

# Metodi computazionali per l'indagine lessicale su testi latini medievali

Un'applicazione al Codex  
Diplomaticus Cavensis

**II<sup>e</sup> Atelier Informatique et Histoire**  
***L'historien, le texte et l'ordinateur***  
**27-28 novembre 2006 – École Normale LSH de Lyon**


LAB.I.U.M. - Laboratorio di Informatica UManistica - Università di Verona - Benvenuti sul sito W - Windows Internet Explorer

http://www.cyllenius.net/labium/

# LAB.I.U.M.

## LABoratorio di Informatica UManistica

Dipartimento di Linguistica Letteratura e Scienze della Comunicazione



Home News Contatti Links Cerc

### MENU PRINCIPALE

- Home
- Ricerca
- Didattica
- Collaborazioni
- Latin WordNet
- Pubblicazioni
- News
- Links
- News Feeds
- Cerca
- Downloads
- Gregorio di Tours
- Staff
- Contatti

### Benvenuti sul sito WEB di LAB.I.U.M.

Scritto da Stefano Minozzi  
Monday 12 September 2005

Il **LAB**oratorio di **I**nformatica **U**Manistica è costituito da un gruppo di ricerca (**staff**) fondato e diretto dal prof. Antonio De Prisco in seno al Dipartimento di Linguistica, Letteratura e Scienze della Comunicazione dell'Università degli Studi di Verona. Questo sito nasce con l'obiettivo di diffondere materiali scientifici prodotti nel contesto dell'attività del laboratorio negli ambiti dell'Informatica umanistica, della Linguistica computazionale e delle Tecnologie didattiche.

### LAB.I.U.M. svolge attività di ricerca nell'ambito dei seguenti settori:

- digitalizzazione, codifica e marcatura delle risorse testuali;
- sviluppo di modelli per l'analisi computazionale dei testi letterari;
- realizzazione di strumenti per l'*information retrieval* nelle collezioni di testi mediolatini;
- redazione di ipertesti per la ricerca e per la didattica univertitaria della lingua e della letteratura latina, classica e medievale;
- didattica del latino con l'ausilio delle nuove tecnologie.

Ultimo aggiornamento ( Saturday 25 November 2006 )  
[\[Indietro\]](#)

### LOGIN

Nome Utente  
Password  
 Ricordami  
[Login](#)

[Hai perso la password?](#)  
Non hai ancora un account?  
[Creane uno!](#)

### WHO'S ONLINE

Abbiamo 4 visitatori online

### STATISTICHE

**Utenti:** 60  
**Notizie:** 21  
**WebLinks:** 78  
**Visitatori:** 9261

### SYNDICATE

[RSS 1.0](#)  
[RSS 2.0](#)  
[ATOM 0.3](#)  
[OPML SHARE IT!](#)

W3C XHTML

Internet 100%

12.51 sabato

# http://www.cyllenius.net/labium/



UNIONE ACCADEMICA NAZIONALE

UNIVERSITÀ DI MILANO  
UNIVERSITÀ DI NAPOLI FEDERICO II  
UNIVERSITÀ DI PALERMO  
UNIVERSITÀ DI ROMA TRE  
UNIVERSITÀ DI VENEZIA CA' FOSCARI  
UNIVERSITÀ DI VERONA

**ALIM**  
ARCHIVIO DELLA  
LATINITÀ ITALIANA DEL MEDIOEVO

## Archivio della Latinità Italiana del Medioevo

[Il Progetto ALIM](#)



[Fonti letterarie](#)



[Fonti documentarie](#)

[Login](#)





UNIVERSITÀ DI MILANO  
UNIVERSITÀ DI NAPOLI FEDERICO II  
UNIVERSITÀ DI PALERMO  
UNIVERSITÀ DI ROMA TRE  
UNIVERSITÀ DI VENEZIA CA' FOSCARI  
UNIVERSITÀ DI VERONA



## Archivio della Latinità Italiana del Medioevo

### Fonti letterarie

[Indice degli Autori](#)

[Indice dei Generi](#)

[Indice dei Titoli](#)

[Opere in Prosa](#)

[Indice dei Periodi](#)

[Opere in Versi](#)



[Ricerche](#)

[Home](#)

[Grafia](#)

[Help](#)



UNIONE ACCADEMICA NAZIONALE

UNIVERSITÀ DI MILANO  
UNIVERSITÀ DI NAPOLI FEDERICO II  
UNIVERSITÀ DI PALERMO  
UNIVERSITÀ DI ROMA TRE  
UNIVERSITÀ DI VENEZIA CA' FOSCARI  
UNIVERSITÀ DI VERONA

# ALIM

ARCHIVIO DELLA  
LATINITÀ ITALIANA DEL MEDIOEVO

## Archivio della Latinità Italiana del Medioevo

### Fonti documentarie

[Indice delle raccolte per denominazione](#)

[Indice dei documenti per raccolta](#)

[Indice delle raccolte per area geografica](#)

[Indice dei documenti per luogo](#)

[Indice delle raccolte per copertura cronologica](#)

[Indice dei documenti per data](#)



[Ricerche](#)

[Home](#)

[Grafia](#)

[Help](#)



FRANCISCUS ARNALDI

PASCHALIS SMIRAGLIA

**LATINITATIS ITALICAE  
MEDII AEVI  
LEXICON**

(saec. V ex. - saec. XI in.)

Editio altera



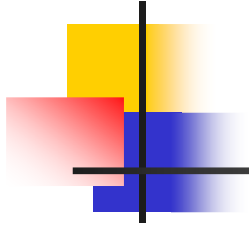
SISMEL  
EDIZIONI DEL GALLUZZO



## Computational analysis. Benefits:

---

- Quantitative support to research hypothesis
- Collection of evidences for hypothesis formulation



- Diagnostic function
- Prognostic function

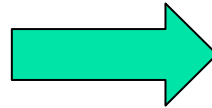




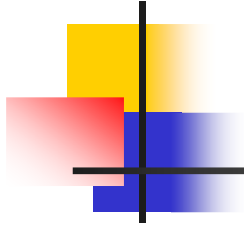
## Use in lexicography

---

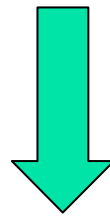
- Text retrieval
- Text mining



**Dictionary  
building**



- 
- Analysis of word usage



**"Snapshot"**  
**of lexical level of a text**



# Can be unveiled:

---

- Permanences in linguistic habits
- Geographical localization
- Chronological localization



# Codex diplomaticus Cavensis

---

- Model of Middle Latin in Campania (Salerno)
- Model of Middle Latin between the VIII and the XI Century



# Building of lexical sample

---

- Dimension of the sample (20MB txt)
- Structure (430 documents of similar length)
- Balancing of parts (building of regular partitions)





## Two “lexical worlds”

---

- City of Salerno: 214 documents;
- Countryside of Salerno: 216 documents;

# Text pre-processing

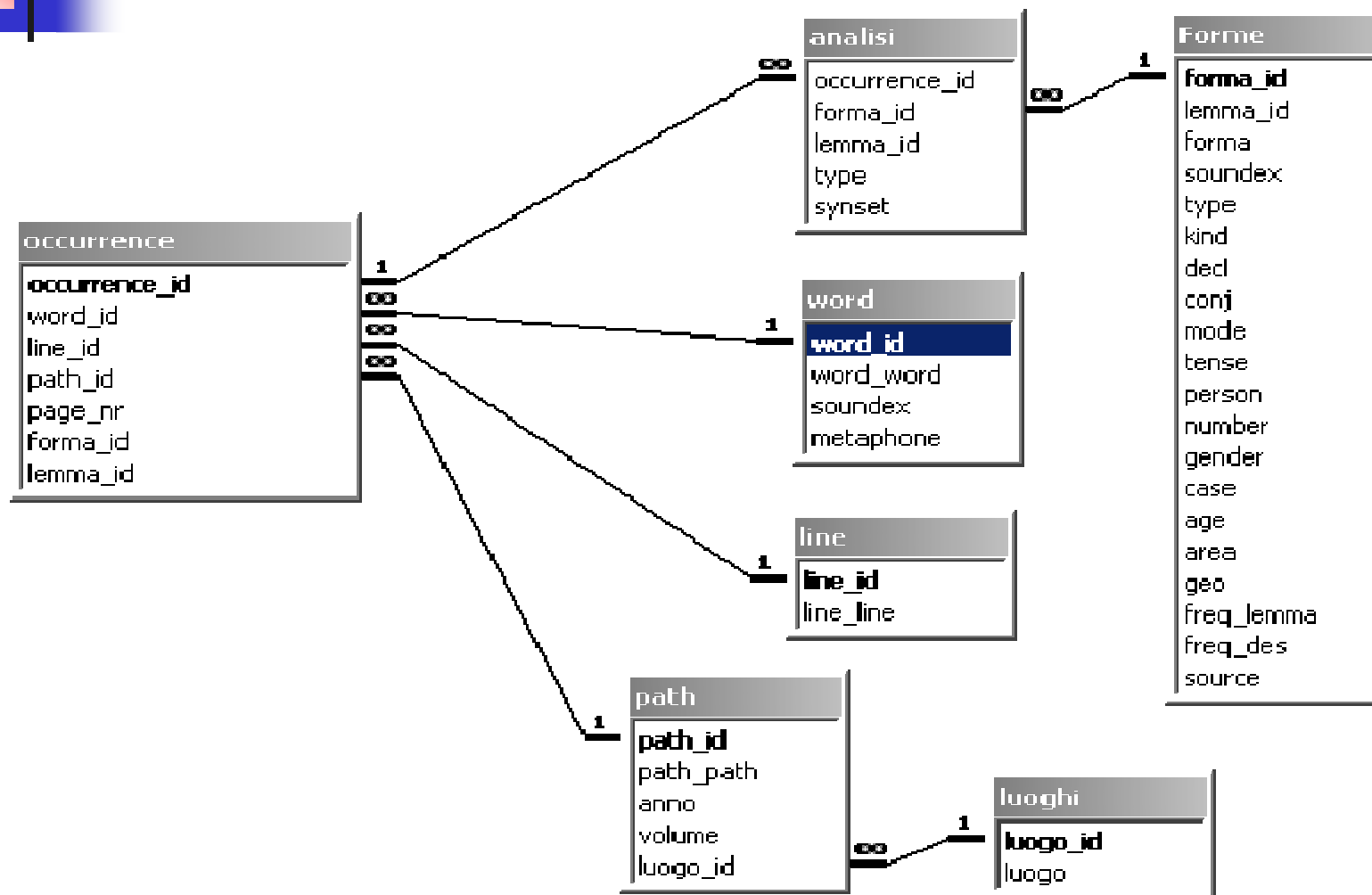
## ***TAGGING OF ESSENTIAL METADATA:***

- **COLLOCATION**
- **DATE**
- **PLACE**

```
<meta vol="1">  
<meta anno="792">  
<meta luogo="Forini">
```

```
1  
+ In nomine domini quinto anno principatum domni nostri  
dux gentis langubardorum mense octobri per indictione p  
consilio et boluntates aldefusi genitori meo havitatori  
quondam roderisi de nuceria in meo sociavit coniugio, t  
et amicus nostros trado adq̄e tradedit tive suprascripte  
portionem substantie mee quod me iuxto hordine a germar  
case quam et intrinsicus case, curte, territoriiis, vine  
planus et in monte, homnia et de omnibus de quitquit vi  
meam una tivi tradedit ad possidendum. et per hunc hell  
rovoratum ostendere previdimus sicundum ritum genti nos  
  
2  
foturo pro hanc causa periurio non percurrat et secundu  
quartam rationem una tive que supra uxori mee tradedit  
in tua sit potestate, quam hanc chartulam morgincaput i  
scribendum rogavimus. Actum forinese mensi et indictior  
signum + manus alderisi viro eius et aldefusi genitore  
morgincaput communiter fierit rogavimus.  
+ MELONIANUS ME TESTE subscripsi.  
+ ego maio filio meloniani sculdais me teste subscribis  
+ ego aceprandu filiu meloniani me teste subscripsi.  
+ ego ermoald notarius rogatus a suprascripti me teste
```

# Pre-processing builds a multidimensional index





# Segmentation of text

---

- Documents
- Phrases
- Lexical units

# Custom interactive software provides support for identification of heterographs

Unknown word:  
"venumdetit"

Suggested word:  
"venumdedit"

The screenshot shows a software interface with a window titled "Indicizzatore" and a sub-window titled "Disambiguazione allografi". The sub-window contains the text "venumdetit" and a table with the following data:

forma_id	lemma_id	forma	soundex	type	kind	decl
1180177	36681	venumdedit	V553	V	TRANS	1.1

Arrows from the text boxes point to the input field and the first row of the table, respectively.



Problems in identification:

*homographs* and *heterographs*



**NEED**

*disambiguation*

**NEED**

*assimilation*



# Kinds of heterographs:

---

- Paradigmatic heterographs
- Orthographic heterographs
- Dialectal and stylistic heterographs
- Syntagmatic heterographs

# Morphological lemmatization

occurrence

occurrence\_id

word\_id

line\_id

path\_id

page\_nr

forma\_id

lemma\_id

Each word occurrence is tagged by the pre-processing engine, making assumption to which lemma each form belongs

	occurrence_id	word_id	line_id	path_id	forma_id	lemma_id
▶	149880	3483	4506	357	PREP1086692	36359 36358
		forma_id	lemma_id	type		
▶		700419	36358	ADV		
		1086692	36359	PREP		
*						
+	149881	2680	4506	357	PREP1086645	706 705
+	149882	2794	4506	357	PREP1086713	19647 19652 19
+	149883	2729	4506	357	PRON1087401	21429
+	149884	2732	4506	357	V1851785 PRO	30951 30733 30
+	149885	4408	4506	357	V1174562	37369
+	149886	6873	4506	357	?	?
+	149887	5899	4506	357	N979637 N9665	16844 16844 16
+	149888	2647	4506	357	PRON1087333	30747 30746 30
+	149889	3559	4506	357	PRON1087505	21535
+	149890	2651	4506	357	PREP1086697	3 4 2 1 5
+	149891	5673	4506	357	N818498 N811E	27478 27478



# Orthographic heterographs

---

- Affect only the graphic form
- Are consequence of editorial choice (distinction U/V; jam/iam etc.)



## Heterographs for “convenientia”:

---

- comvenientia
- combenientia
- conbenientia
- comvenihentia
- convenihentia
- cumvenihentia





# Syntagmatic heterographs

---

- Are effect of *sandhi* (“chaining of words”):
  - *Ecceos, viden, bonast ecc.*



# Homographs

---

- Intralemmatic homographs → same word, same spelling, different function (*not a problem if our interest is only in quantitative analysis of lemma's usage*)
- Esalemmatic homographs → same spelling, different word (*can be a problem in quantitative analysis*)



# Statistic analysis

---

- Two parameters:
  - Frequency of words in the “lexical universe”
  - Dispersion of words in the partitions of the “lexical universe”



## Dispersion as standard deviation from the mean

---

$$D = 1 - \frac{\sigma}{2\bar{x}}$$

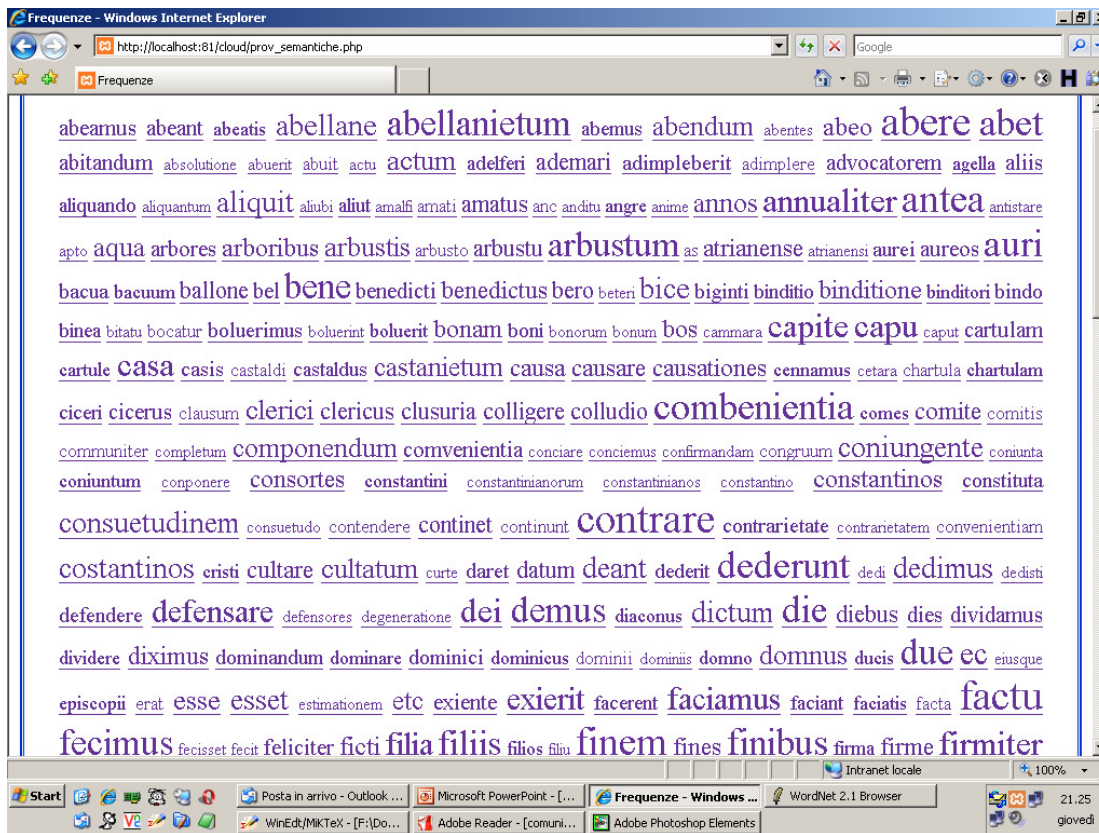
Standard deviation of occurrences of a word in partitions of the "lexical universe"

Mean value of occurrences in partitions of the "lexical universe"

$D = 0 \rightarrow$  dispersion *pessima*

$D = 1 \rightarrow$  dispersion *optima*

# A sort of game... a "tag cloud" based on frequency and dispersion:



Usage of word can be seen as a combination of both parameters:

$$U = \frac{FD}{100}$$





# Analysis of co-occurrences

---

- Interest for lexicographers:
  - Better understanding of contexts of use
- Interest for historians:
  - Reconstruction of material culture



## Cosine calculation for co-occurrent words

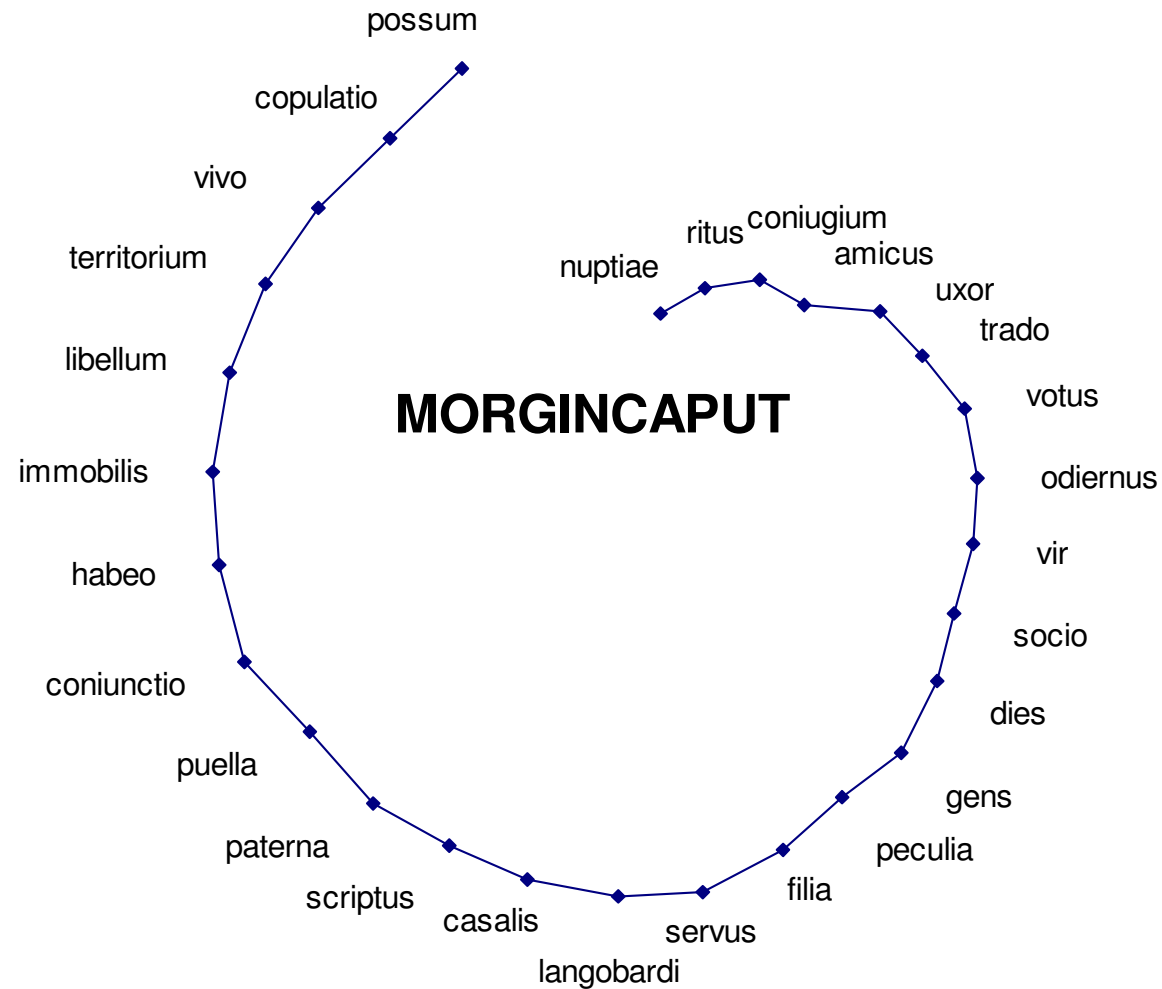
---

$$C(X, Y) = \frac{X \cap Y}{\sqrt{X} \sqrt{Y}}$$

# First 30 words co-occurrent with "guadia" (distance is inversely proportional to association)

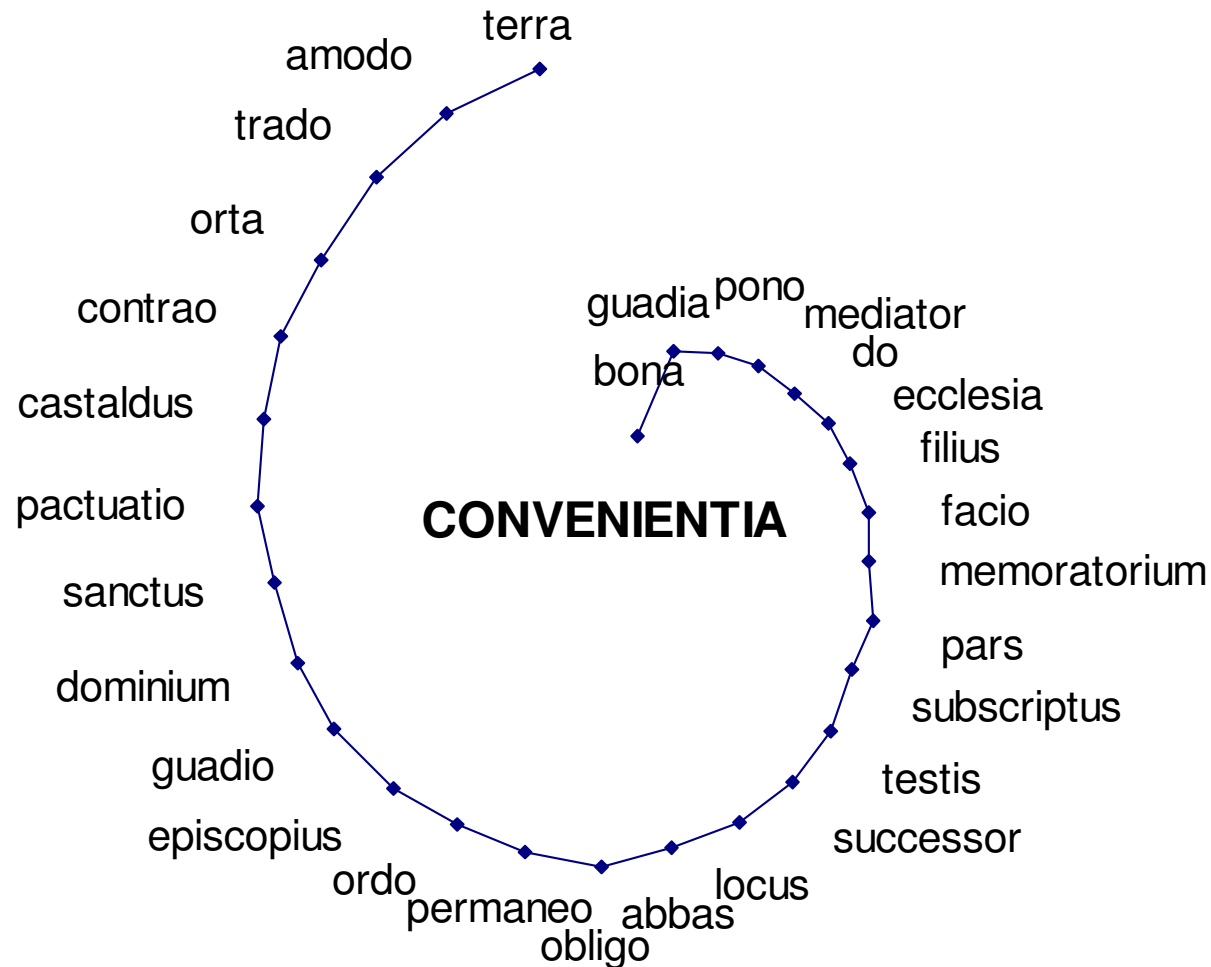


# First 30 words co-occurrent with "morgincaput" (distance is inversely proportional to association)



# First 30 words co-occurrent with "convenientia"

(distance is inversely proportional to association)



# First 30 words co-occurrent with "bonus"

(distance is inversely proportional to association)

