Linguistic Networks

Alexander Mehler, Rüdiger Gleim, Alexandra Ernst, Andy Lücking

Networking Words

First hits of a Google picture search for German “Bank”:

















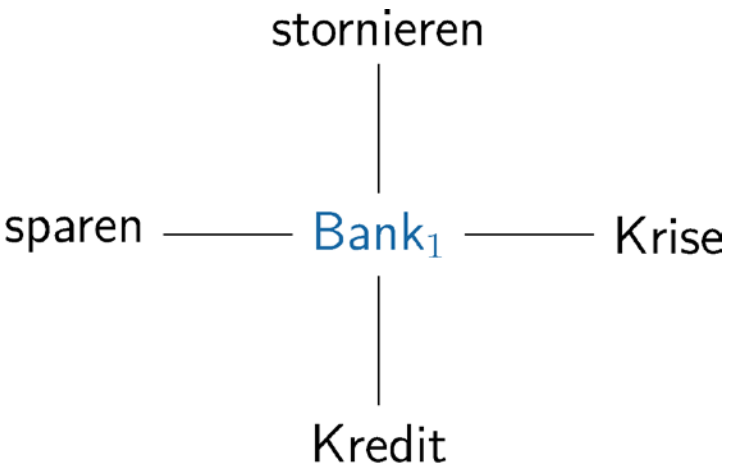


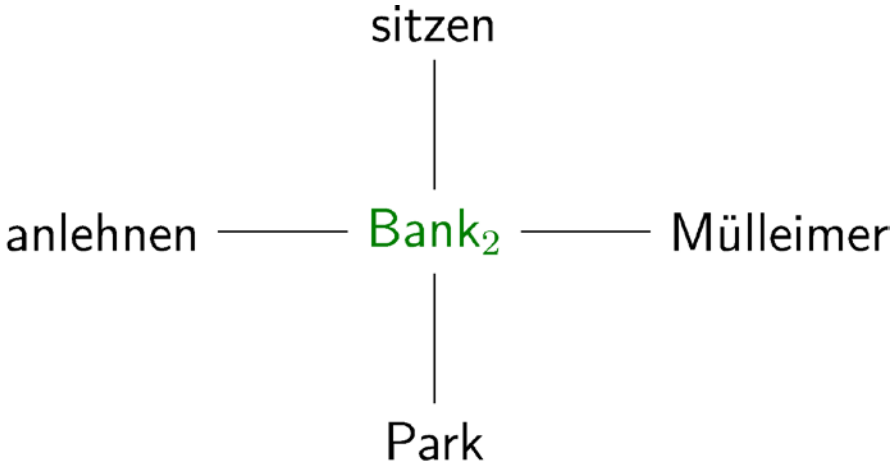
Networking Words

Neighbors of “Bank”:

User Jleon, CC-By-SA

Florian Hurlbrink, CC-By-SA





Agenda

1. Corpus Example: Patrologia Latina
2. eLexicon
3. Time Series
4. Sample Analyses
5. Linguistic Networks Workflow

The Corpus

Patrologia Latina

- Compiled by Jaques Paul Migne
- Latin documents from the 4th to the 13th century
- Multiple text types
- Digitized by Mark D. Jordan and colleagues (since 1993)

The Corpus

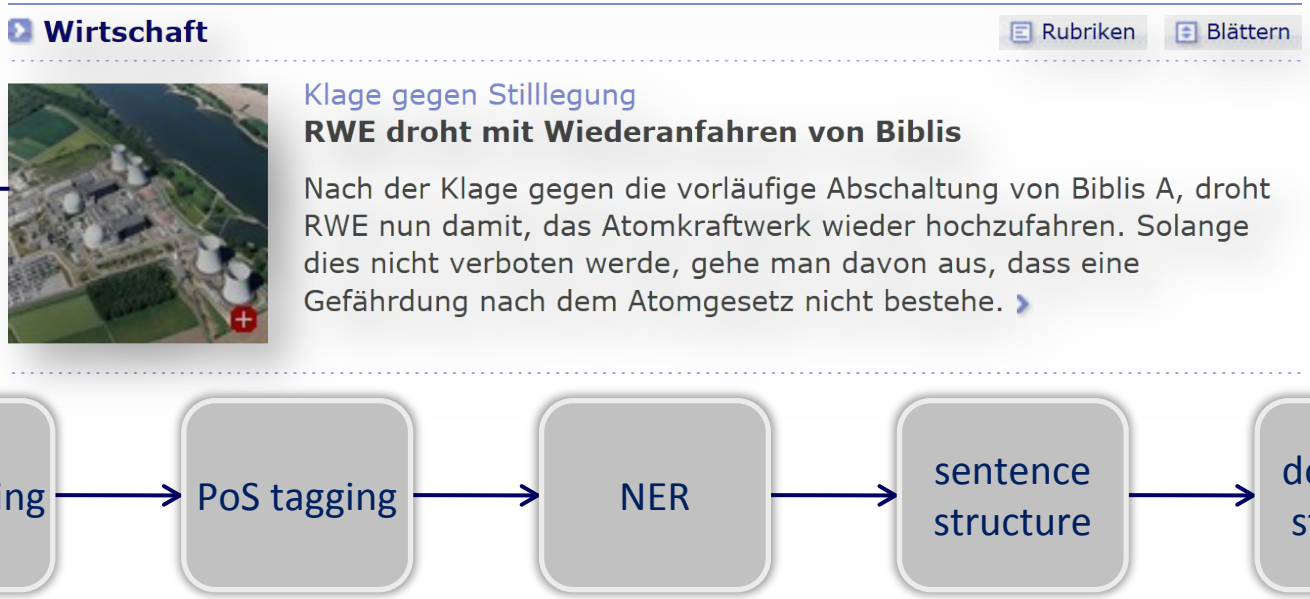
Patrologia Latina

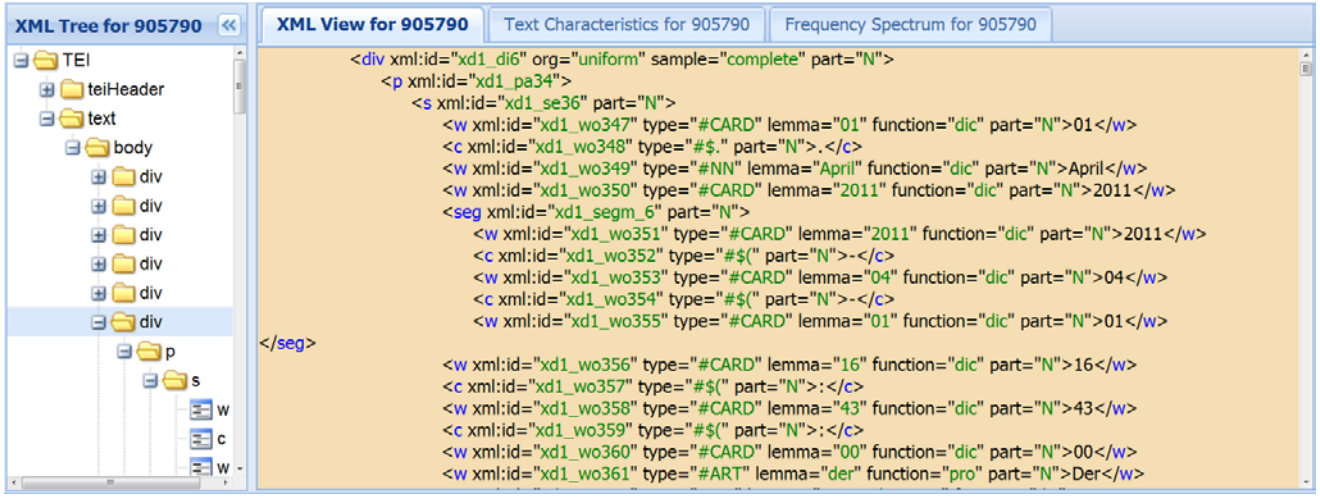
Object	Number
author	3,586,131
text	875,404
paragraph	10,431,961
sentence	56,677,686
word form	4,592,145
token	436,439,087

Object	Number
author	1,320
text	4,555
paragraph	674,718
sentence	7,727,864
word form	1,094,850
token	121,722,687

Preprocessing – Corpus

Beyond Conversion: Preprocessing Steps



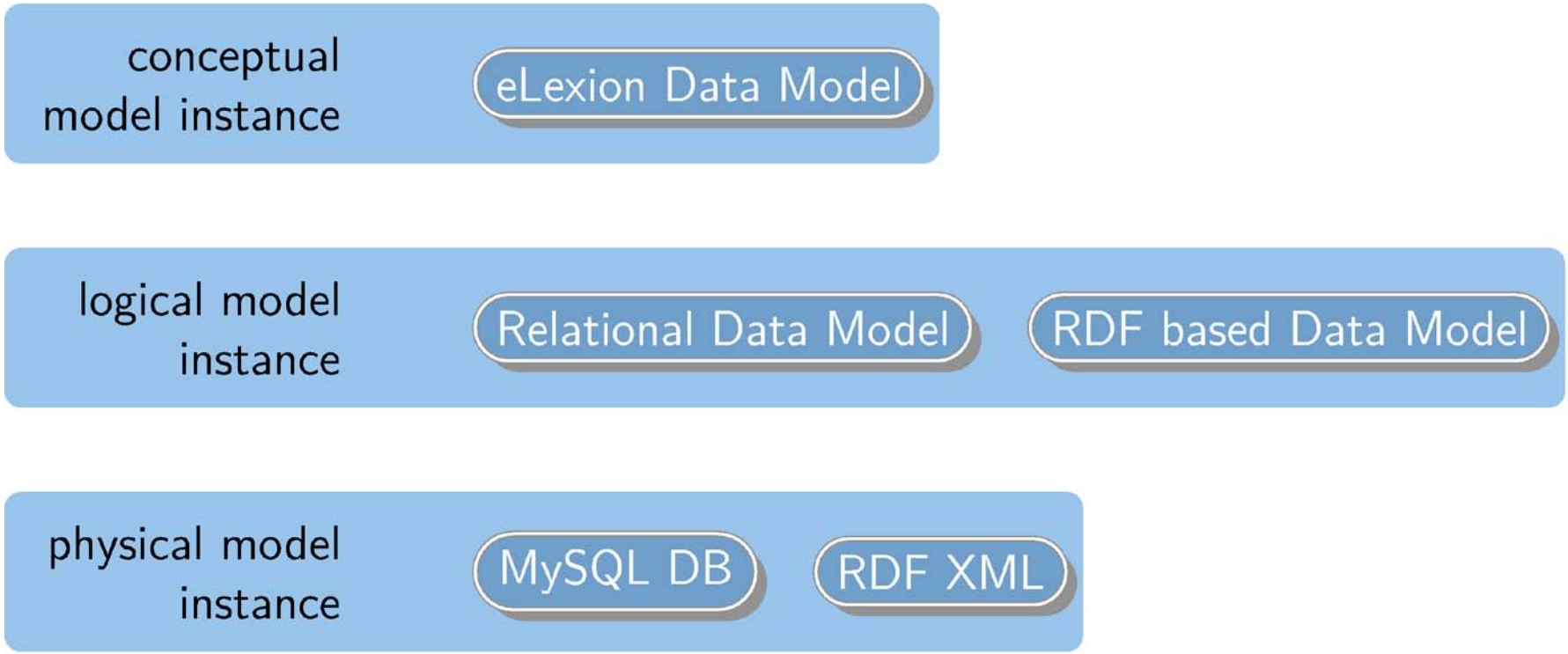


1. Corpus Example: Patrologia Latina
2. eLexicon
3. Time Series
4. Sample Analyses
5. Linguistic Networks Workflow

eLexicon Data Model Architecture

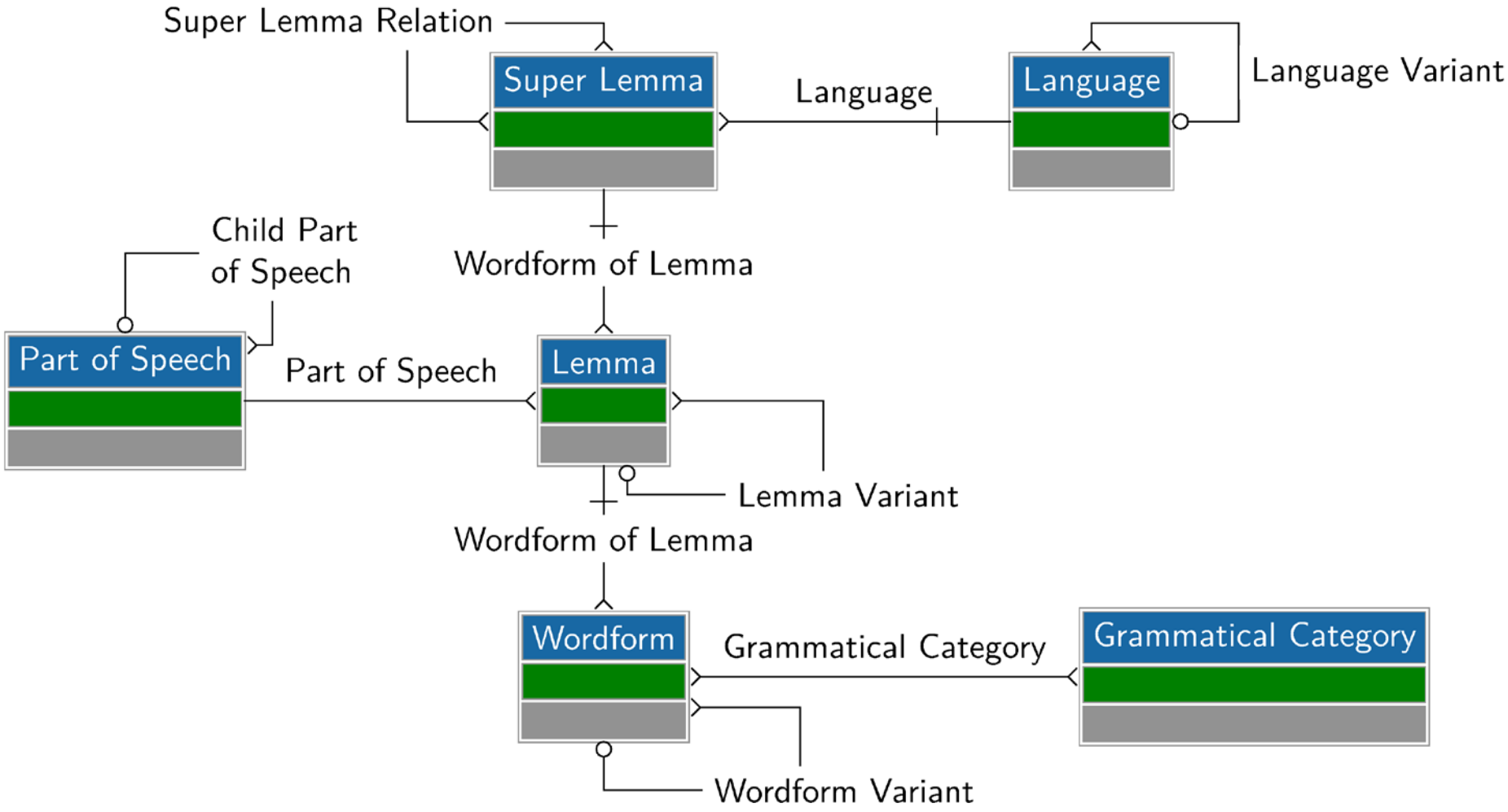
Three layers

ANSI/X3/SPARC Study Group on Data Base Management Systems (ANSI, 1975)



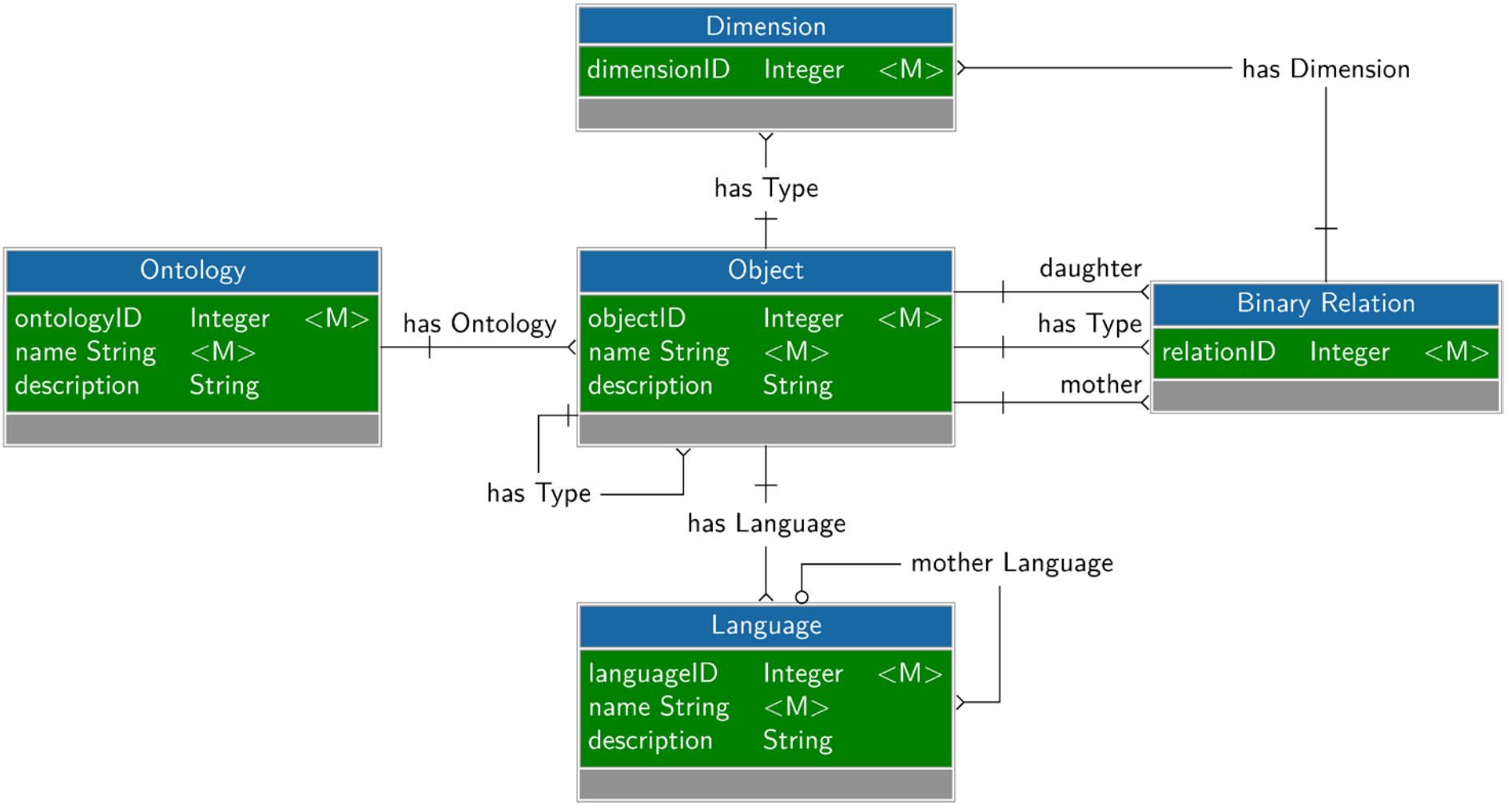
eLexicon Conceptual Data Model

Simplified Entity-Relationship Diagram



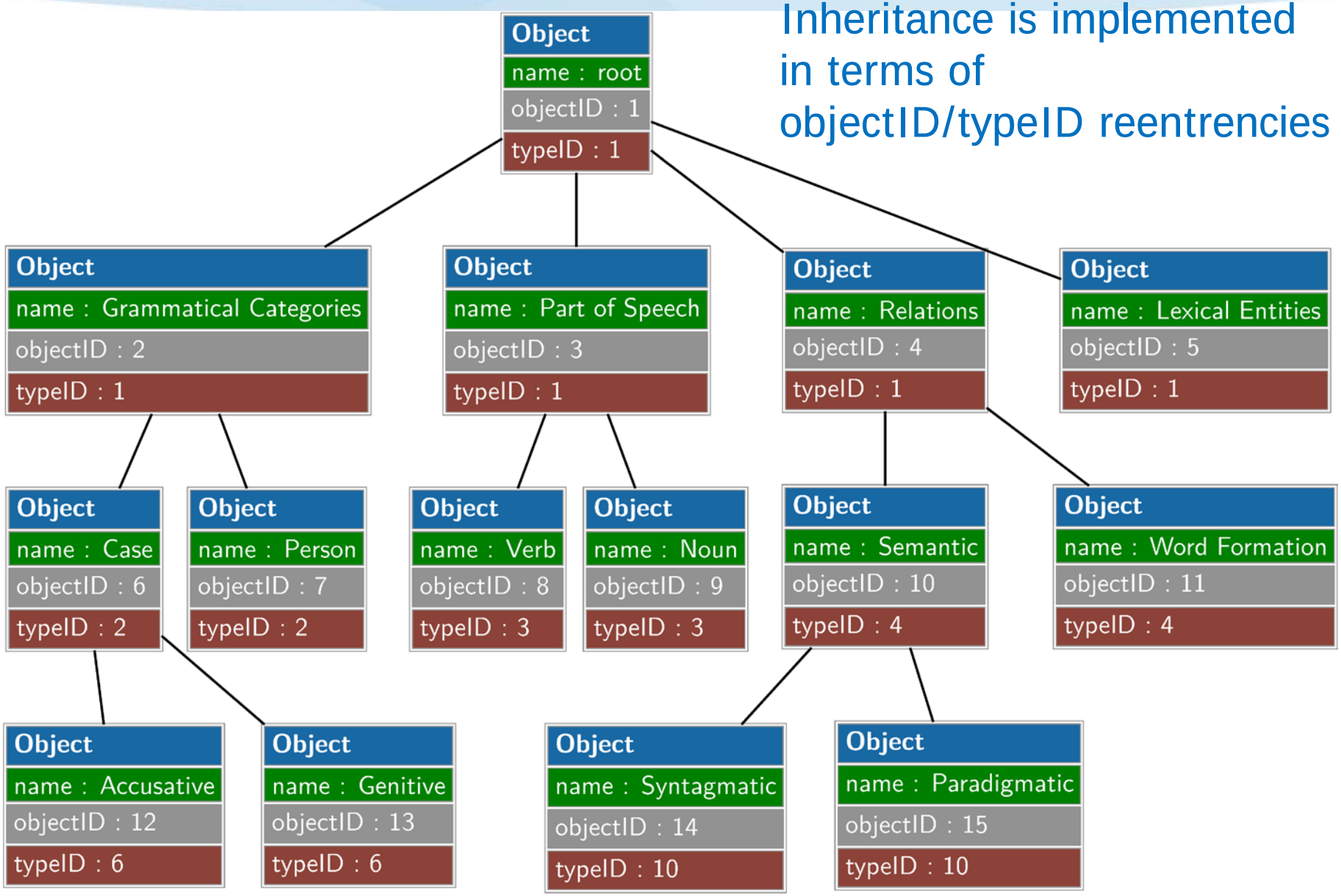
eLexicon Logical Data Model

Entity-Relationship Diagram

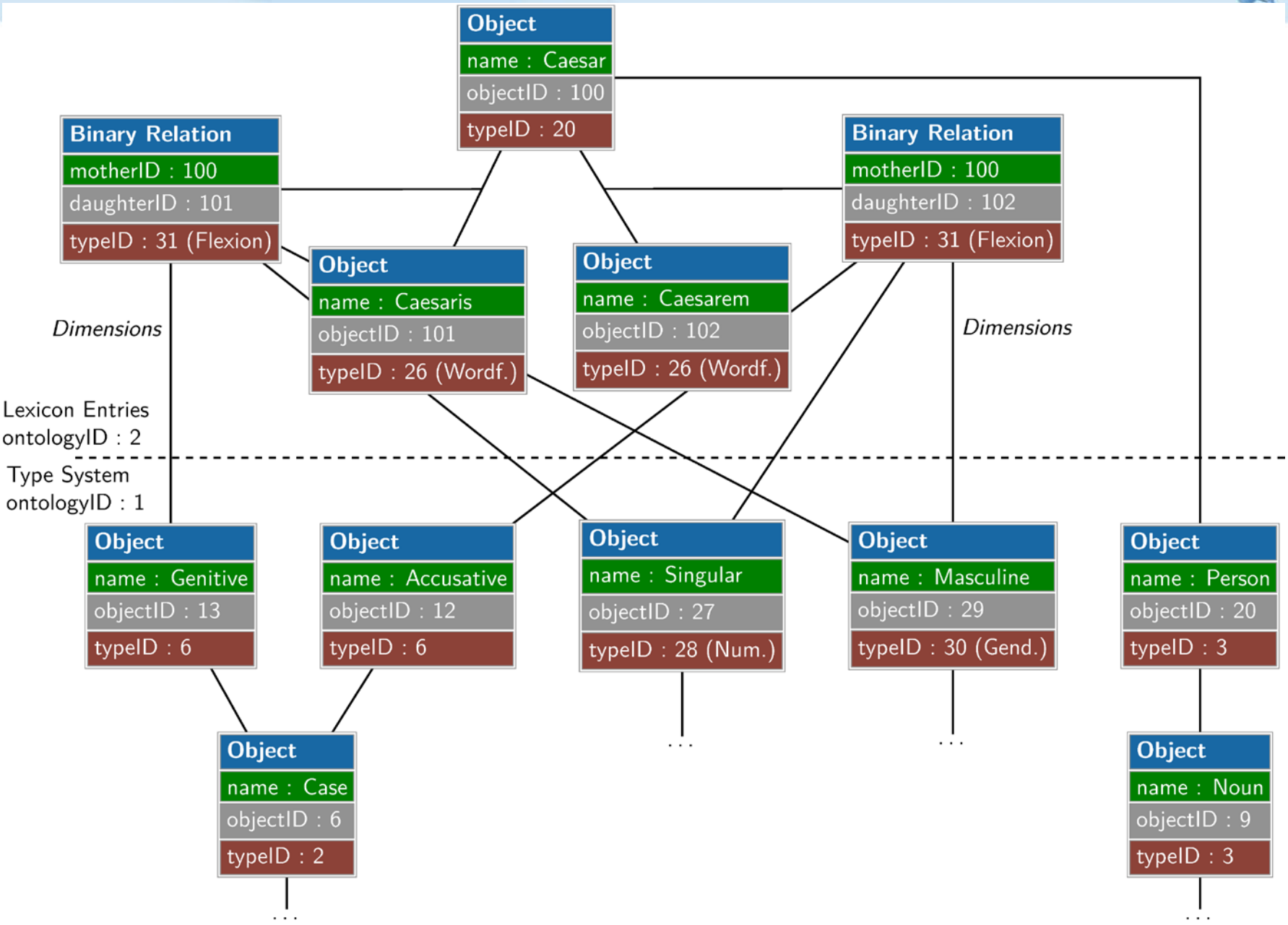


eLexiconType Hierarchy: Excerpt

Inheritance is implemented in terms of objectID/typeID reentrencies



eLexicon Data Model: Example Entry *Caesar*



eLexicon Size

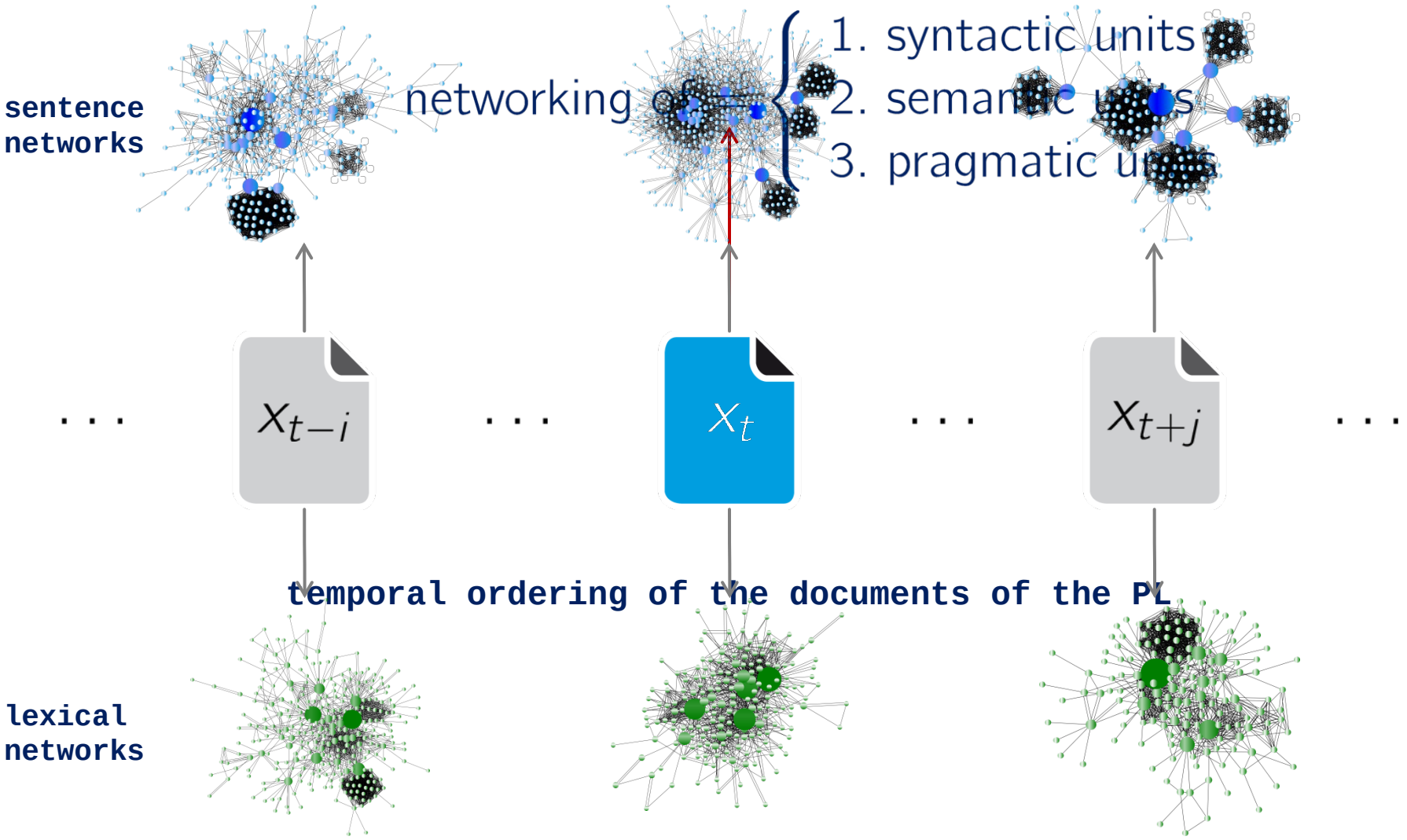
Stock-taking

POS	# lemma	# word form
ADJ	23 998	2 396 700
ADV	25 314	90 505
APPR (prep. left)	71	138
CARD	71	1 474
ITJ	140	140
KO (conjunction)	166	166
NN	33 649	442 623
NE	1 052	16 204
NEP	28 213	264 382
V	10 005	3 429 535
ORD	71	2 892
DIST	69	1 590
sums:	122 819	6 646 349

1. Corpus Example: Patrologia Latina
2. eLexicon
3. Time Series
4. Sample Analyses
5. Linguistic Networks Workflow

Network Induction

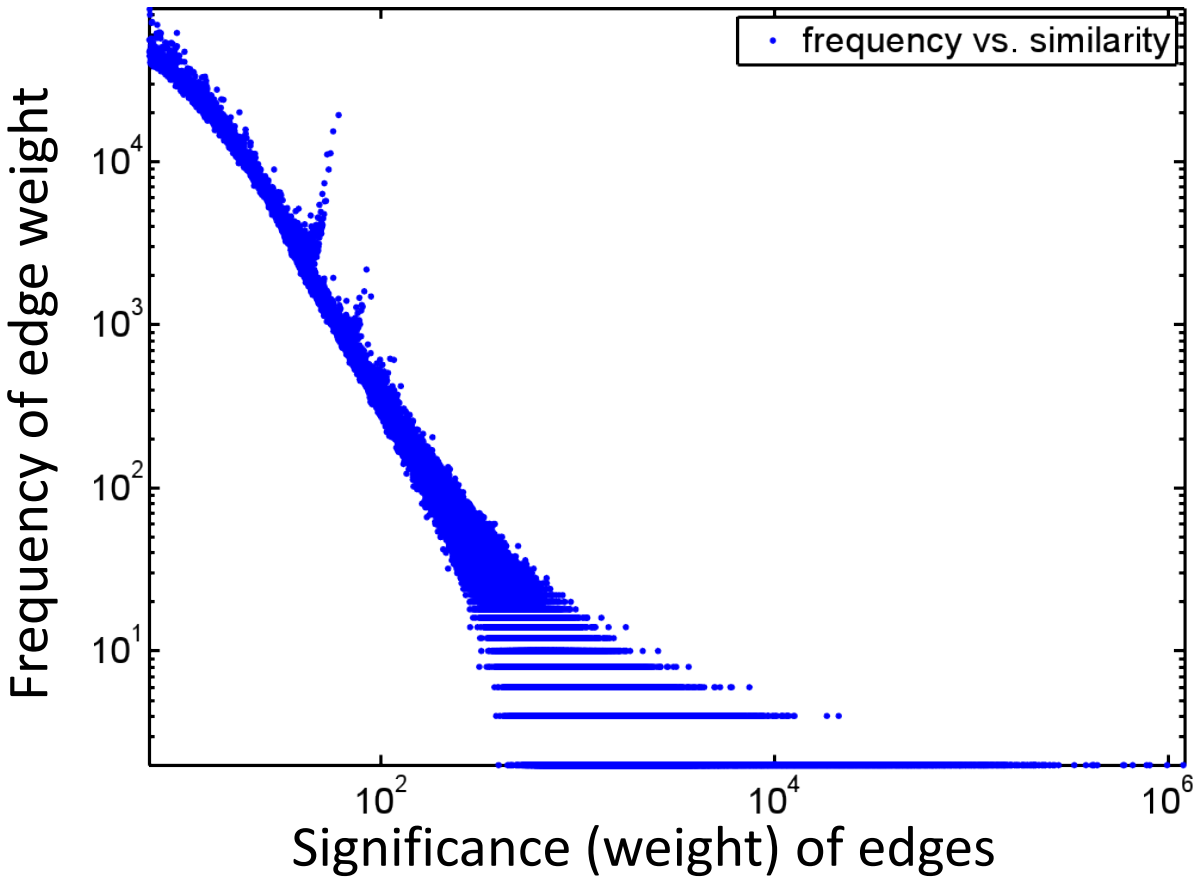
Multilevel Networks



Network Induction

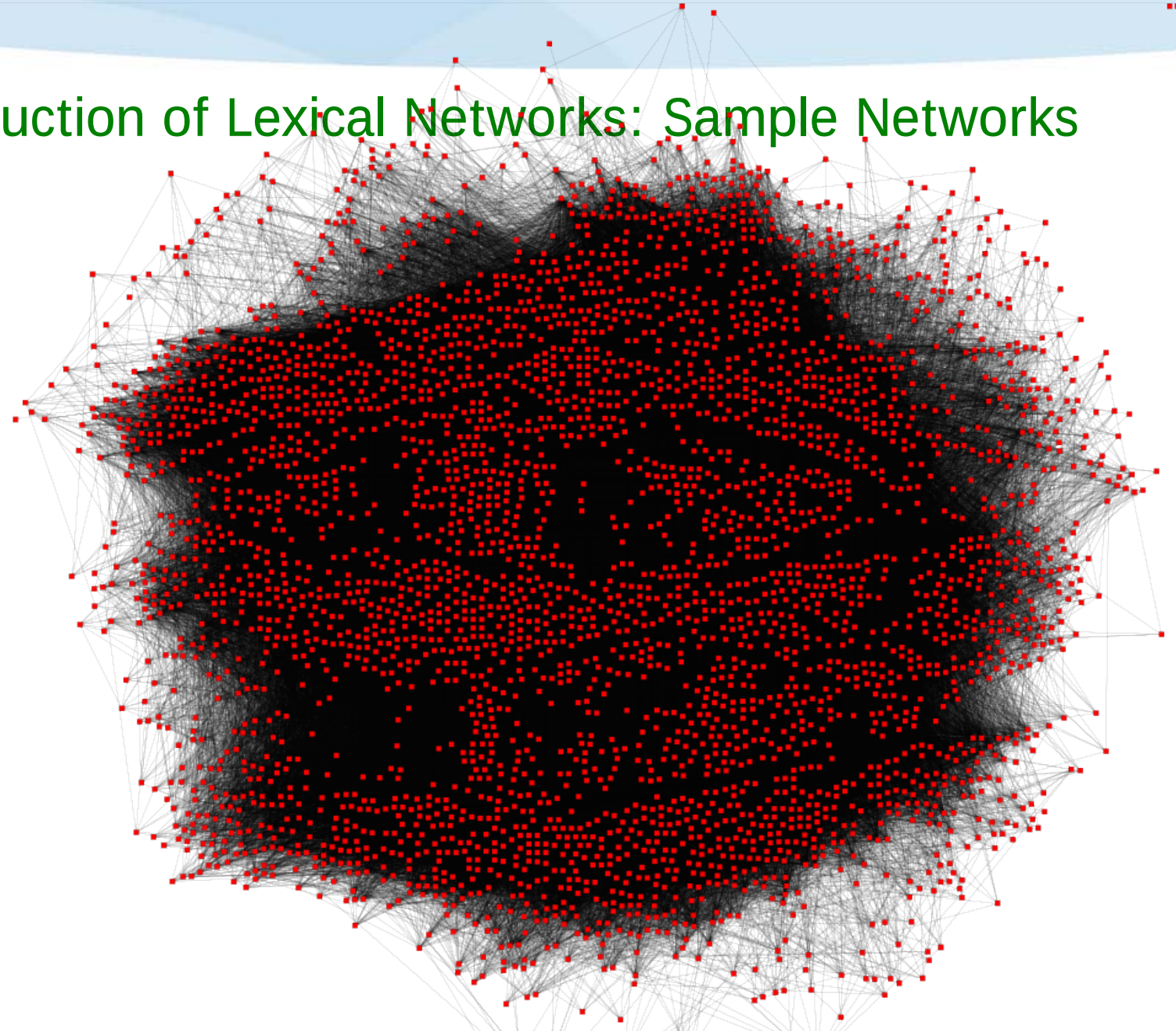
Induction of Lexical Networks (Heyer et al. 2006)

$$sig(w_i, w_j) = sig(A, B) = \frac{k(\ln k - \ln \lambda - 1)}{\ln n}$$



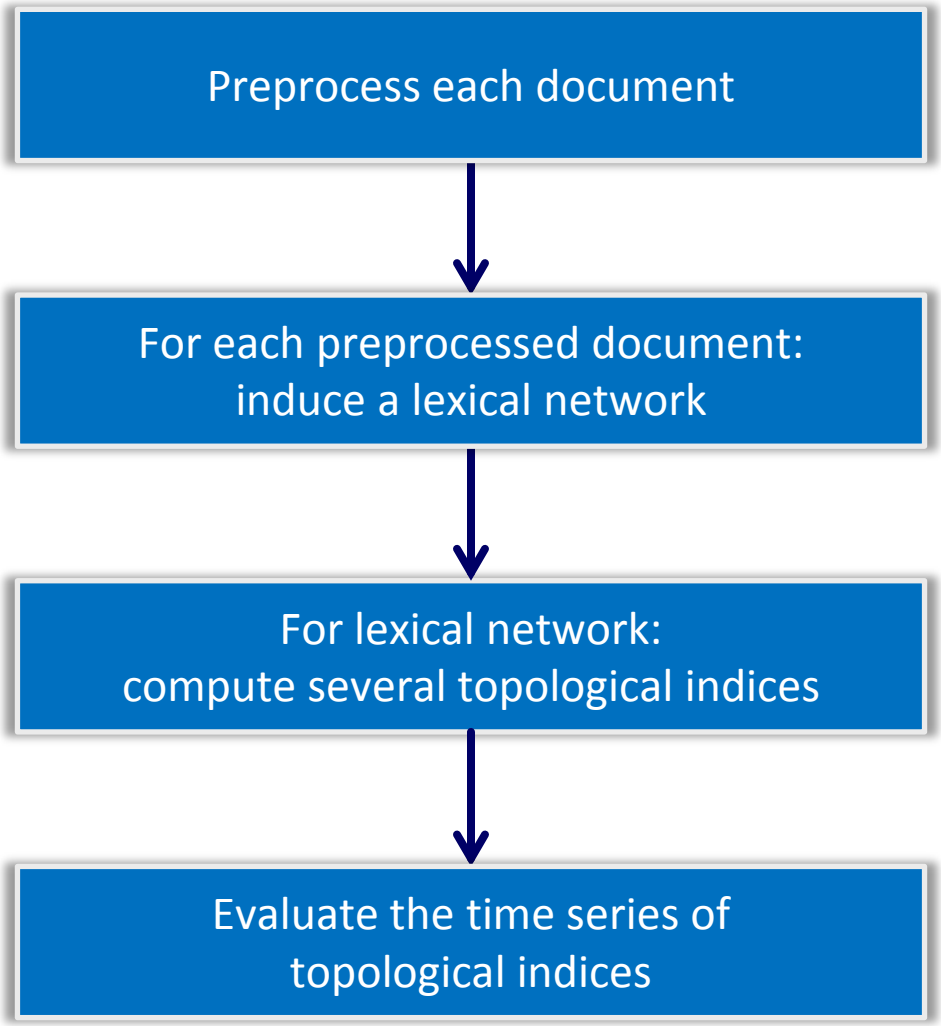
Network Induction

Induction of Lexical Networks: Sample Networks



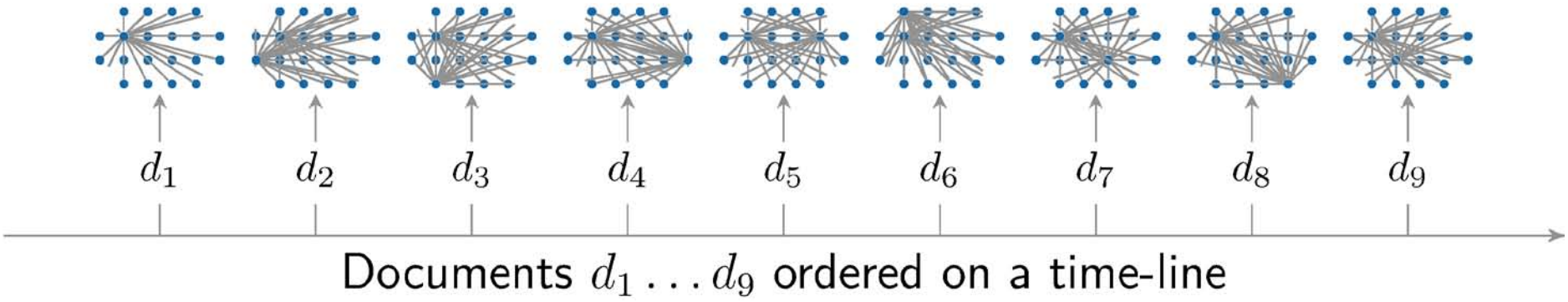
Time Series of Lexical Networks

Approach



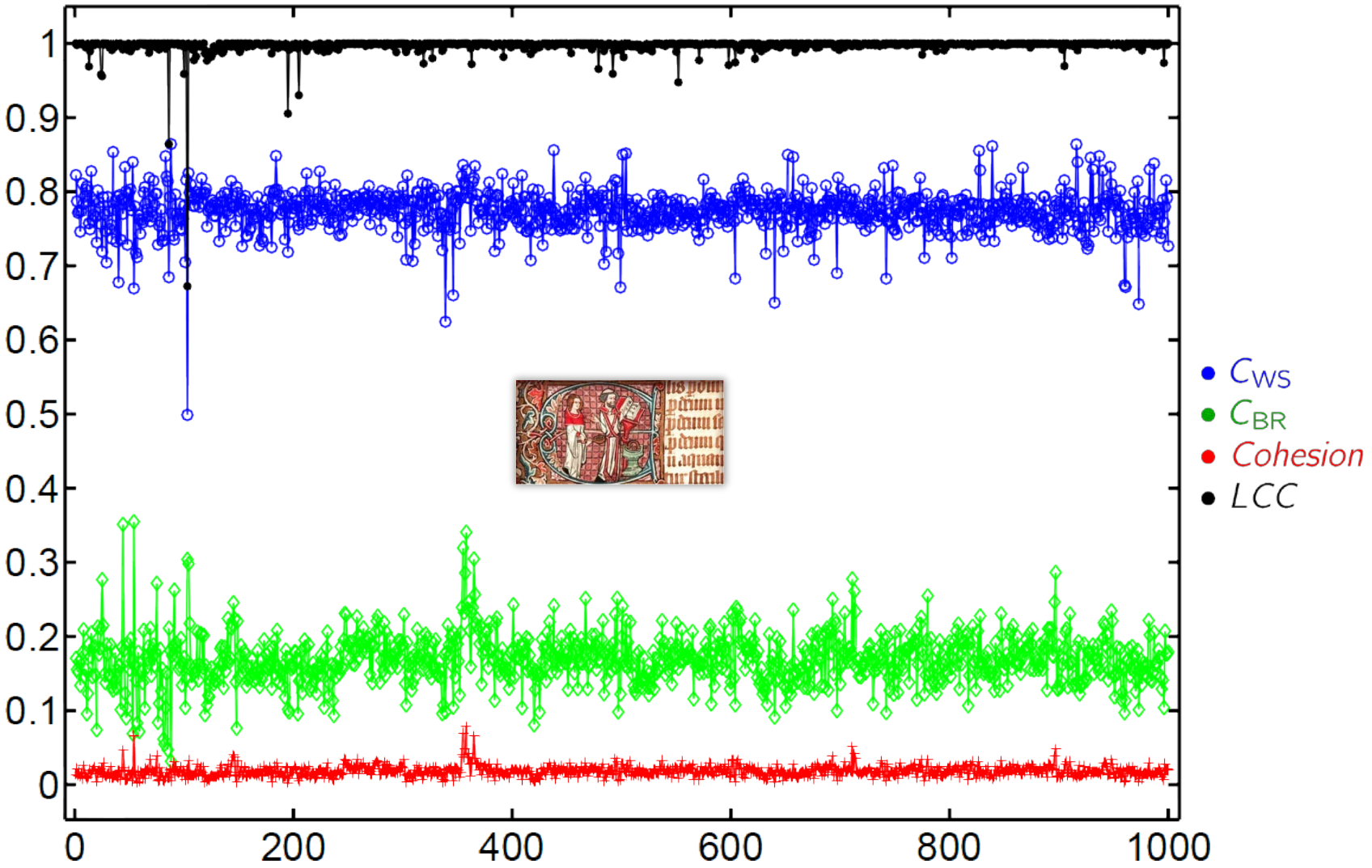
Time Series of Lexical Networks

- Illustration of a time series of lexical networks
  - Documents are ordered according to an underlying time line
  - For each document a lexical network is induced



Time Series of Lexical Networks

Real life example: Patrologia Latina

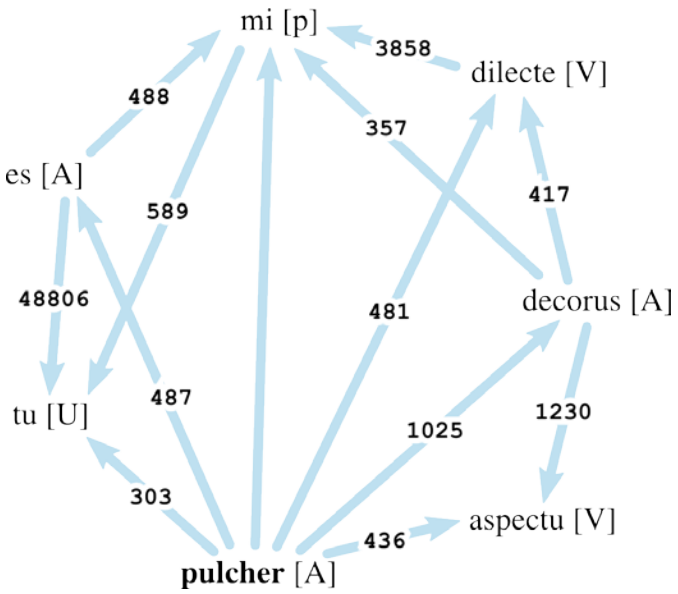


Agenda

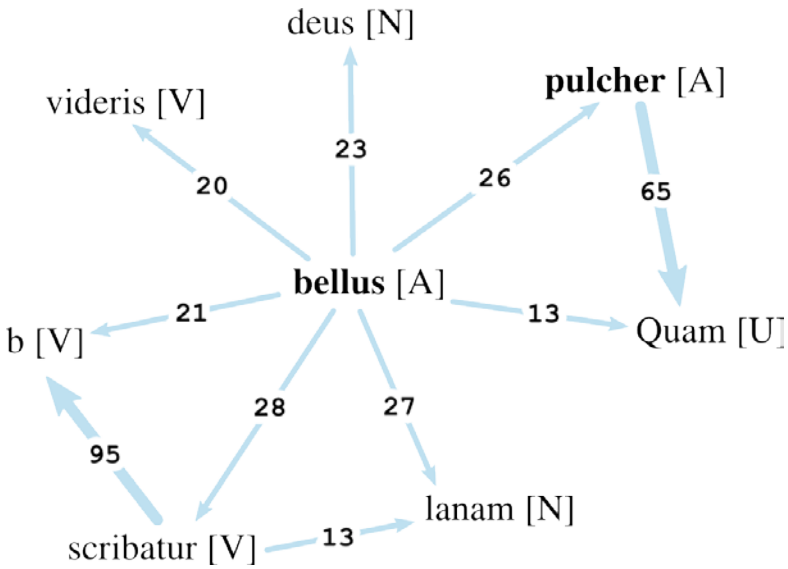
1. Corpus Example: Patrologia Latina
2. eLexicon
3. Time Series
4. Sample Analyses
5. Linguistic Networks Workflow

Language Change in a Network Perspective

Example: Word Usage in Time



pulcher
'pretty' or 'nice'
old form



bellus
'pretty' or 'nice'
new form

Example of Word Usage

(…)

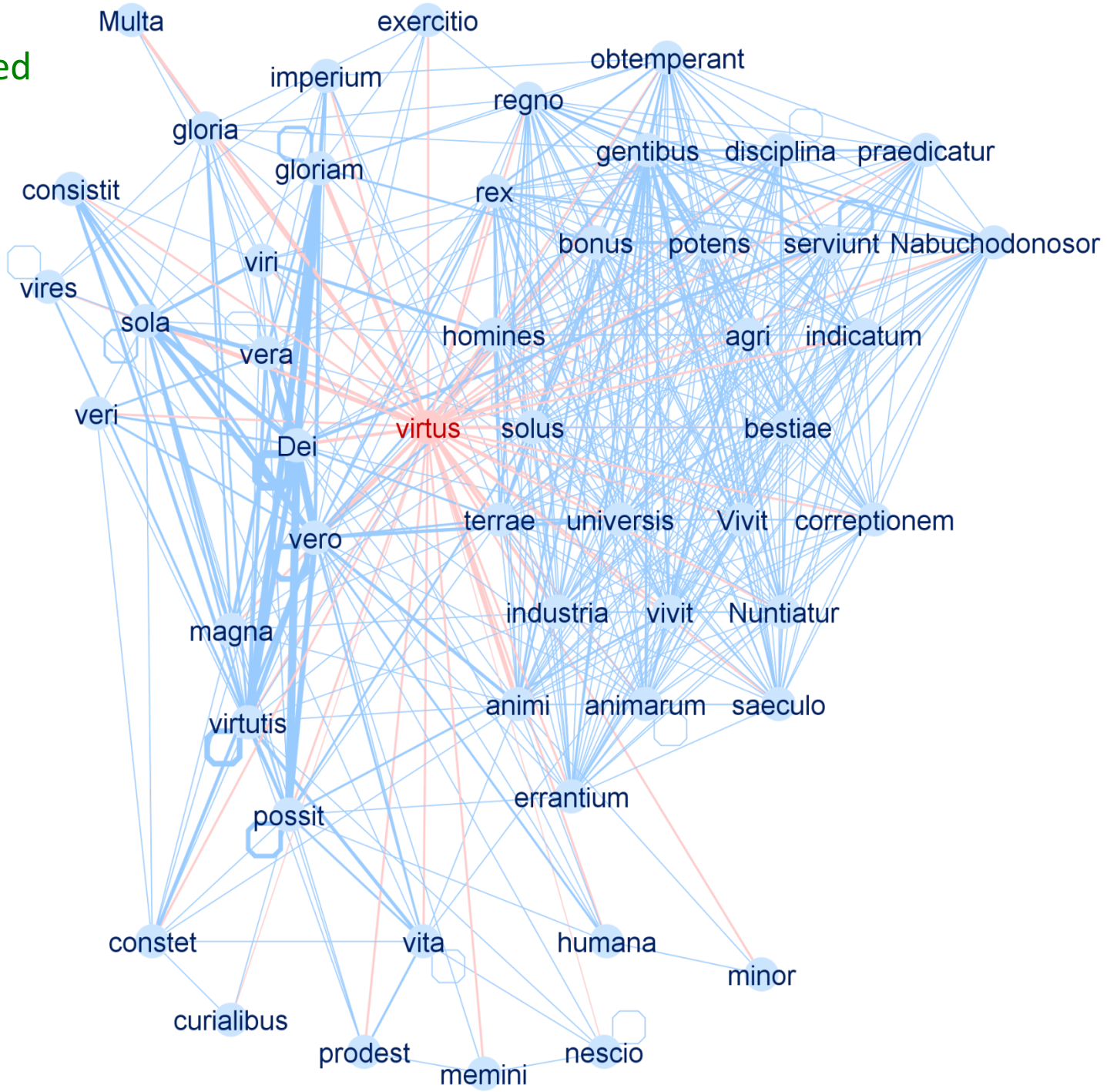
Vellus, si lanam significat, per v;
si **bellus**, id est, **pulcher**, per b scribatur.

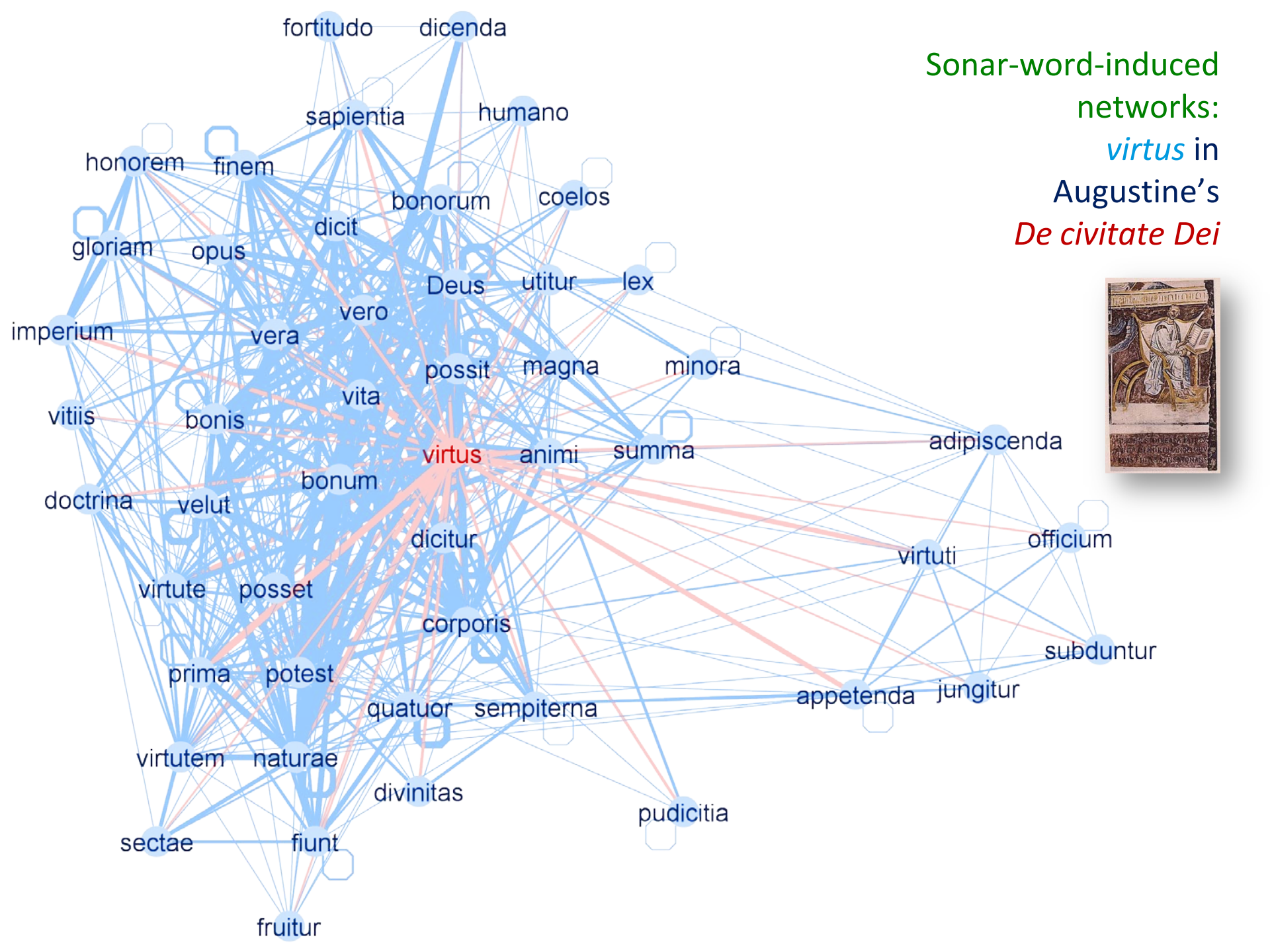
(…)

Alcuinus: De Orthographia, Vol. 101, ~ 735 - 804

Vellus, if it means *lana (wool)*, is written with a *v*; if it's *bellus*, in the meaning of *pulcher*, it should be written with a *b*.

Sonar-word-induced networks:
virtus in
John of Salisbury's
Polycraticus

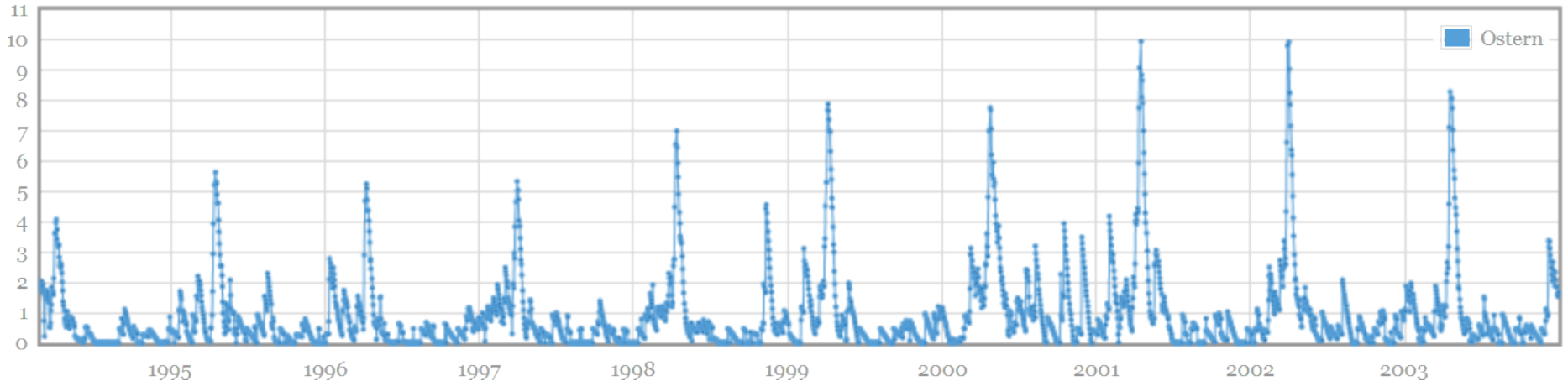


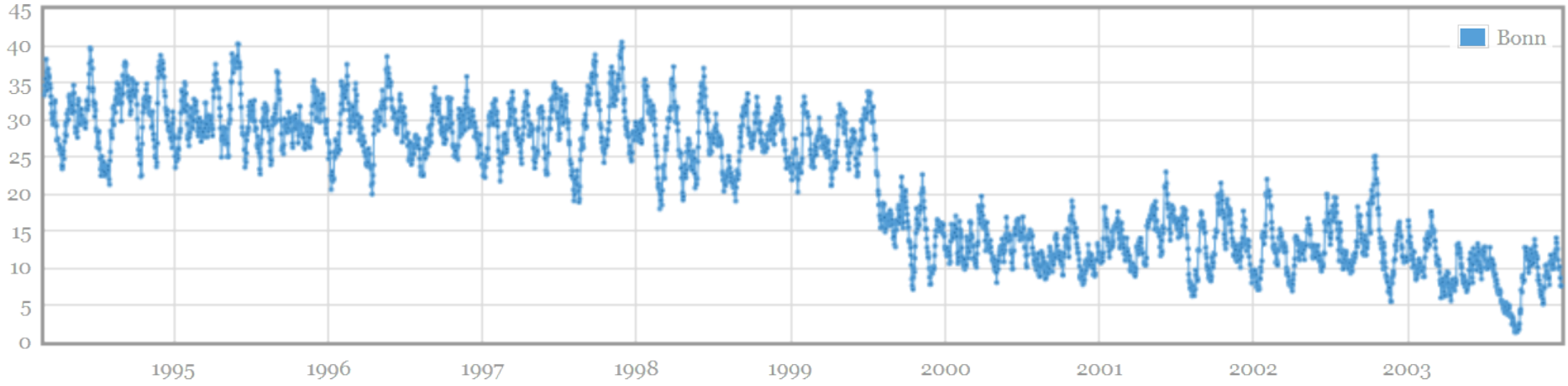

fortitudo
dicenda
sapientia
humano
honorem
finem
bonorum
coelos
dicit
gloriam
opus
Deus
utitur
lex
vero
imperium
vera
possit
magna
minora
vita
vitiis
bonis
virtus
animi
summa
adipiscenda
bonum
doctrina
velut
dicitur
officium
virtuti
virtute
posset
corporis
subduntur
prima
potest
appetenda
jungitur
quatuor
sempiterna
virtutem
naturae
divinitas
pudicitia
sectae
fiunt
fruitur
Sonar-word-induced networks:
virtus in Augustine's
De civitate Dei

Time Series Analyses

Temporal Variability of Word Meanings

(Laußmann 2010)





Agenda

1. Corpus Example: Patrologia Latina
2. eLexicon
3. Time Series
4. Sample Analyses
5. Linguistic Networks Workflow

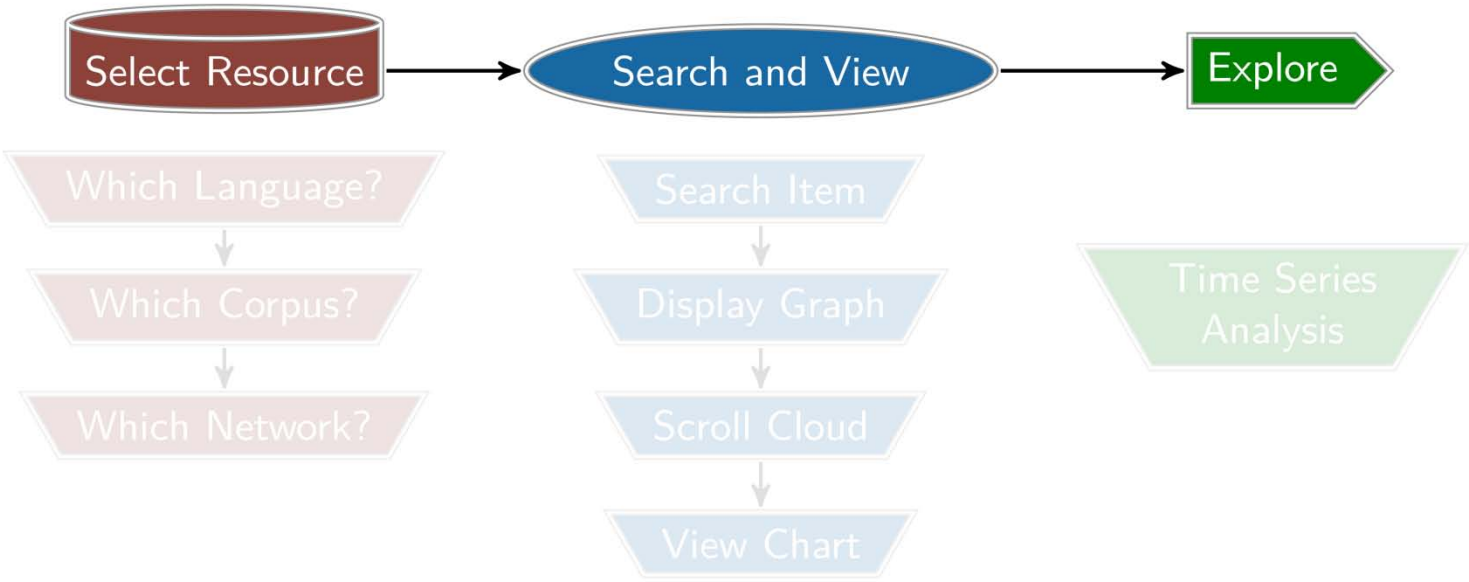
www.linguistic-networks.net

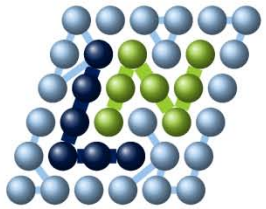
Coverage

Stock-taking

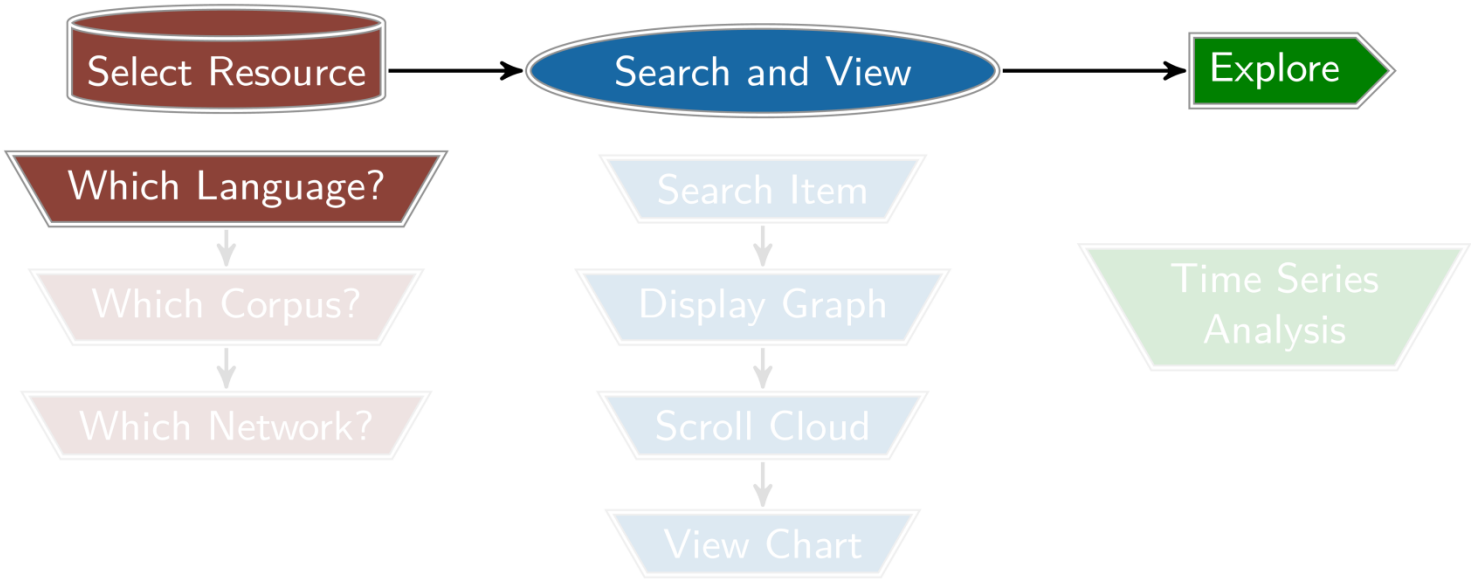
Corpus	Language	Network	# Nodes	# Edges
Althochdeutsch	German	lemma network	5 787	103 402
Althochdeutsch	German	word form network	4 336	82 652
AmeisenWiki	German	lemma network	29 046	1 299 079
AmeisenWiki	German	word form network	36 083	1 614 081
Avestan	Avestan	lemma network	5 405	157 465
Avestan	Avestan	word form network	1 868	75 398
CoNLL Shared Task English Treebank	English	lemma network	33 373	278 260
Enron Email Dataset	English	lemma network	743 459	1 154 792
Enron Email Dataset	English	word form network	512 048	1 352 563
Franz Kafka: Bericht	German	lemma network	1 097	24 655
Franz Kafka: Bericht	German	word form network	1 299	27 622
Franz Kafka: Erzählungen	German	lemma network	9 463	520 064
Franz Kafka: Erzählungen	German	word form network	11 796	629 996
Franz Kafka: Strafkolonie	German	lemma network	2 097	63 901
Franz Kafka: Strafkolonie	German	word form network	2 499	76 328
Friedrich Nietzsche: GdM (Exc.)	German	lemma network	3 186	178 841
Friedrich Nietzsche: GdM (Exc.)	German	word form network	3 678	201 091
Heinrich Rauchberg: Zählmaschine [...]	German	lemma network	2 134	80 755
Heinrich Rauchberg: Zählmaschine [...]	German	word form network	2 500	95 146
Hofmannsthals Werk	German	lemma network	35 048	53 226
Hofmannsthals Werk	German	word form network	49 813	74 614
Patrologia Latina	Latin	word form network	967 554	42 112 842
Patrologia Latina (tagged new)	Latin	lemma network	512 574	9 072 484
Patrologia Latina (tagged new)	Latin	word form network	886 794	12 712 662
SZ 1994	German	lemma network	552 163	5 185 775
SZ 1994	German	word form network	761 853	6 539 301
Test	Latin	lemma network	2 097	774
Test	Latin	word form network	2 499	775
Wikipedia	German	lemma network	7 933 813	24 819 804
Wikipedia	German	word form network	8 465 244	29 665 388
		sum word form:	11 922 715	94 879 333
		sum lemma:	9 657 891	43 374 403
		total:	21 580 606	138 253 736

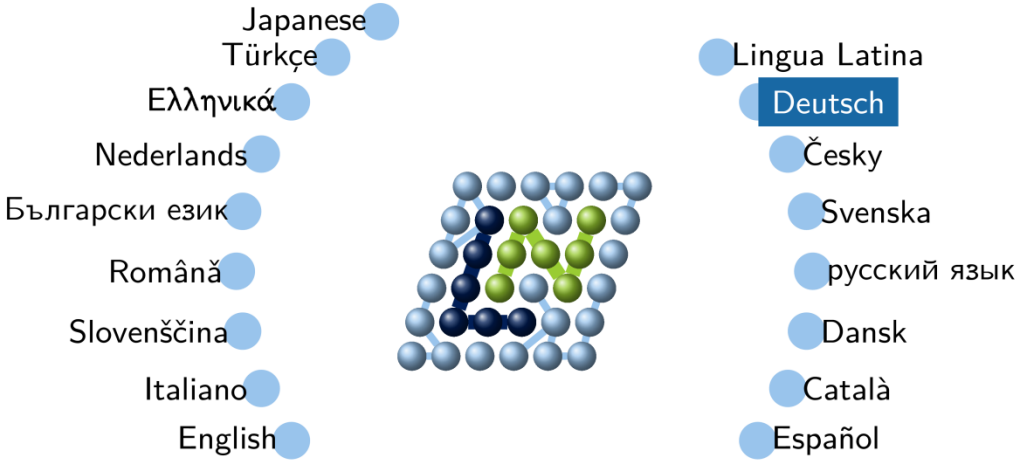
Linguistic Networks: Workflow



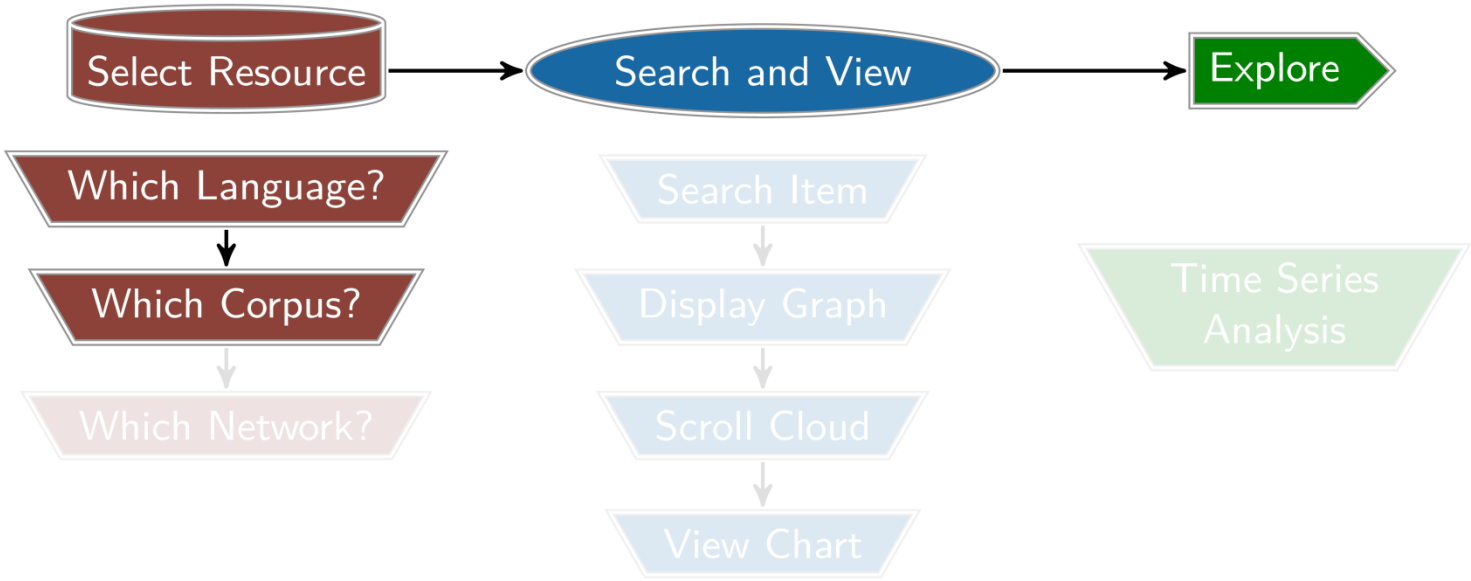


Linguistic Networks: Workflow





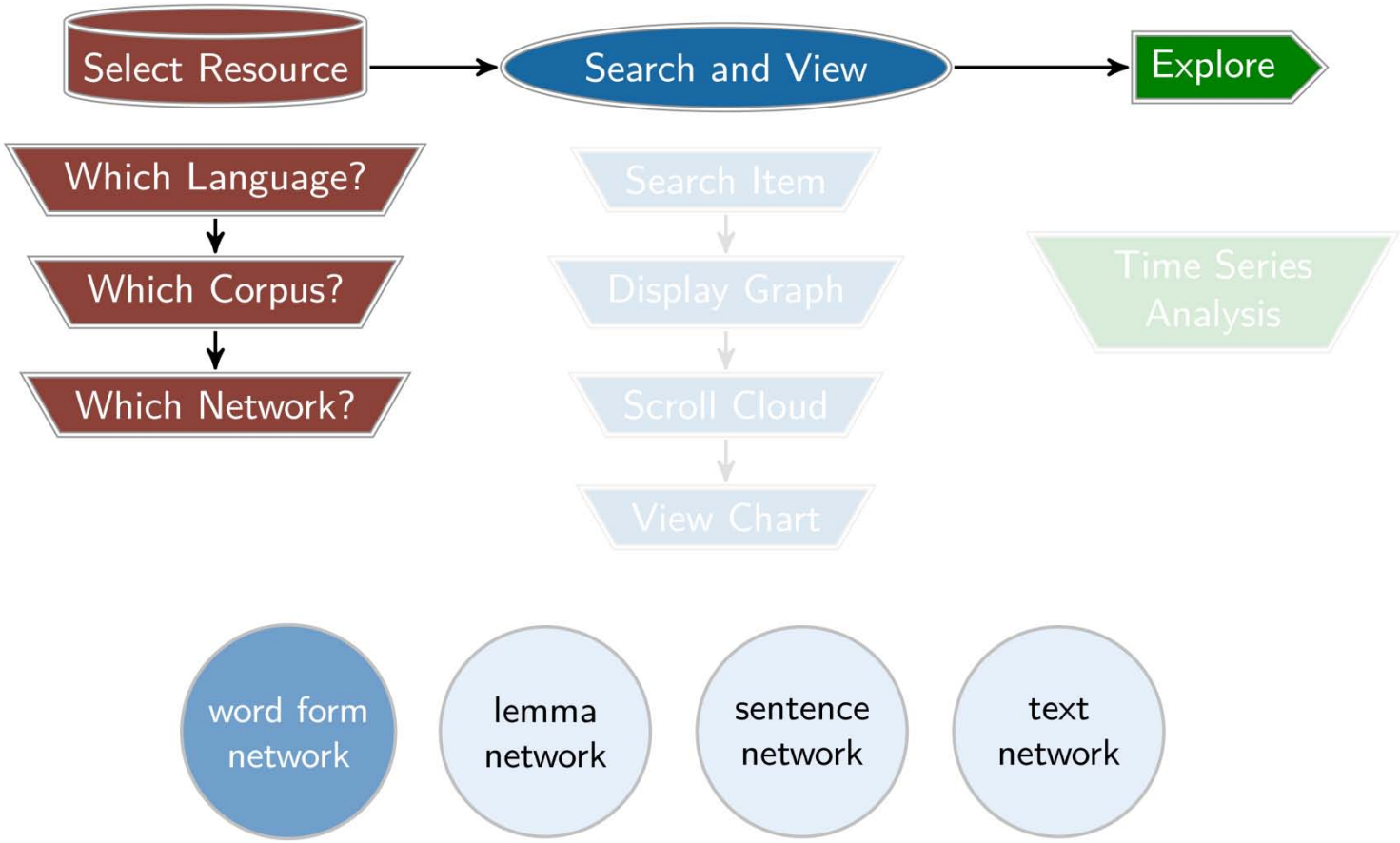
Linguistic Networks: Workflow



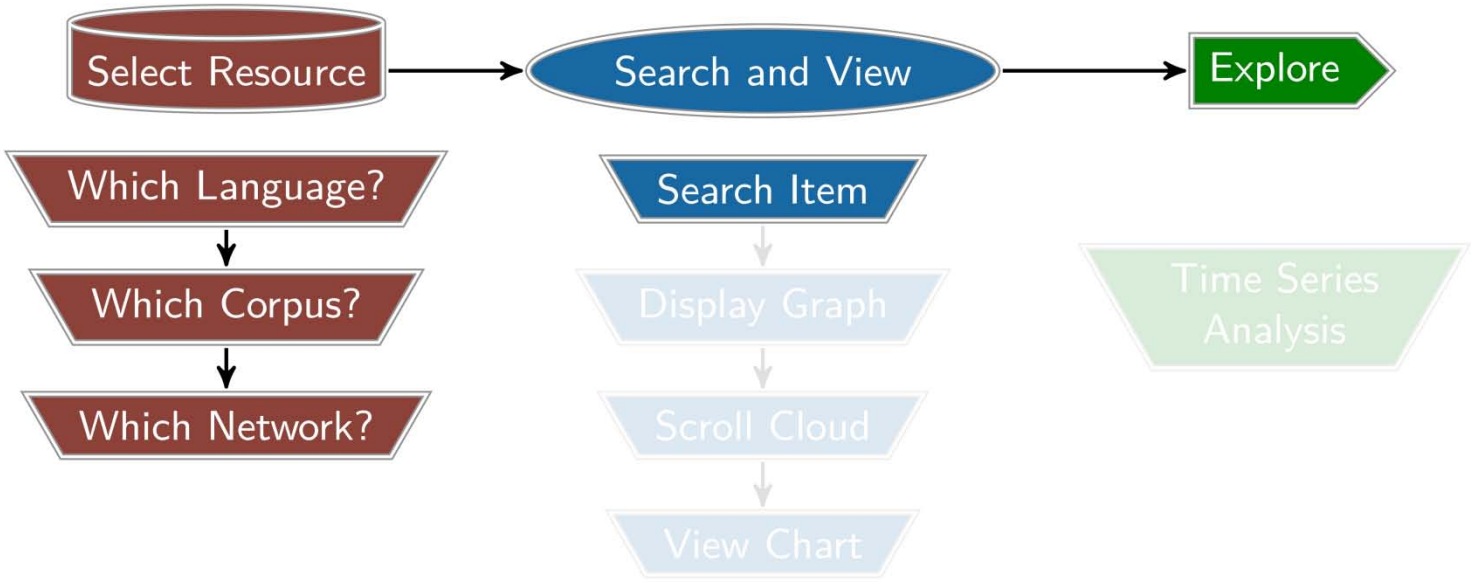
Choose corpus

SZ 1994
Wikipedia
Hofmannsthals Werk
Althochdeutsch
Franz Kafka: Erzählungen
Friedrich Nietzsche: Genealogie der Moral (Ausschnitt)
GuttenPlag Wiki
MemoryAlpha Wiki
VroniPlag Wiki

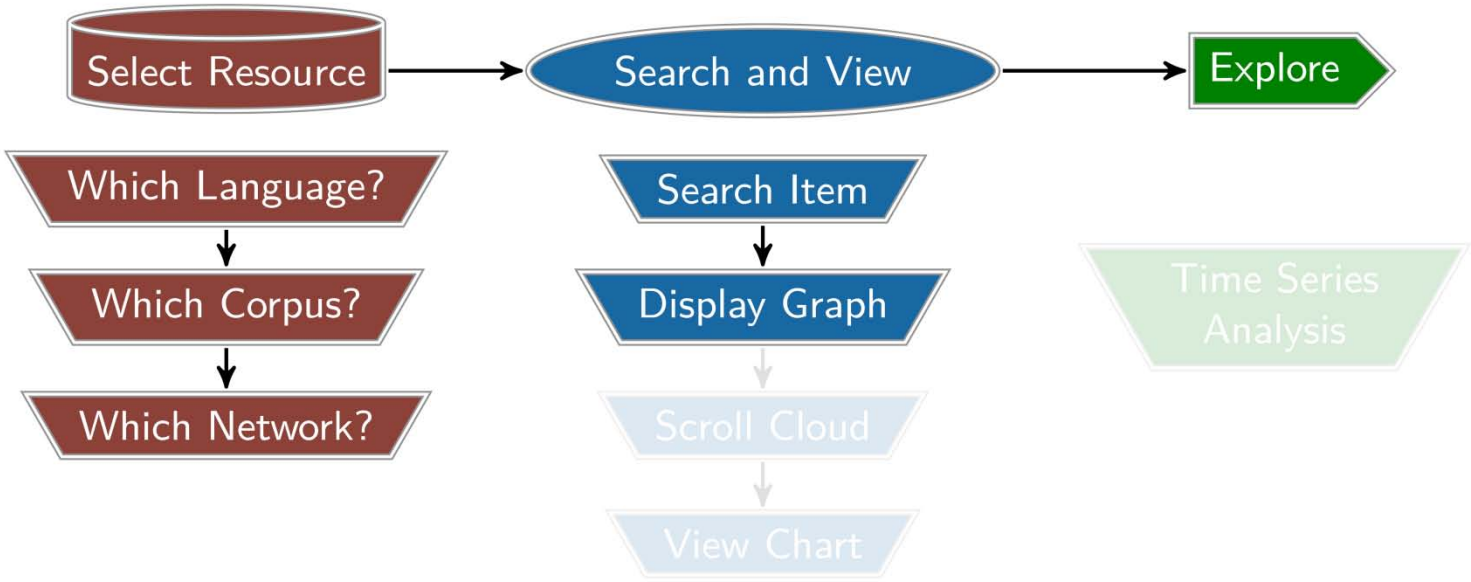
Linguistic Networks: Workflow

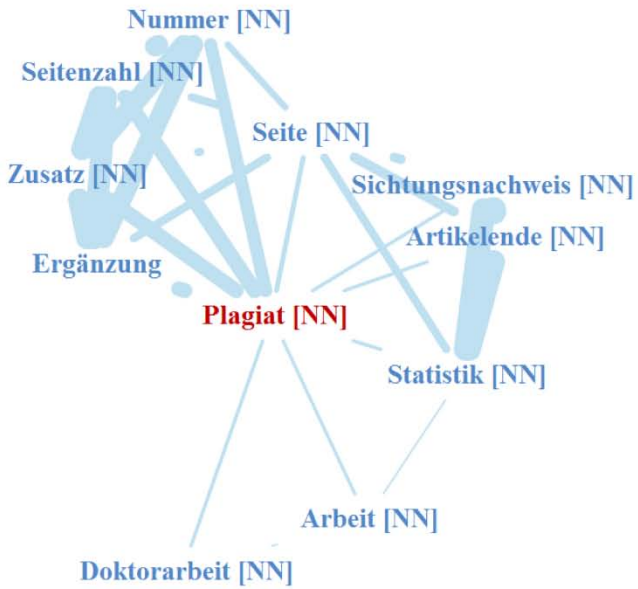


Linguistic Networks: Workflow

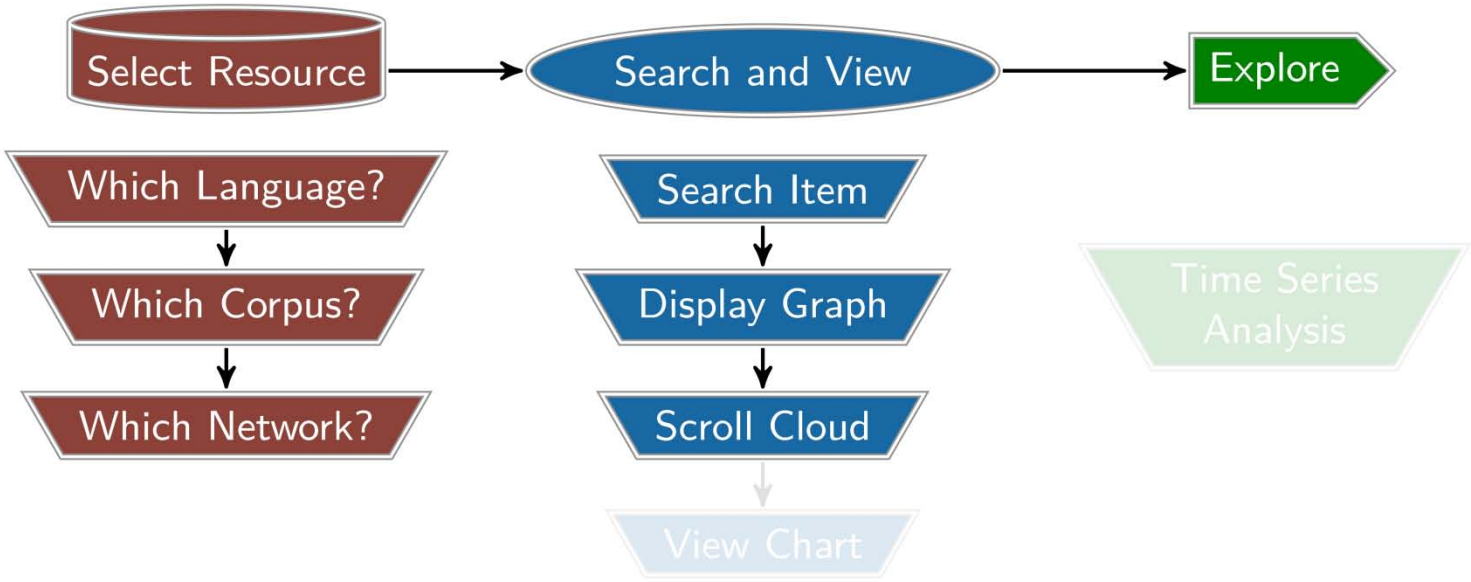


Linguistic Networks: Workflow

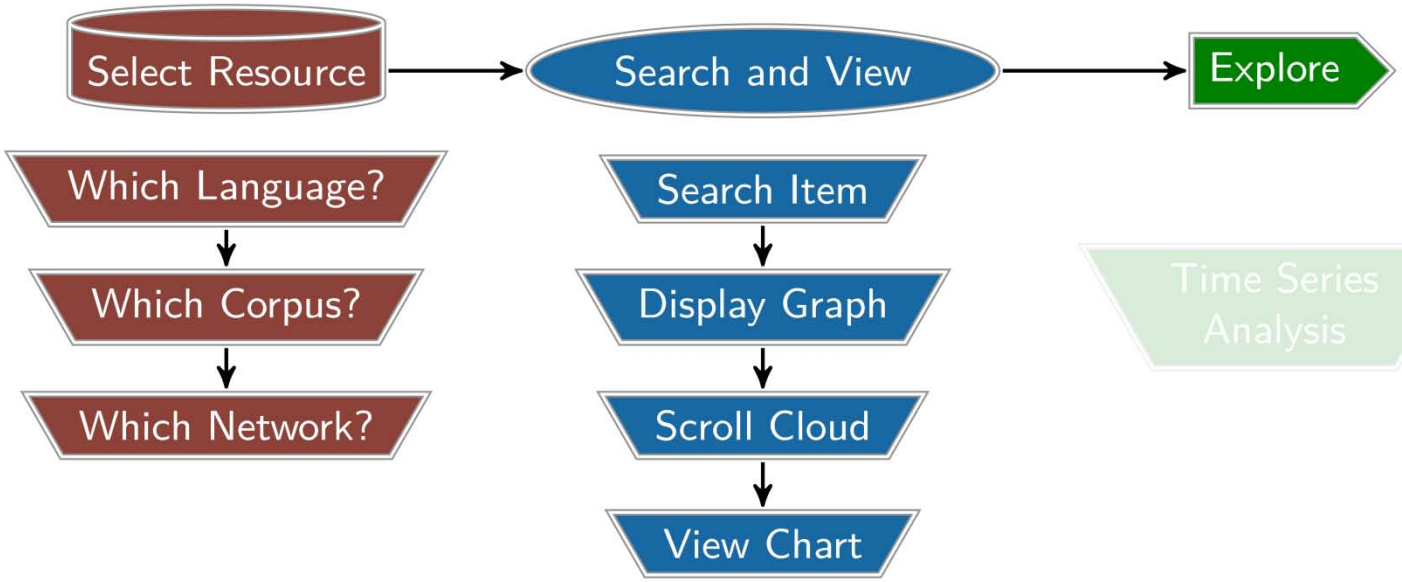


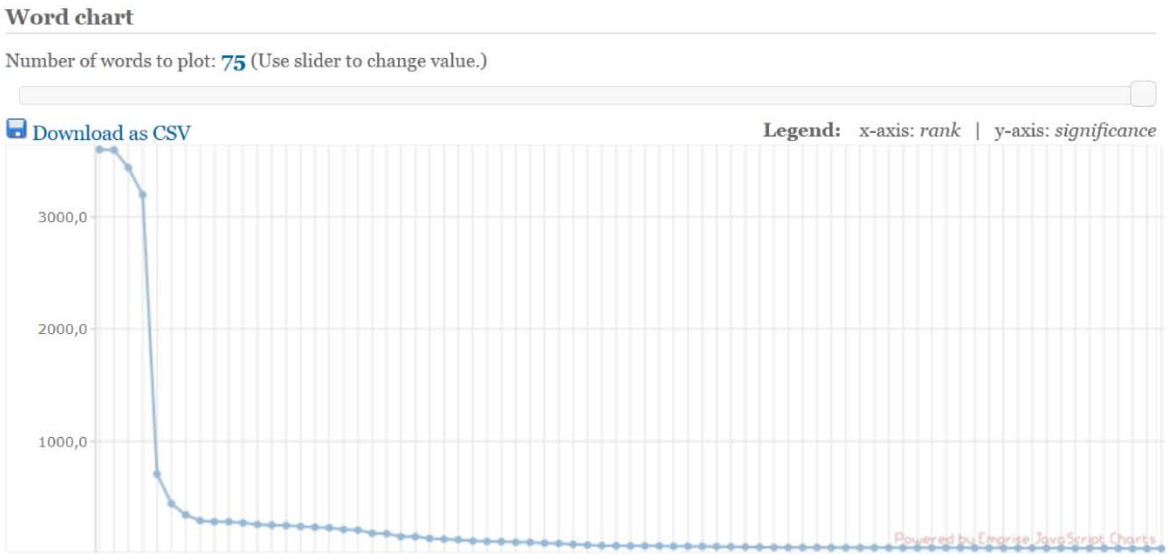


Linguistic Networks: Workflow

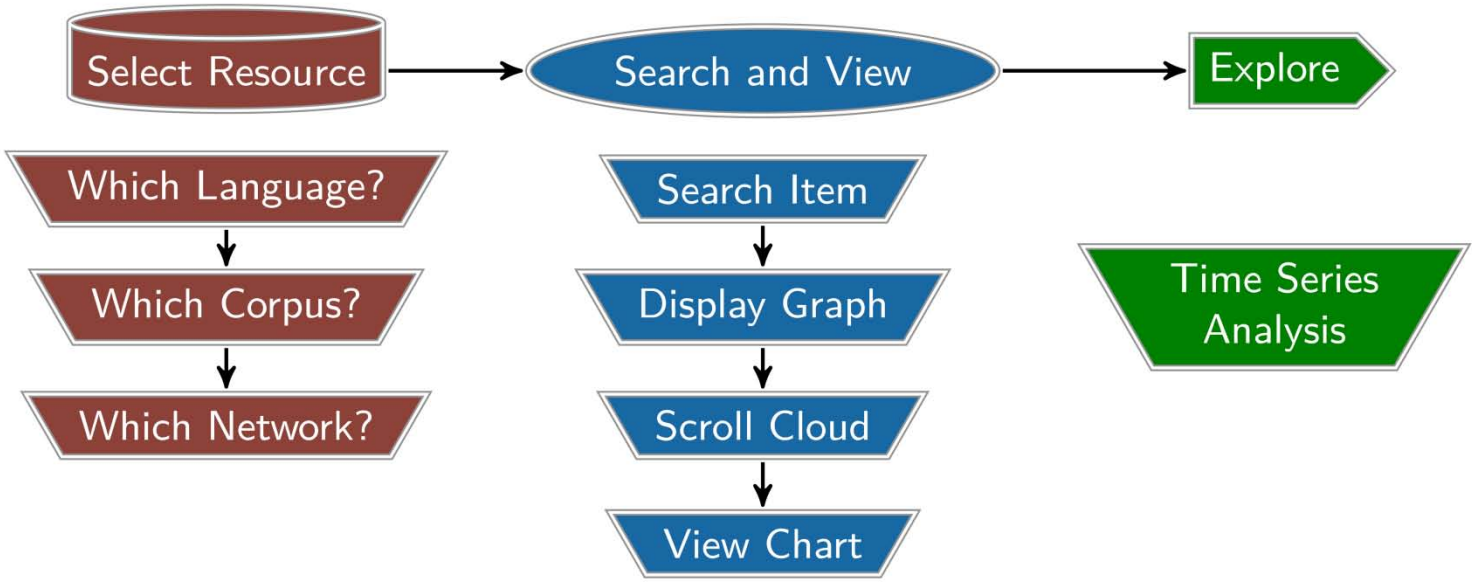


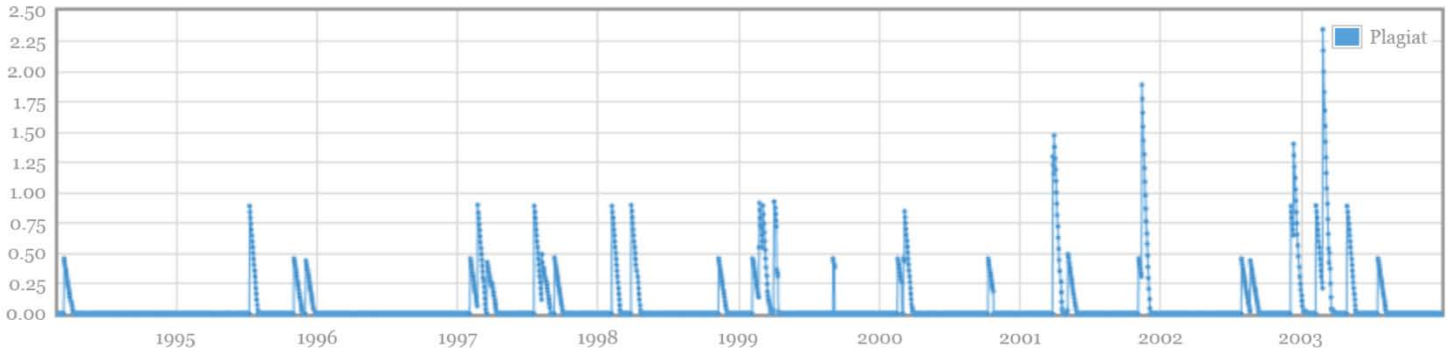
Linguistic Networks: Workflow





Linguistic Networks: Workflow





www.linguistic-networks.net

LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz

SPONSORED BY THE

Federal Ministry of Education and Research