

Linguistic Networks



Alexander Mehler, Rüdiger Gleim, Alexandra Ernst, Andy Lücking

Networking Words



First hits of a Google picture search for German "Bank":

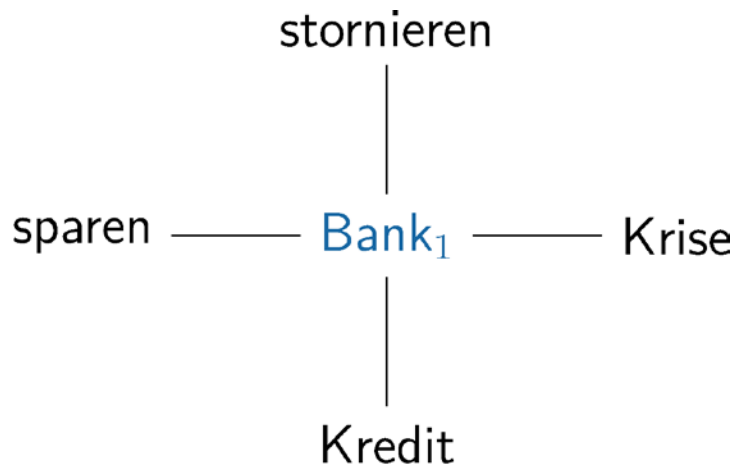


Networking Words

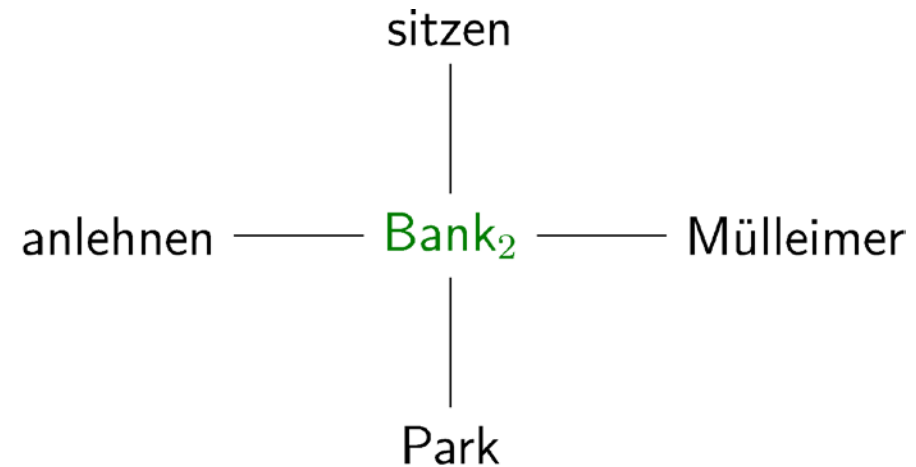


Neighbors of "Bank":

User Jleon, CC-BY-SA



Florian Hurlbrink, CC-BY-SA





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



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)





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





Beyond Conversion: Preprocessing Steps

Wirtschaft

Rubriken

Blättern



Klage gegen Stilllegung

RWE droht mit Wiederanfahren von Biblis

Nach der Klage gegen die vorläufige Abschaltung von Biblis A, droht RWE nun damit, das Atomkraftwerk wieder hochzufahren. Solange dies nicht verboten werde, gehe man davon aus, dass eine Gefährdung nach dem Atomgesetz nicht bestehe. >



The screenshot displays a software interface with two main panels. The left panel, titled 'XML Tree for 905790', shows a hierarchical tree structure with folders for 'teiHeader', 'text', 'body', and 'p'. The right panel, titled 'XML View for 905790', shows the corresponding XML code. The XML code includes various tags for words and segments, such as <w>01</w>, <w>April</w>, <w>2011</w>, <w>2011</w>, <w>04</w>, <w>16</w>, <w>43</w>, <w>00</w>, and <w>Der</w>. The XML code also includes attributes like 'org="uniform"', 'sample="complete"', 'part="N"', 'type="#CARD"', 'lemma="April"', 'lemma="2011"', 'lemma="16"', 'lemma="43"', 'lemma="00"', and 'lemma="der"'. The interface also shows tabs for 'Text Characteristics for 905790' and 'Frequency Spectrum for 905790'.



1. Corpus Example: Patrologia Latina

2. eLexicon

3. Time Series

4. Sample Analyses

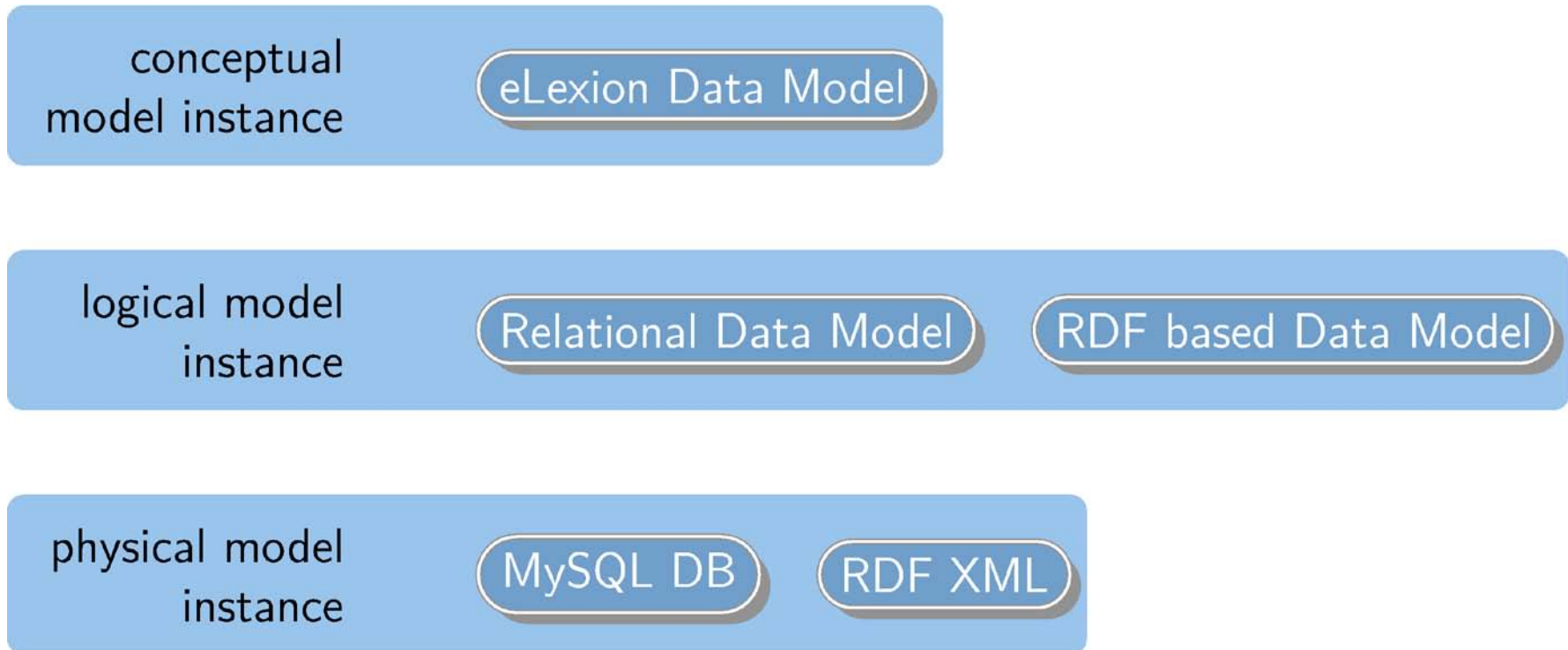
5. Linguistic Networks Workflow

eLexicon Data Model Architecture



Three layers

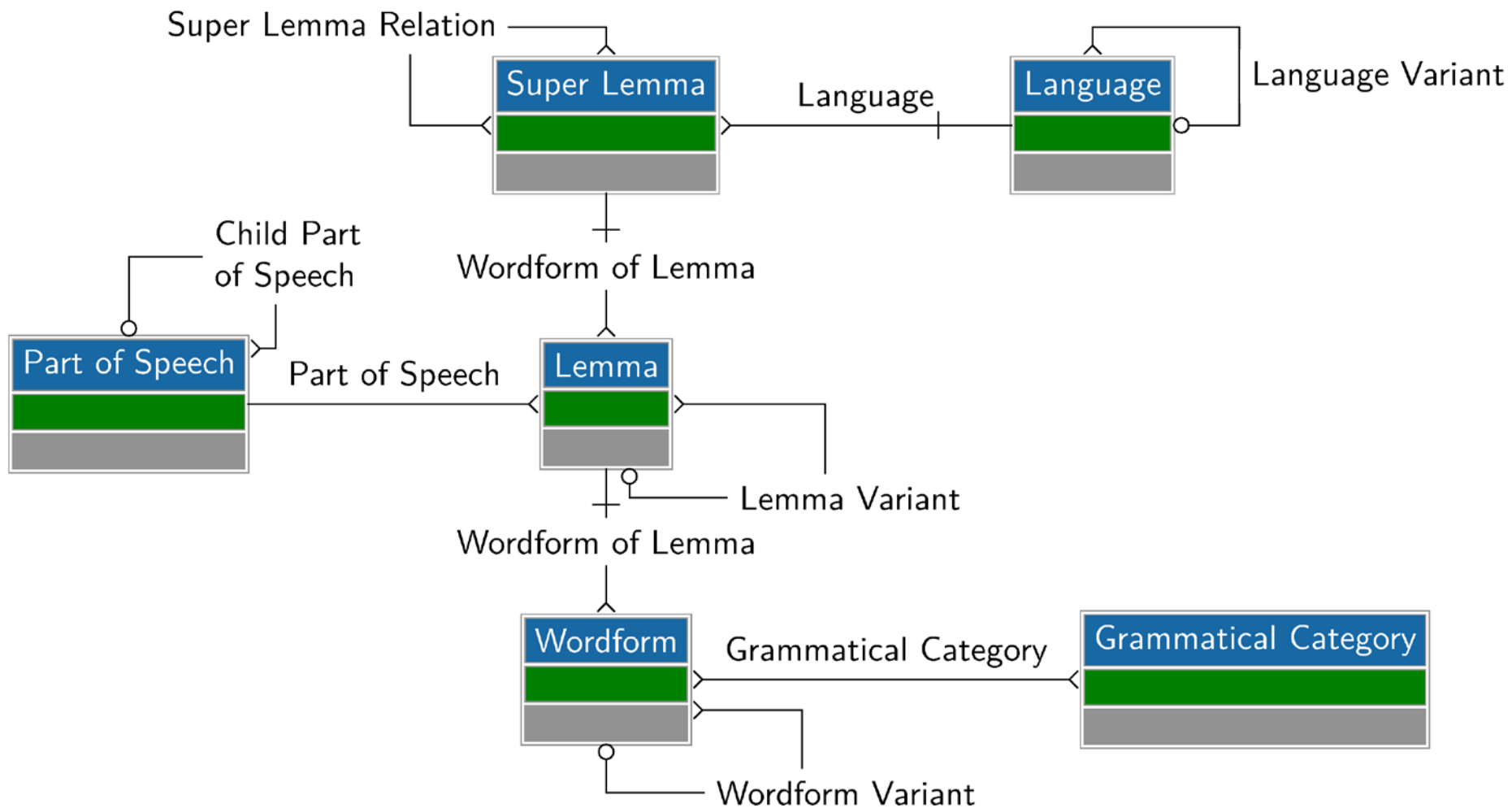
ANSI/X3/SPARC Study Group on Data Base Management Systems (ANSI, 1975)



eLexicon Conceptual Data Model



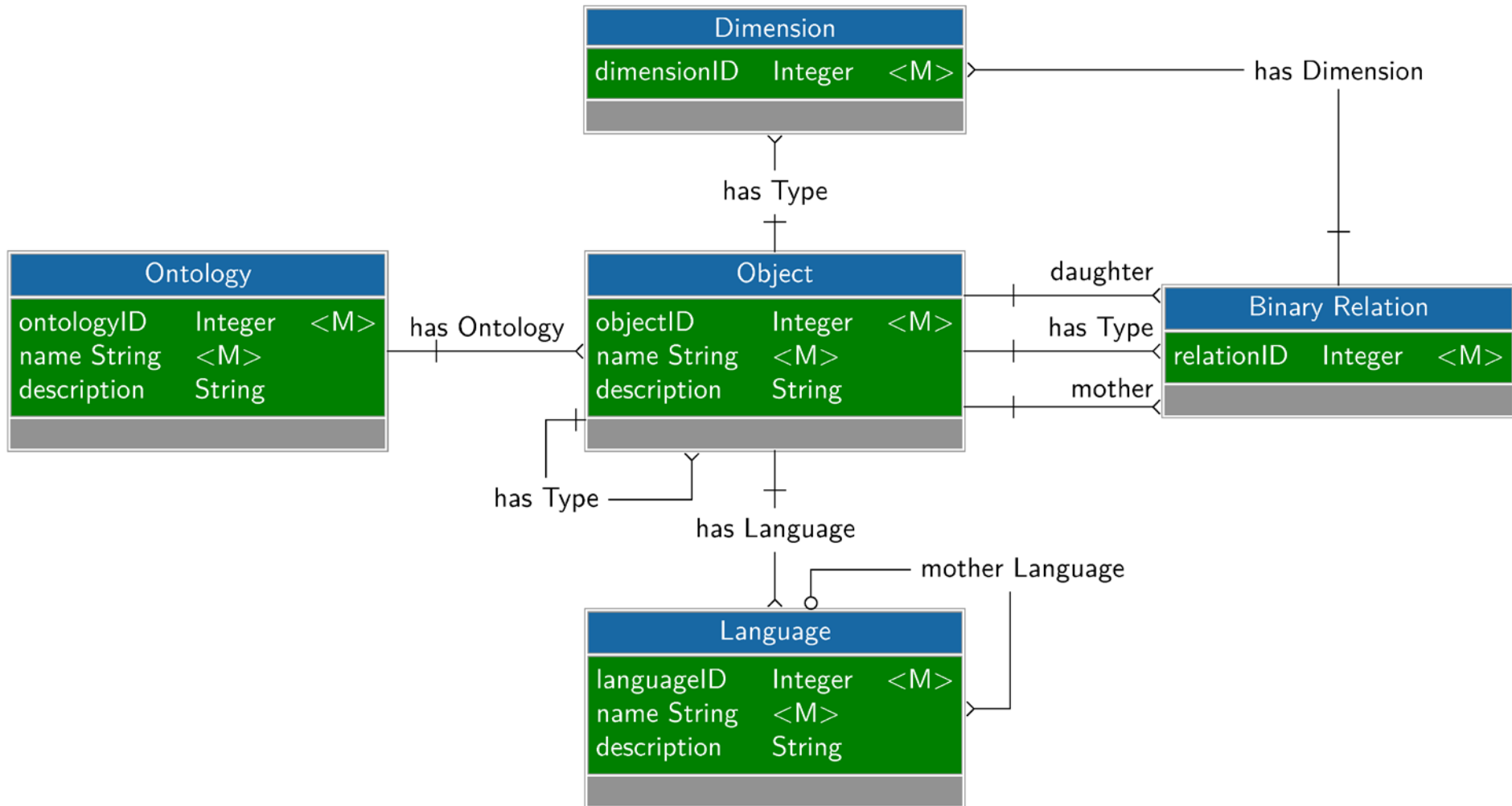
Simplified Entity-Relationship Diagram



eLexicon Logical Data Model



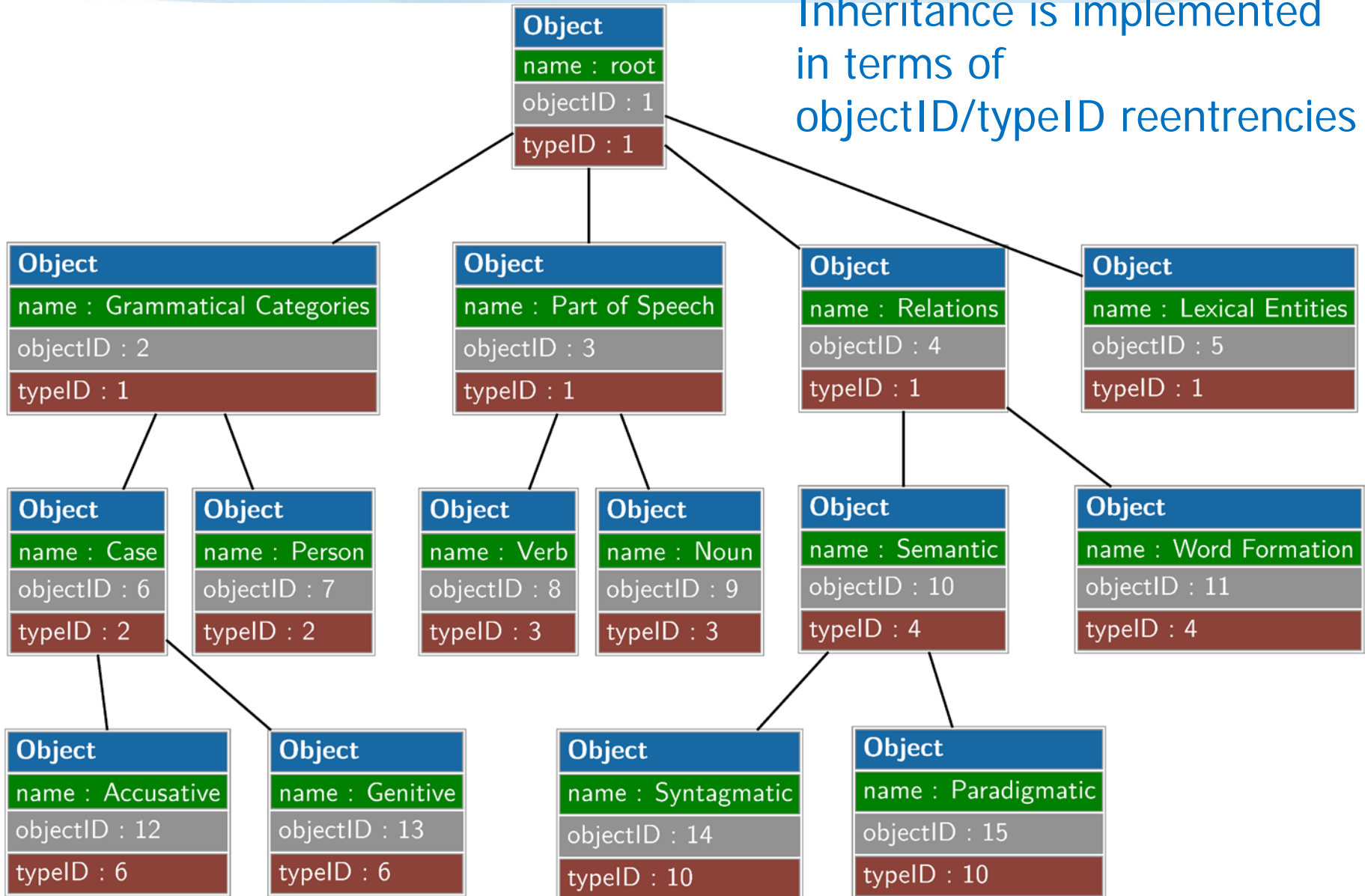
Entity-Relationship Diagram



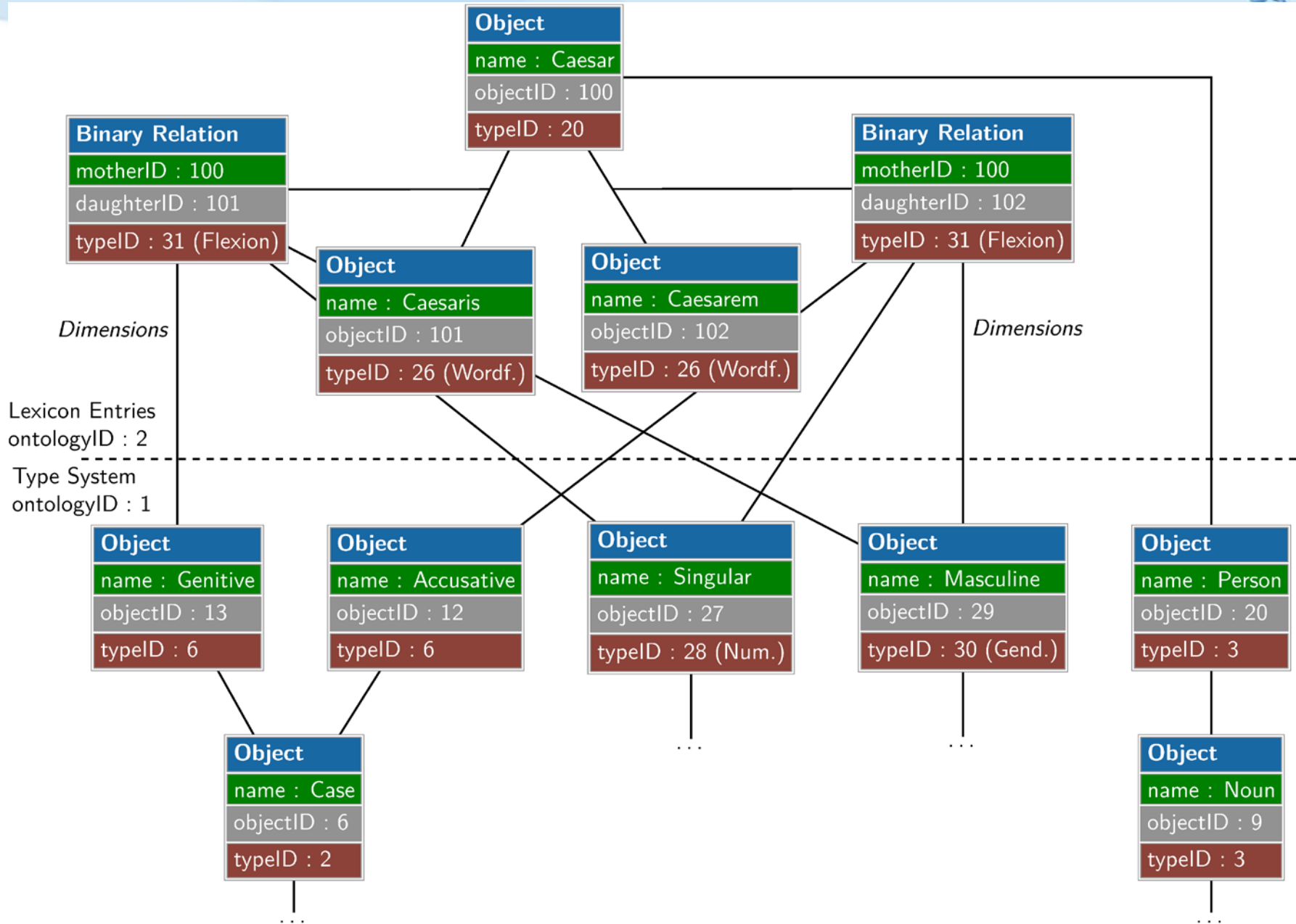
eLexiconType Hierarchy: Excerpt



Inheritance is implemented
in terms of
objectID/typeID reentrancies



eLexicon Data Model: Example Entry *Caesar*





Stock-taking

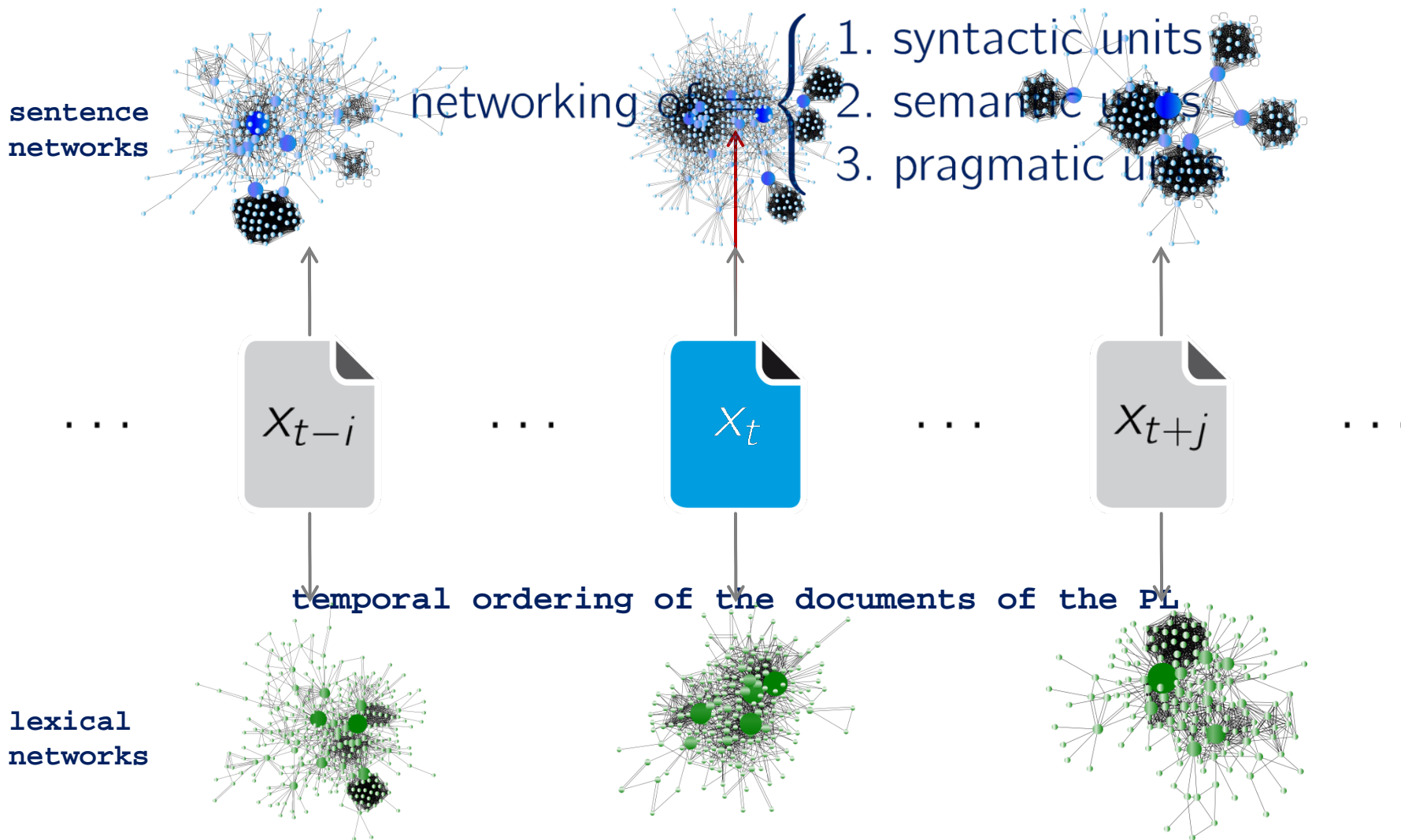
POS	# lemma	# word form
ADJ	23 998	2 396 700
ADV	25 314	90 505
APPR (prep. left)	71	138
CARD	71	1 474
ITJ	140	140
KO (conjunction)	166	166
NN	33 649	442 623
NE	1 052	16 204
NEP	28 213	264 382
V	10 005	3 429 535
ORD	71	2 892
DIST	69	1 590
sums:	122 819	6 646 349



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



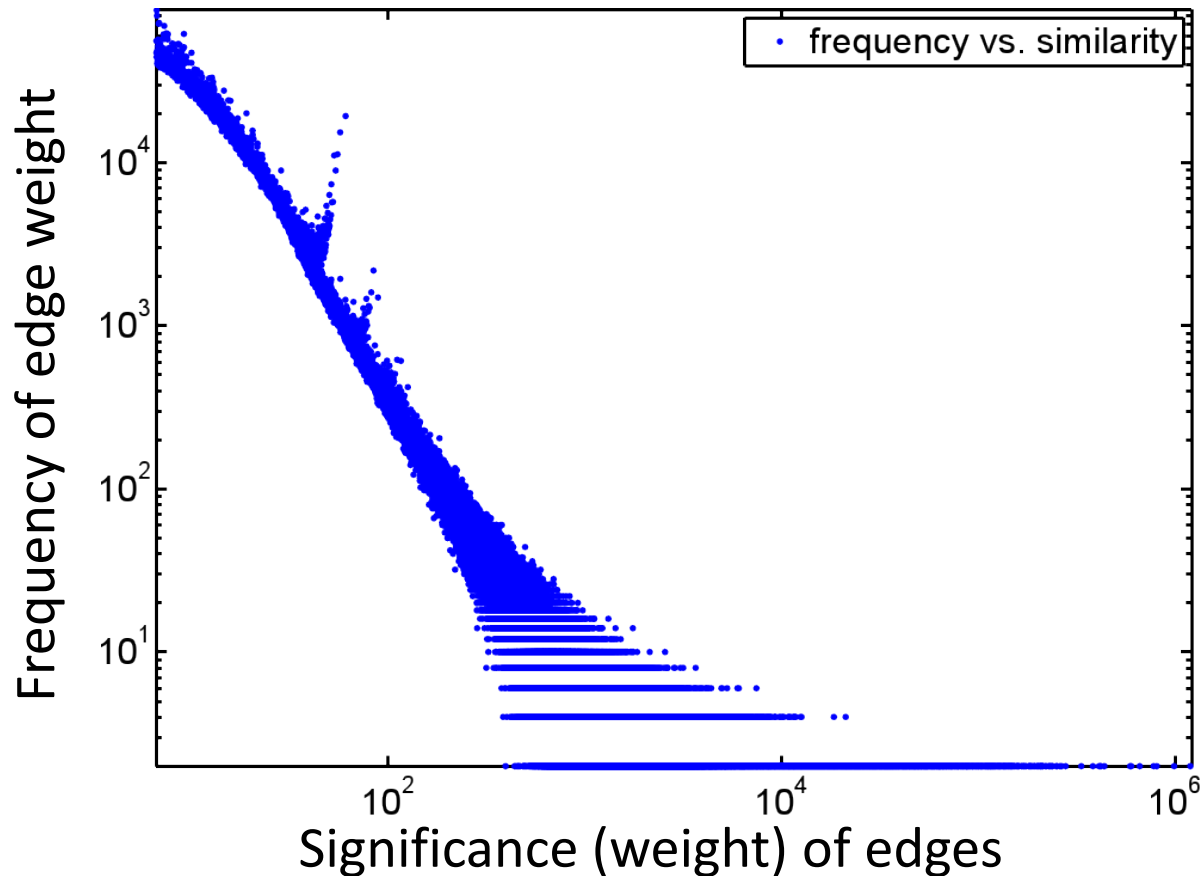
Multilevel Networks





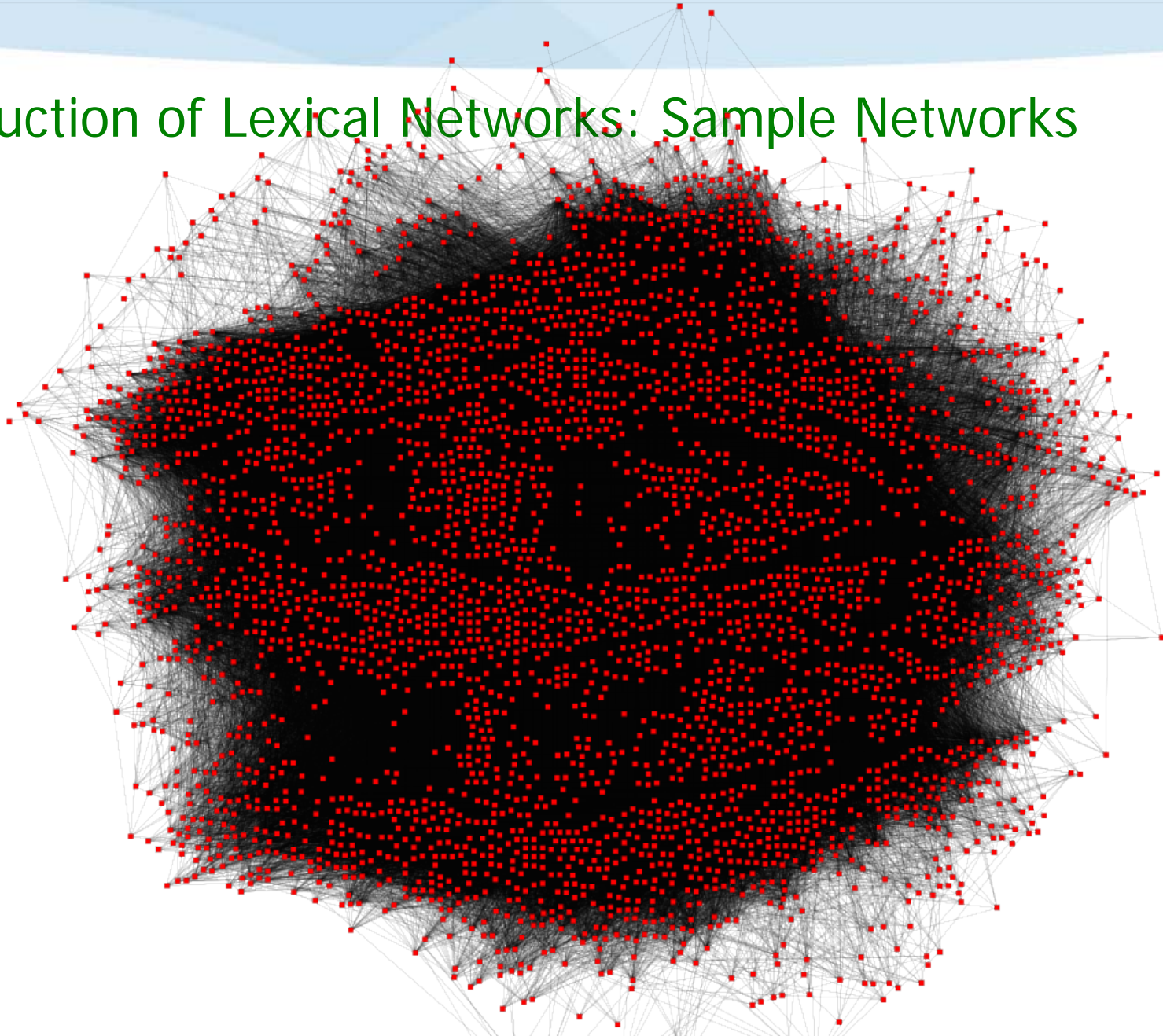
Induction of Lexical Networks (Heyer et al. 2006)

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





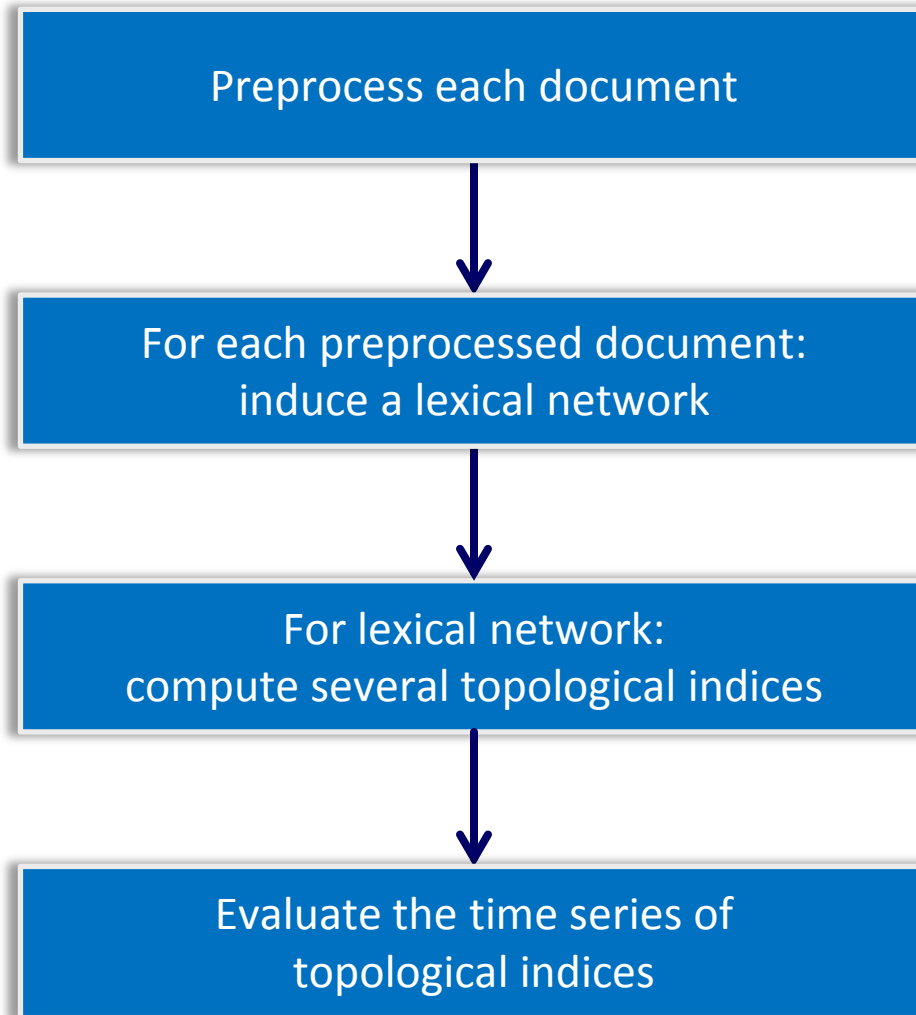
Induction of Lexical Networks: Sample Networks



Time Series of Lexical Networks



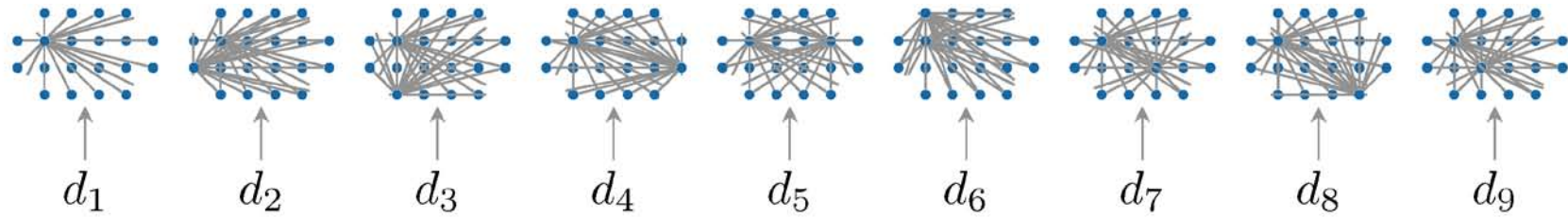
Approach



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

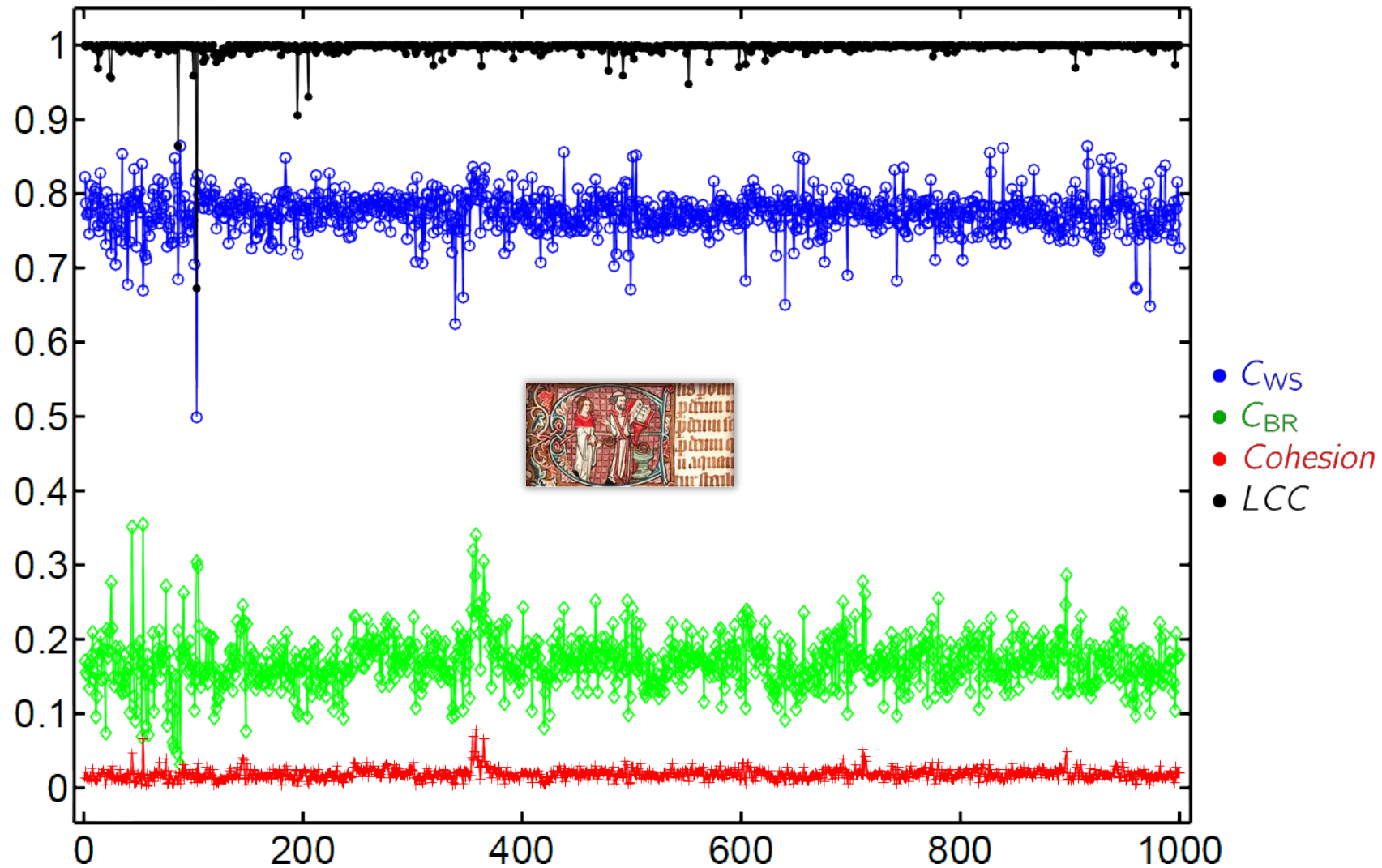


Documents $d_1 \dots d_9$ ordered on a time-line

Time Series of Lexical Networks



Real life example: Patrologia Latina



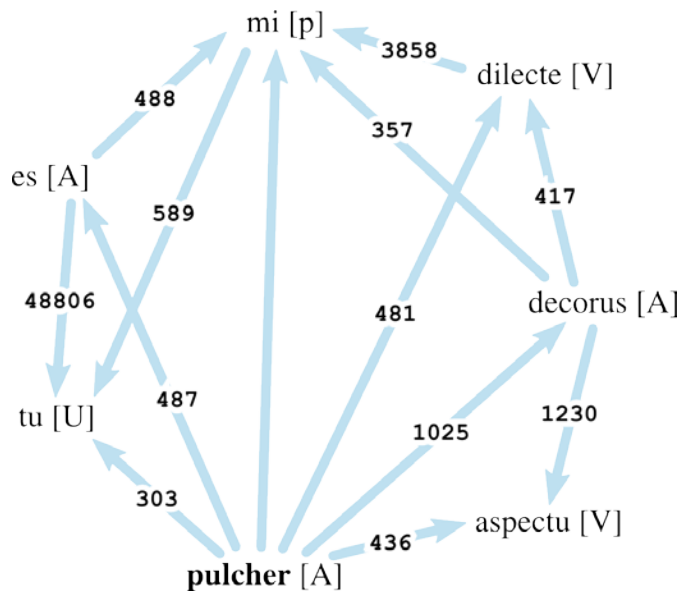


1. Corpus Example: Patrologia Latina
2. eLexicon
3. Time Series
4. Sample Analyses
5. Linguistic Networks Workflow

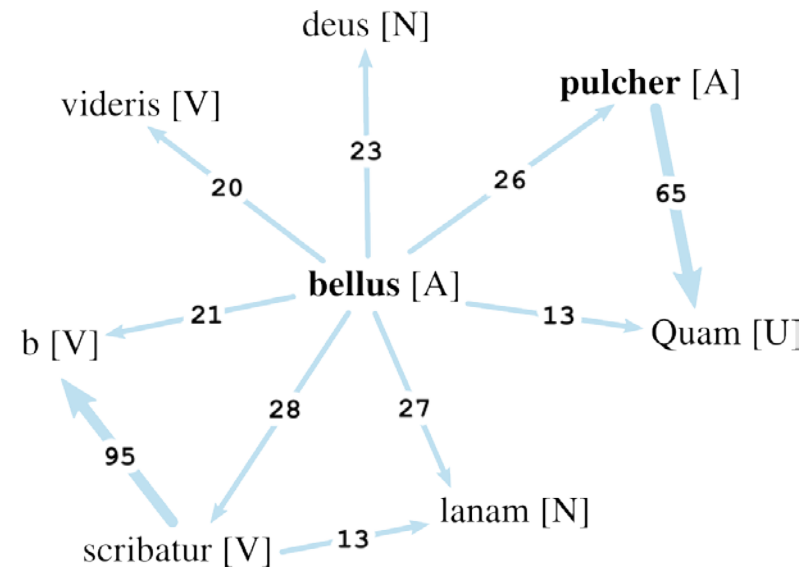
Language Change in a Network Perspective



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

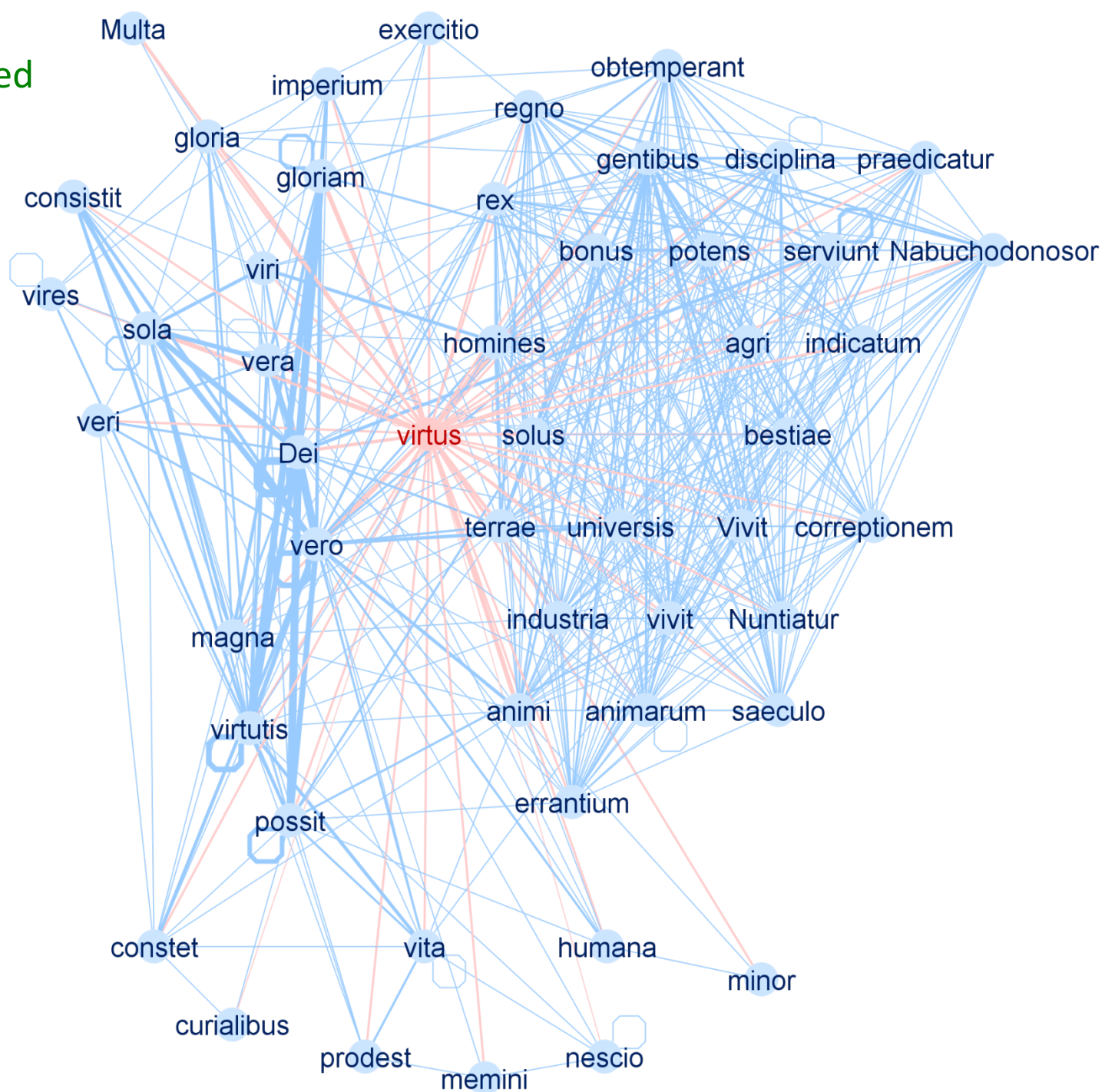
Vellus, if it means *lana* (*wool*), is written with a *v*; if it's *bellus*,
in the meaning of *pulcher*, it should be written with a *b*.

Sonar-word-induced networks:

virtus in

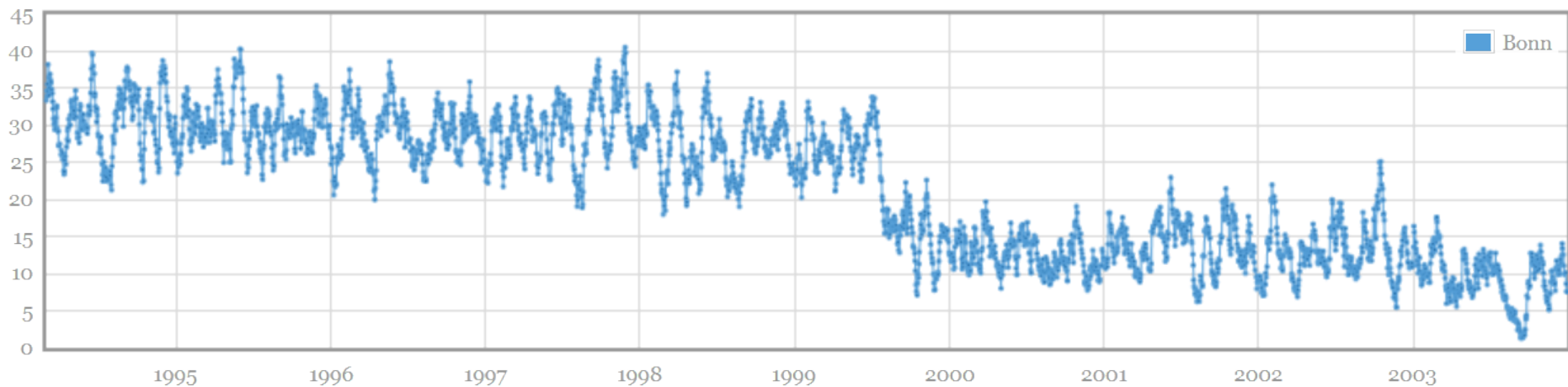
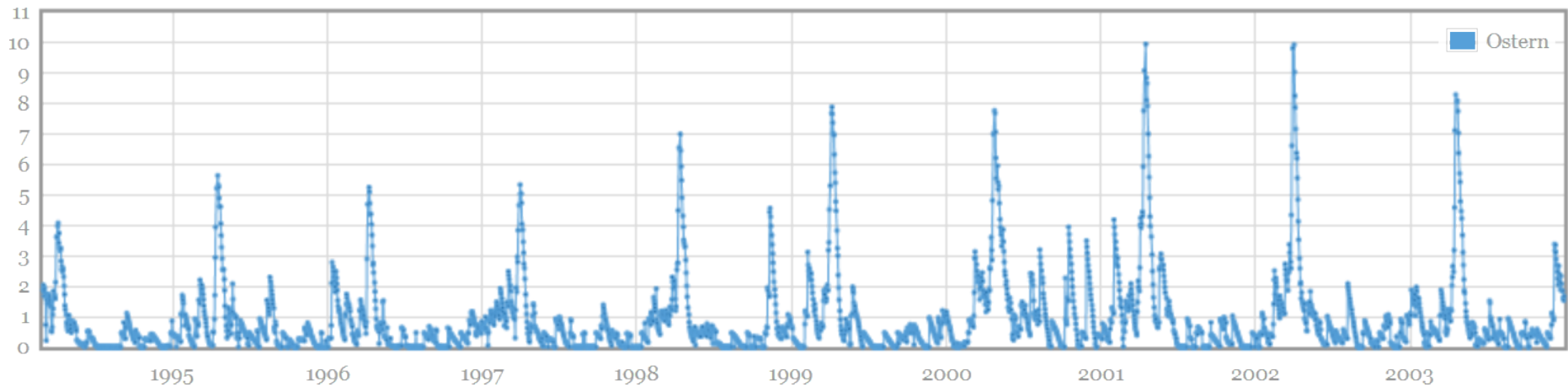
John of Salisbury's

Polycraticus





Temporal Variability of Word Meanings (Laußmann 2010)





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



[more languages](#)

Български език

Română

Slovenščina

Italiano

English

Lingua Latina

Deutsch

Česky

Svenska

русский язык

Dansk

Català

Español



word form
network

lemma
network

sentence
network

text
network

Filter:

Choose language

Choose corpus

Choose network

caesar

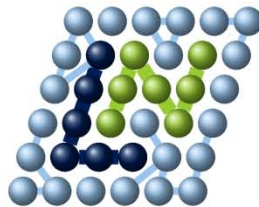
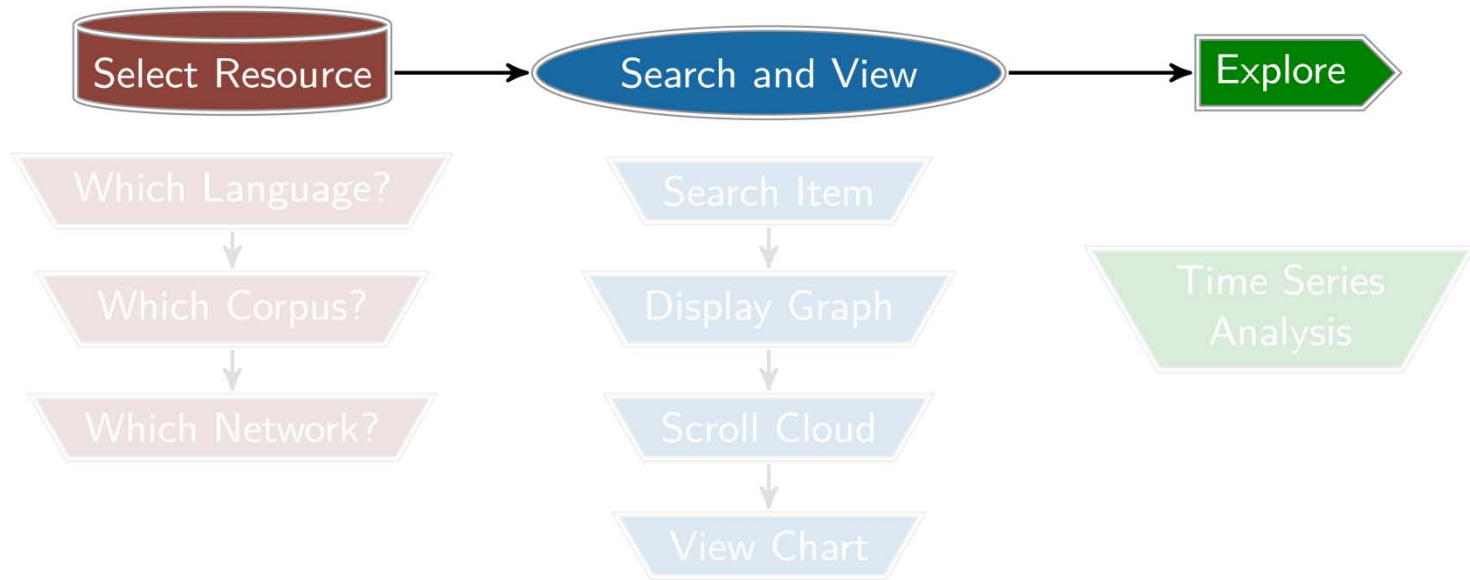
Search

Coverage

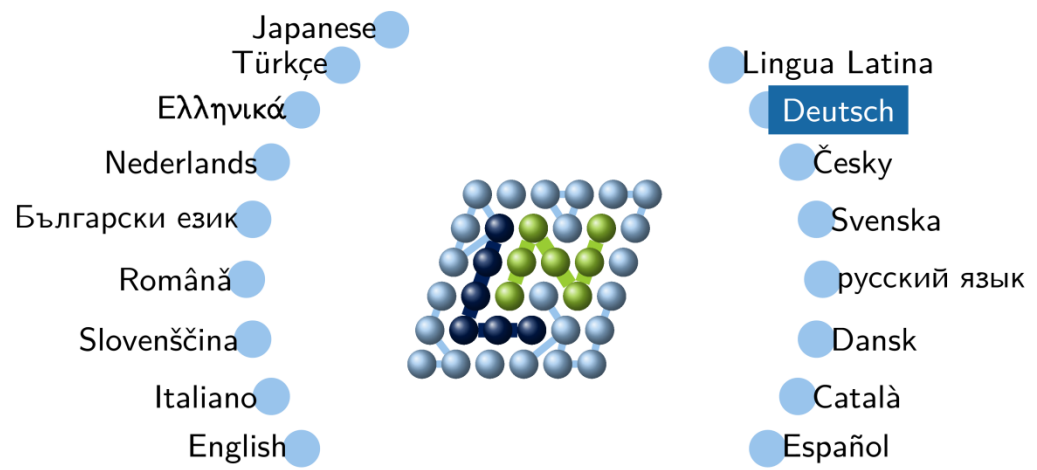
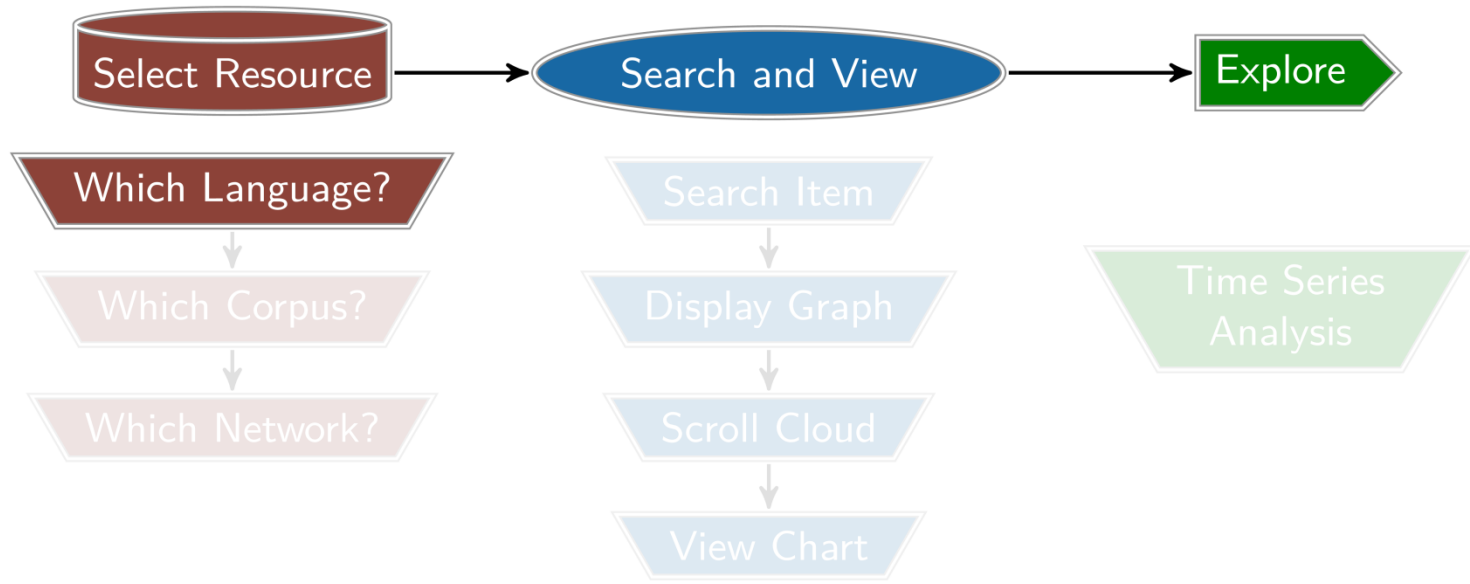
Stock-taking

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

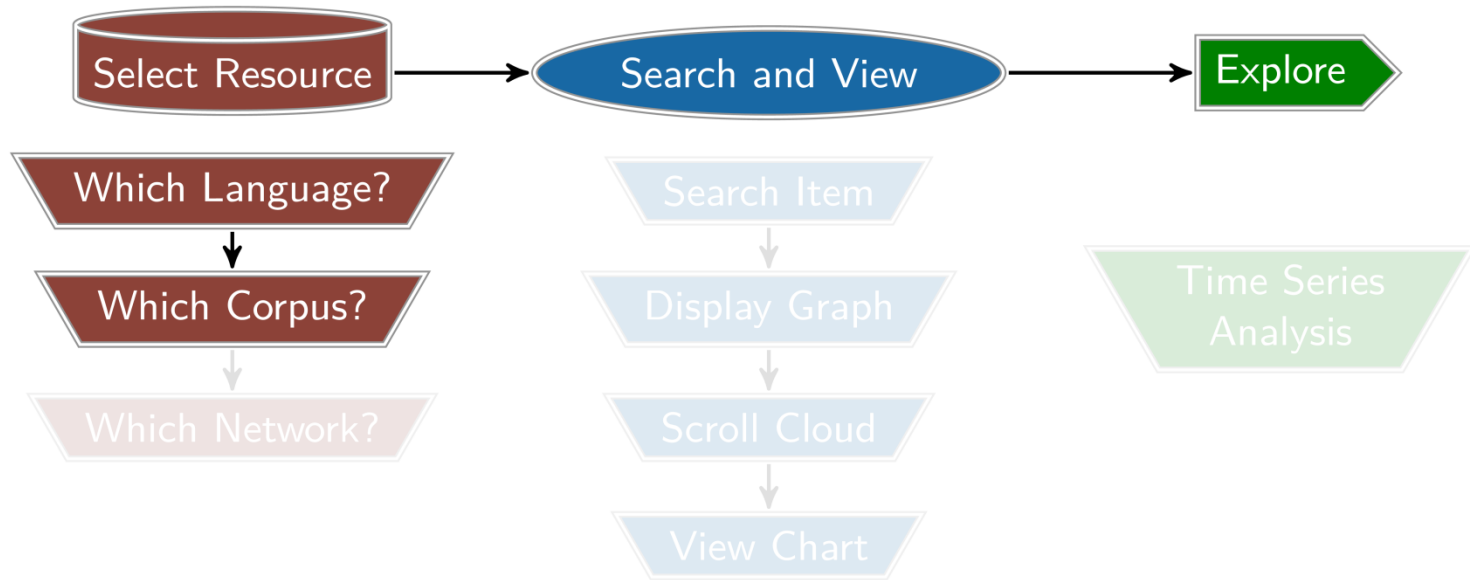
Linguistic Networks: Workflow



Linguistic Networks: Workflow



Linguistic Networks: Workflow



Choose corpus

SZ 1994

Wikipedia

Hofmannsthals Werk

Althochdeutsch

Franz Kafka: Erzählungen

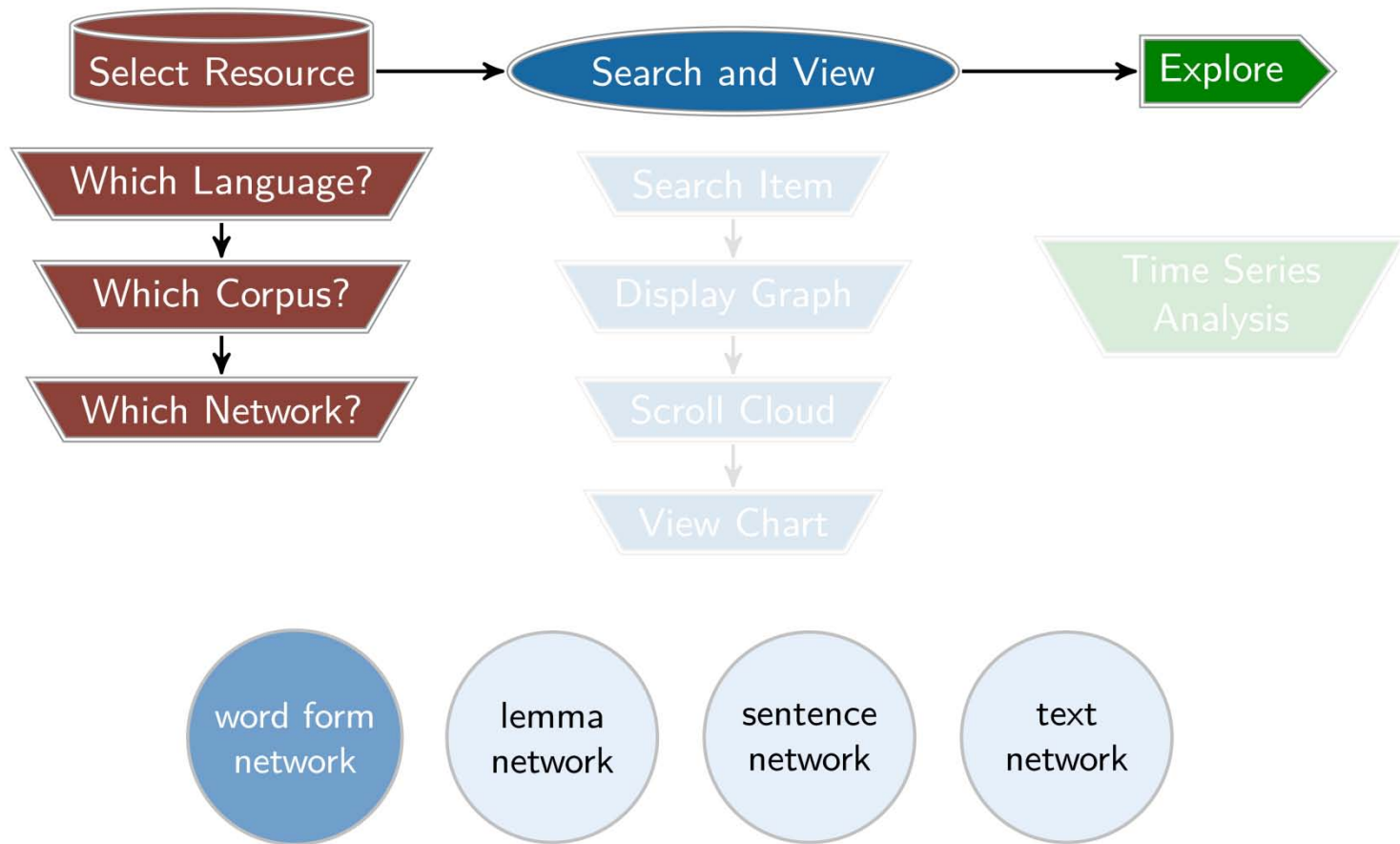
Friedrich Nietzsche: Genealogie der Moral (Ausschnitt)

GuttenPlag Wiki

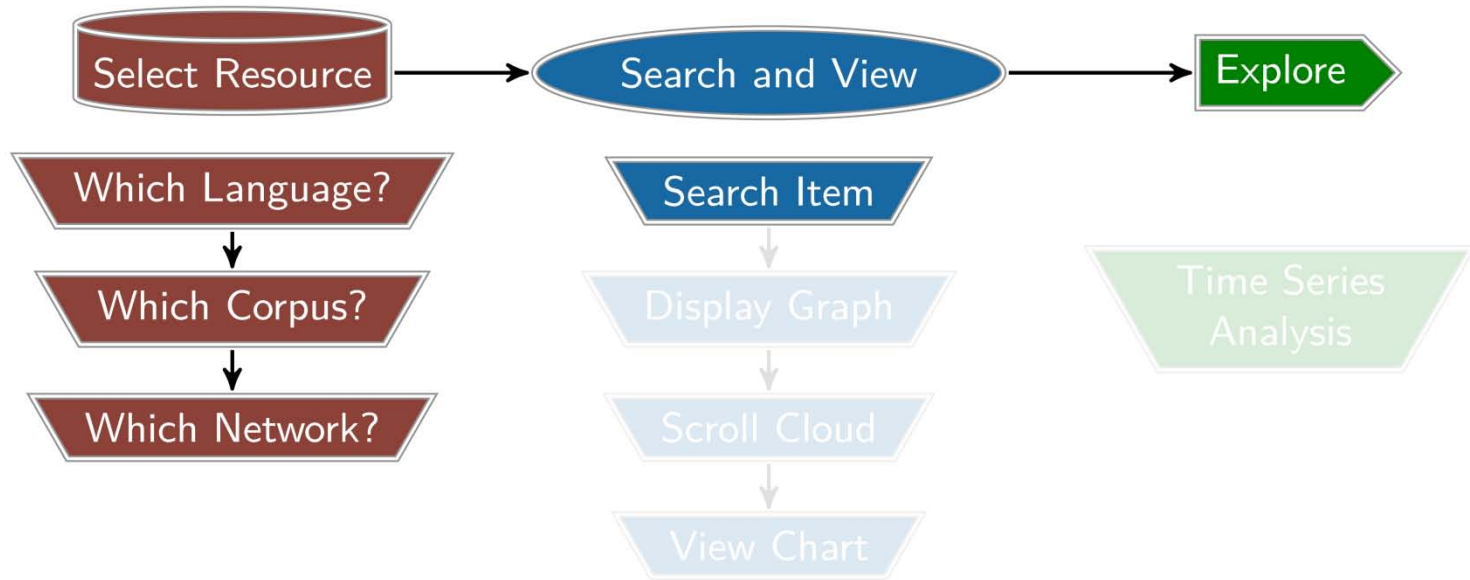
MemoryAlpha Wiki

VroniPlag Wiki

Linguistic Networks: Workflow



Linguistic Networks: Workflow



ABOUT | PROJECT HOME | GRAPHML

LINGUISTIC NETWORKS

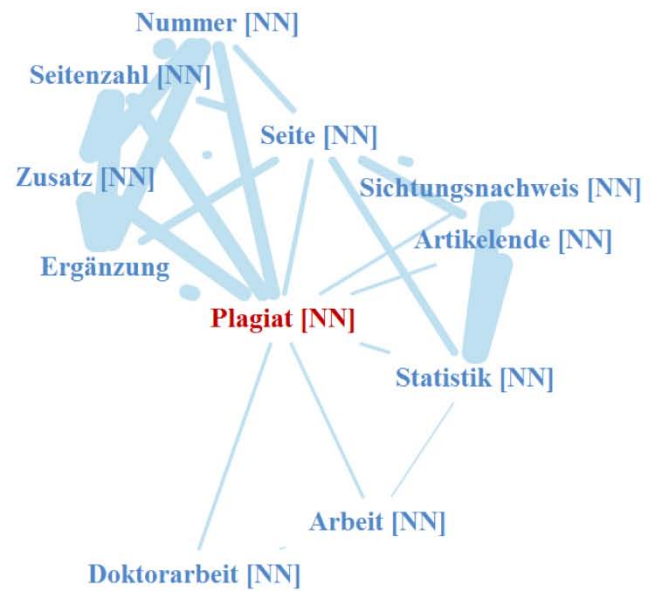
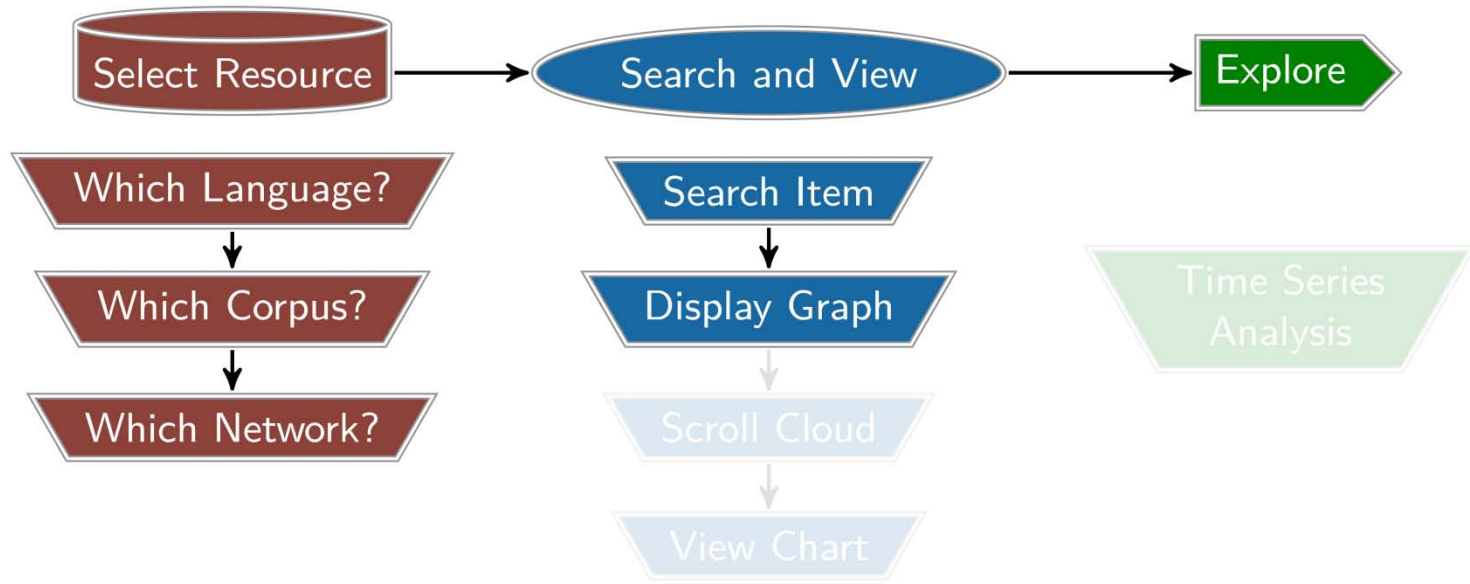
Filter: Deutsch de.guttenplag.wikia.cc word form network Plagiat Search

Plagiat clear history

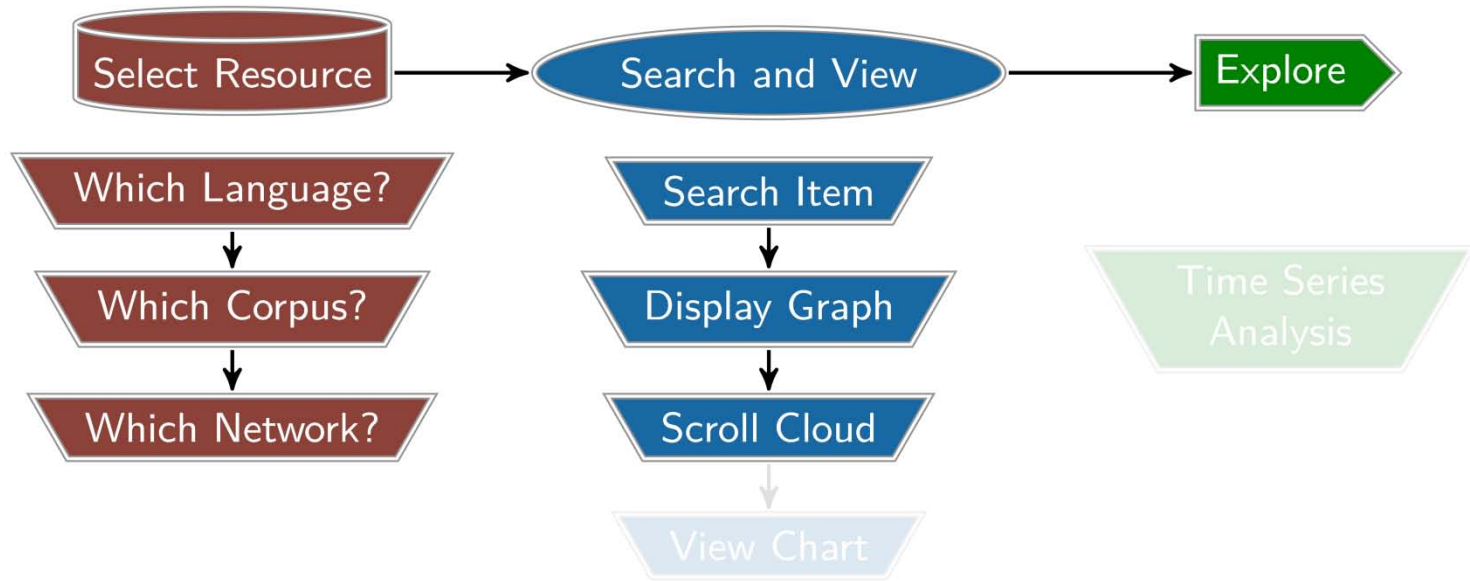
Show part-of-speech-filter

Similar words: ADJA: Plagiat-Spotting (*Plagiat-spotting*),
nachrichten.t-online.de/plagiat-affaeren-in-der-union-guttenberg-schreibt-nicht-als-erster-ab/id_44588866/index
(nachrichten.t-online.de/plagiat-affaeren-in-der-union-guttenberg-schreibt-nicht-als-erster-ab/id_44588866/index)
, Komplet-Plagiat (*Komplet-plagiat*),
multipunkt.blog.de/2011/02/19/karl-plagiat-guttenberg-10634619 (multipunkt.blog.de/2011/02/19/karl-plagiat-guttenberg-
10634619)
, Plagiat-freier (*Plagiat-freier*),
faz-community.faz.net/blogs/antike/archive/2011/04/02/sozialgeschichte-abgeschrieben-erinnerung-an-ein-plagiat.aspx
faz-community.faz.net/blogs/antike/archive/2011/04/02/sozialgeschichte-abgeschrieben-erinnerung-an-ein-plagiat.aspx)
, general-plagiat (*general-plagiat*),
nachrichten.rp-online.de/politik/plagiat-die-taktik-der-doktoren-1.1322942 (nachrichten.rp-online.de/politik/plagiat-
die-taktik-der-doktoren-1.1322942)

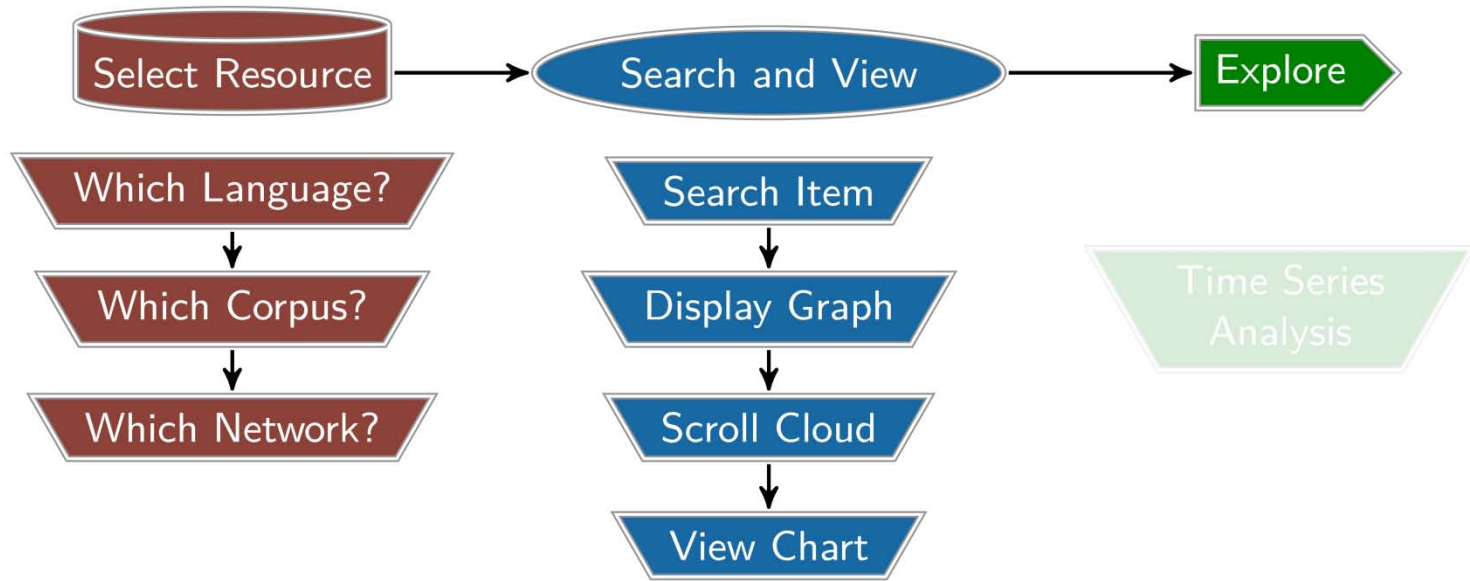
Linguistic Networks: Workflow



Linguistic Networks: Workflow

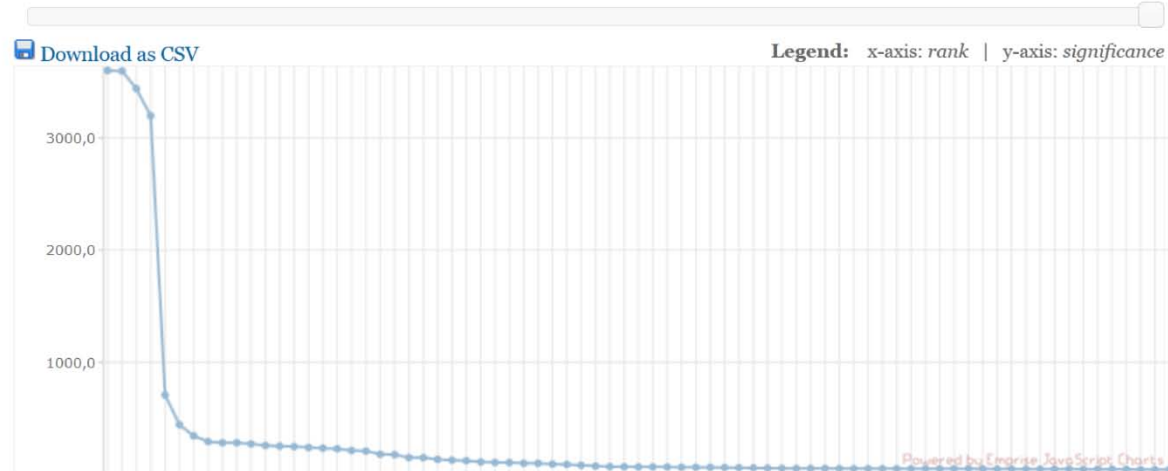


Linguistic Networks: Workflow

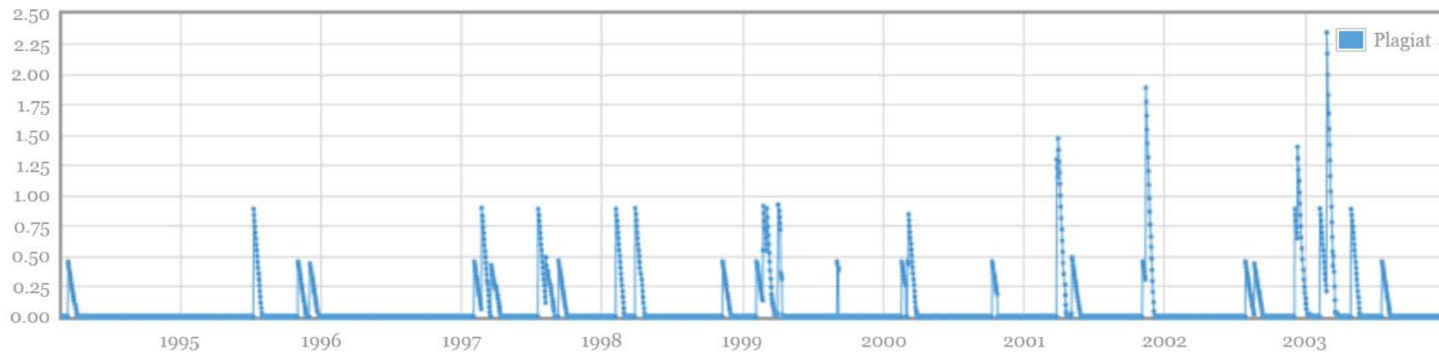
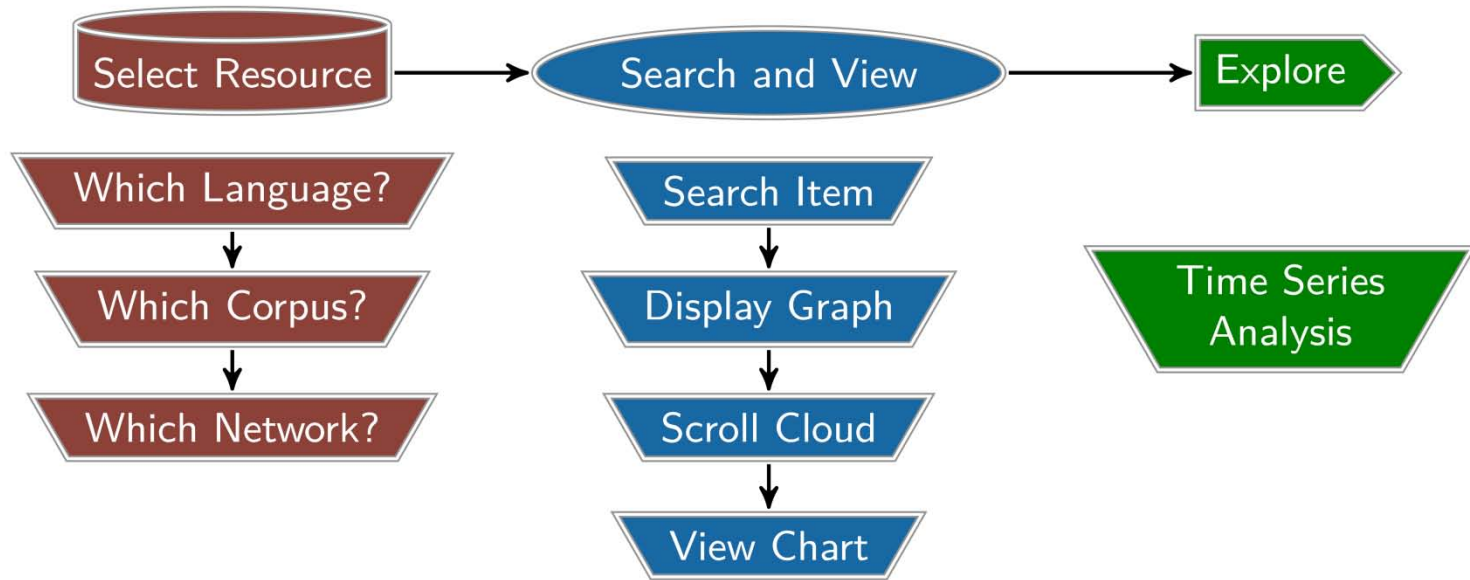


Word chart

Number of words to plot: **75** (Use slider to change value.)



Linguistic Networks: Workflow





www.linguistic-networks.net

 **LOEWE** – Landes-Offensive zur
Entwicklung Wissenschaftlich-
ökonomischer Exzellenz

SPONSORED BY THE



Federal Ministry
of Education
and Research