? Are there types of texts that do not use this word?

...

What do ~~we~~ I need?

- *Caveat*: a theoretically simple task may get quite untractable in real life (standards? searches? references? openness? computer science? etc.)
- If possible, use tools that already exist (learn about them)
- If possible, connect with projects that already exist (idem)
- Attract users, who will also help keep the project alive (corrections? reviews? research? teaching?)
- Hear what others think!