Natural Language Processing of Historical Languages

Dr. Thomas Köntges,
University of Leipzig,
Open Philology Project,
Perseus Research Group

Digital Humanities
UNIVERSITÄT LEIPZIG

UNIVERSITÄT LEIPZIG
Digital Humanities
UNIVERSITÄT LEIPZIG

Tufts UNIVERSITY

Perseus Research Group (PRG)

Open Philology
UNIVERSITÄT LEIPZIG

PERSEUS DIGITAL LIBRARY

Open Greek and Latin (OGL)

Open Persian and Arabic (OPA)

Historical Languages eLearning Project (HLEP)

CapiTainS

PERSEUS DIGITAL LIBRARY

The Perseus Catalog

PERSEIDS

Leipzig Open Fragmentary Text Series (LOFTS)

Ancient Greek and Latin Dependency Treebank

OPP Manuscript Catalogue

Topic-Modelling for Historical Languages

OGL Digitisation and OCR

AncientGeek

Perseids Collaborative Editing Environment

Perseids Morphology Services

Dolphin

CSEL

PL and PG

Out-of-copyright text-critical editions

OCR Toolset

Digital Athenaeus

Digital Fragmenta Historicorum Graecorum (DFHG)

Digital Marmor Parium

Digital Petronius

Perseids Fragmentary Text Editor

Projects

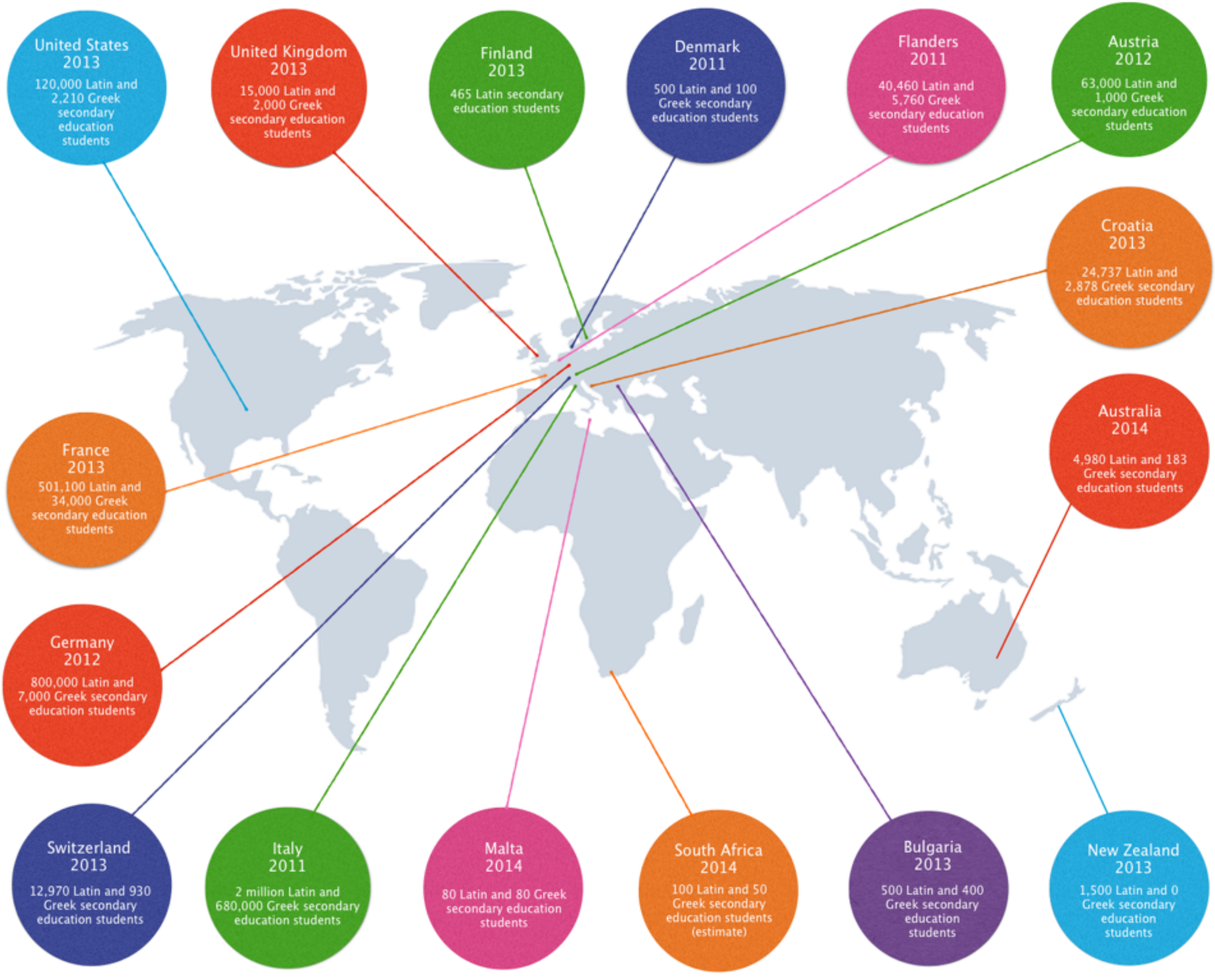
Archives / Editions

Tools / Programmes

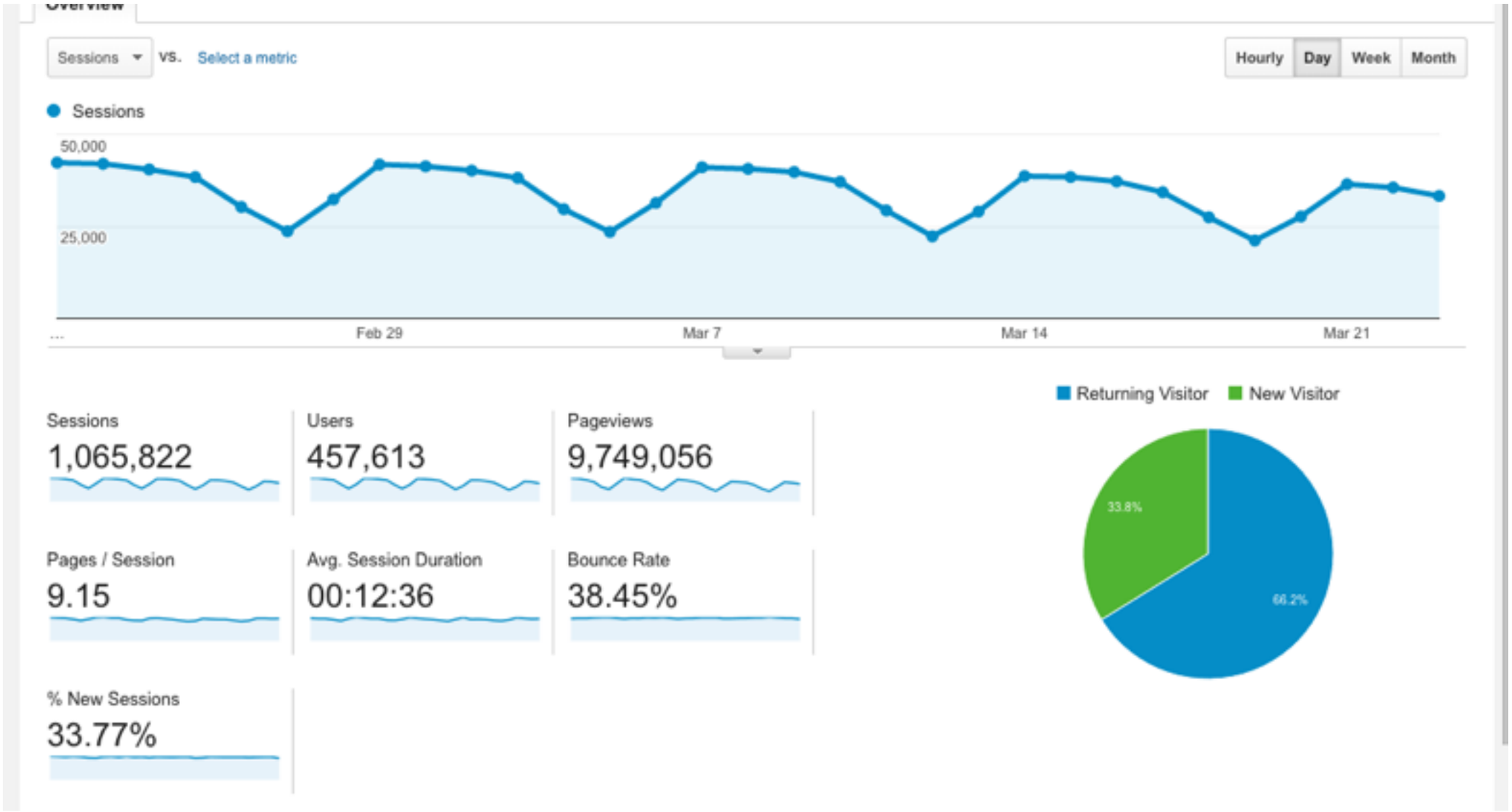
February 2016

Digital Renaissance

- Carolingian renaissance
- Italian renaissance
- digital renaissance
- access to literature, education, and reuse of knowledge


United States
2013
120,000 Latin and 2,210 Greek secondary education students
United Kingdom
2013
15,000 Latin and 2,000 Greek secondary education students
Finland
2013
465 Latin secondary education students
Denmark
2011
500 Latin and 100 Greek secondary education students
Flanders
2011
40,460 Latin and 5,760 Greek secondary education students
Austria
2012
63,000 Latin and 1,000 Greek secondary education students
Croatia
2013
24,737 Latin and 2,878 Greek secondary education students
Australia
2014
4,980 Latin and 183 Greek secondary education students
France
2013
501,100 Latin and 34,000 Greek secondary education students
Germany
2012
800,000 Latin and 7,000 Greek secondary education students
Switzerland
2013
12,970 Latin and 930 Greek secondary education students
Italy
2011
2 million Latin and 680,000 Greek secondary education students
Malta
2014
80 Latin and 80 Greek secondary education students
South Africa
2014
100 Latin and 50 Greek secondary education students (estimate)
Bulgaria
2013
500 Latin and 400 Greek secondary education students
New Zealand
2013
1,500 Latin and 0 Greek secondary education students

Perseus Project visitors



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Natural Language Processing

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

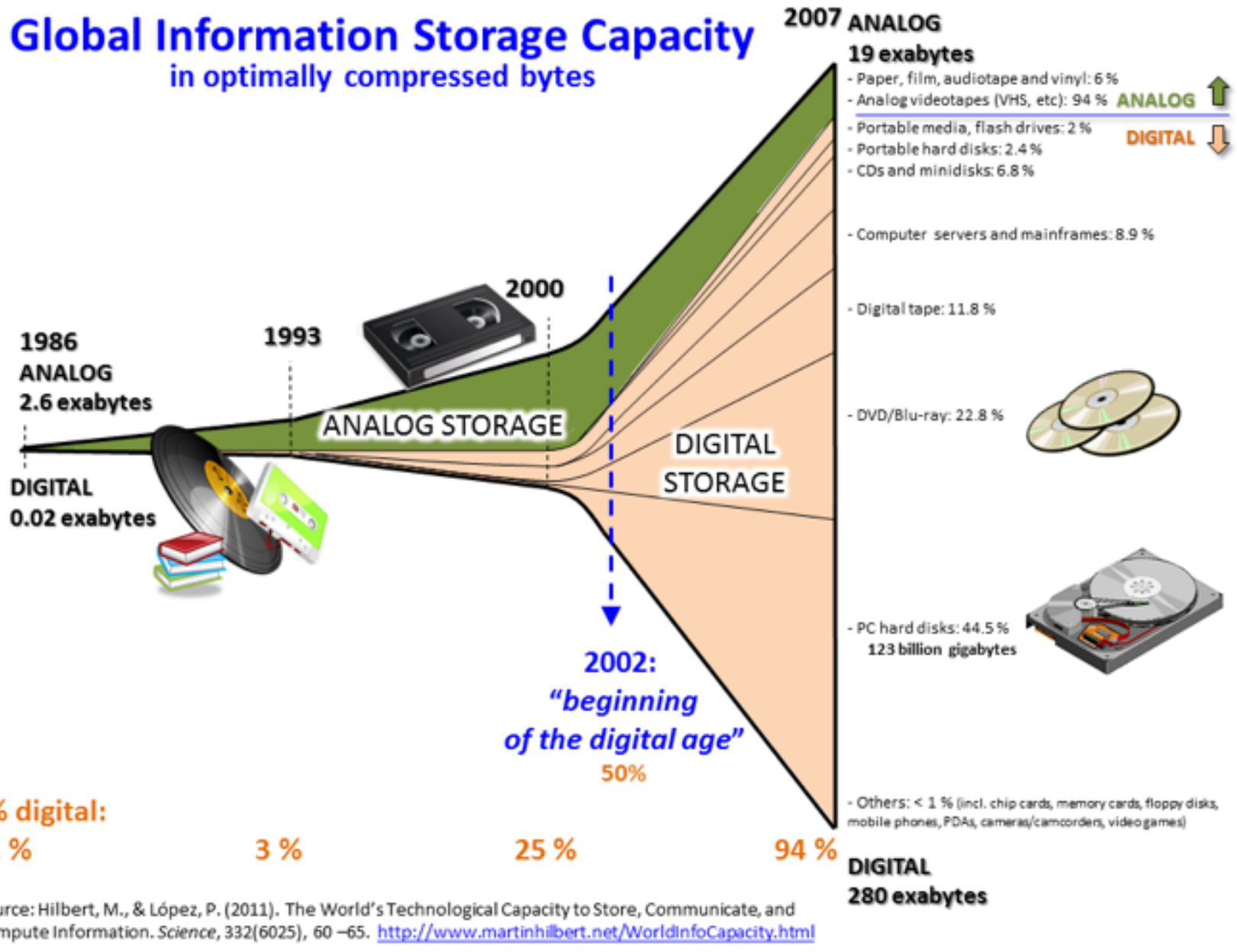
NLP

- NLP refers to the understanding and creation of human languages by a machine
- Today we focus on the former:
  - OCR
  - Machine Translation
  - Part-of-Speech Tagging
  - Topic and word sense disambiguation
  - Information Retrieval

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

The Digital Age



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

There is a lot of Data

The **2.8 Zettabytes** of data created in **2012** alone
are the storage equivalent of
550 editions of Homer per day
for each human being on the planet…

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Textual Analysis

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Analysis

Split the texts into basic units.

E.g. morphemes, words.

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Synthesis

Re-assemble those units
into a new text (in the widest possible sense).

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Examples

- Metrical analysis
- Morphological analysis
- Tree-banking or other morpho-syntactic analyses
- Translation alignment
- Named-Entity mark-up

Historical Languages

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

The complexity of a language changes over time.

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Grammar

- Indo-European languages have the tendency to become less morphologically complex.

- English barely 3 cases, German 4, Ancient Greek 4–5, Latin 5–6, Proto-Indo-European 8.

- English, German, Latin 2 Numbers, Ancient Greek 3.

- NB: Morphological complexity often replaced by syntactic or idiomatic complexity.

- NB: Languages often flatten if number of speakers is increased.

PROBLEM OF DISTANCE & CONTEXT

Time flies like an arrow.

Fruit flies like an apple.

Written Sources

A mono-directional communication
of text-data, graphics, and meta-data
from the past to the present.

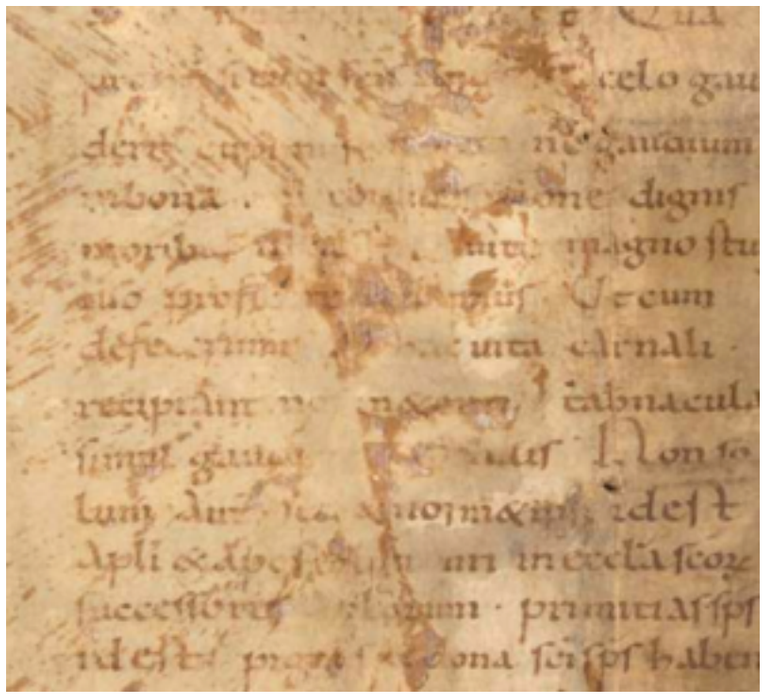
Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

TEXTUAL TRANSMISSION

A mono-directional communication
of text-data, graphics, and meta-data
from the past to the present.

Distribution is Essential



Thomas Köntges
NLP of Historical Languages
June 2016

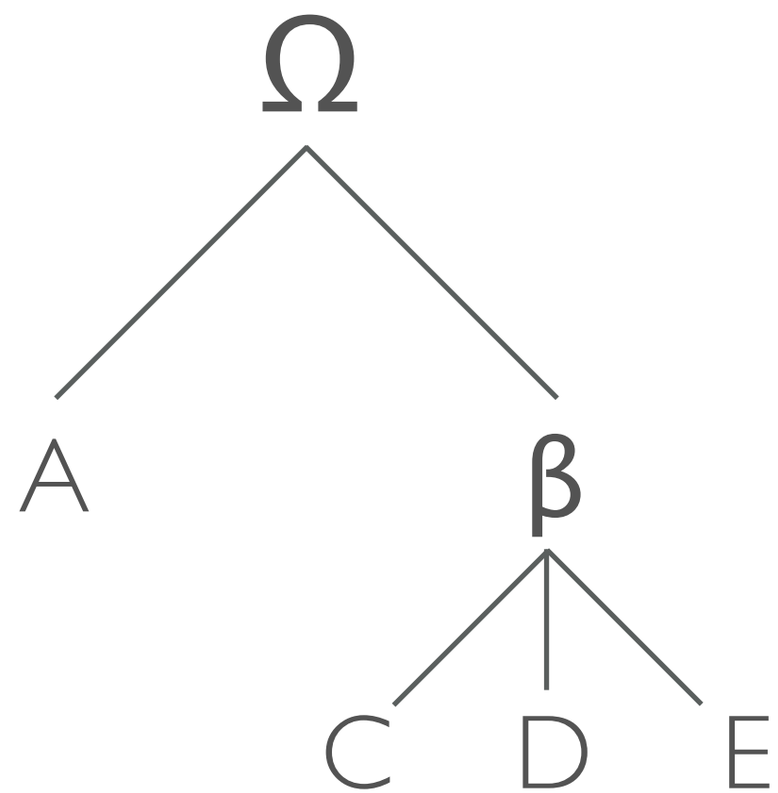
Digital Humanities
UNIVERSITÄT LEIPZIG

Traditional Text-Criticism

- text-criticism is the skill of finding mistakes in ancient sources and the art of correcting — A. E. Housman (1922)

- the main task of text-criticism is to produce a text that resembles as closely as possible the autograph (original text). this process is called *constitutio textus* — P. Maas (1950)

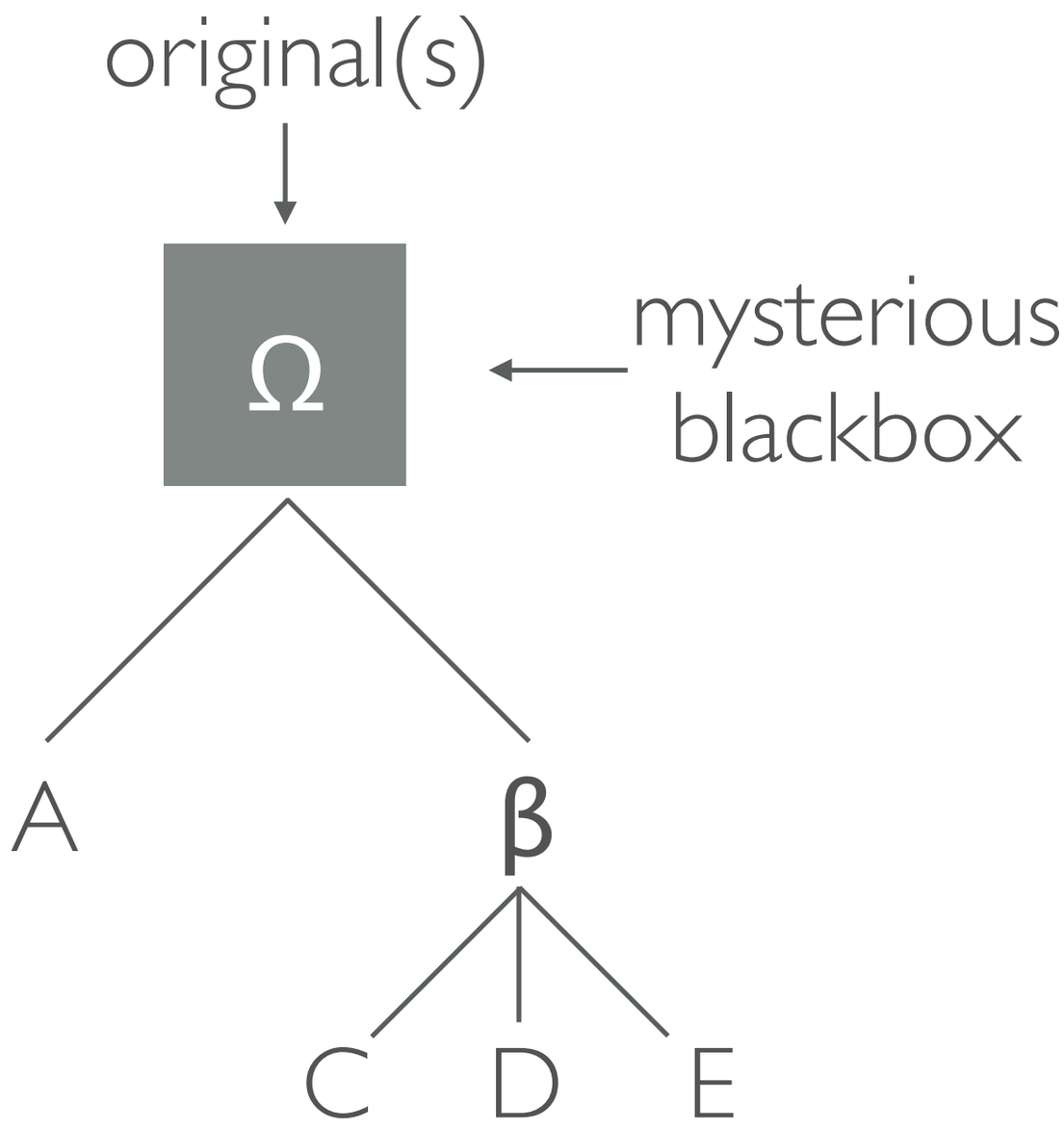
Archetype



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Archetype



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Archetype

an archetype represents a **virtual entity** that is the result and **summary** of all **unknown transmission processes** and changes the **original text** were subject to and that usually predates the oldest known/extant manuscripts of that text.

because an original text can have multiple originals (e.g. multitext, growing texts, canonical texts), it can have more than one archetype (as well as possible transmission-dependent hyparchetypi of each transmission line).

Example: Petronius

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Why Petronius?

- fragments of a fictional **prosimetric** text (35,000 words)
- fragments which look **connected** and build the base for our modern text (33,000)
- **references** in the text to **earlier passages** (23 references)
- **unconnected fragments** and carmina (51 fragments, at least 5 to 8 refer to pre-cena plot)
- transmitted along complex transmission lines in at least 60 manuscripts (4 main families) and hundreds of editions.

Prosimetric

prose

(§108.13) *data ergo acceptaque ex more patrio fide protendit ramum oleae a tutela navigii raptum, atque in colloquium venire ausa*

‘Quis furor’ exclamat ‘pacem convertit in arma’

prose sentence’s predicate

verse

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

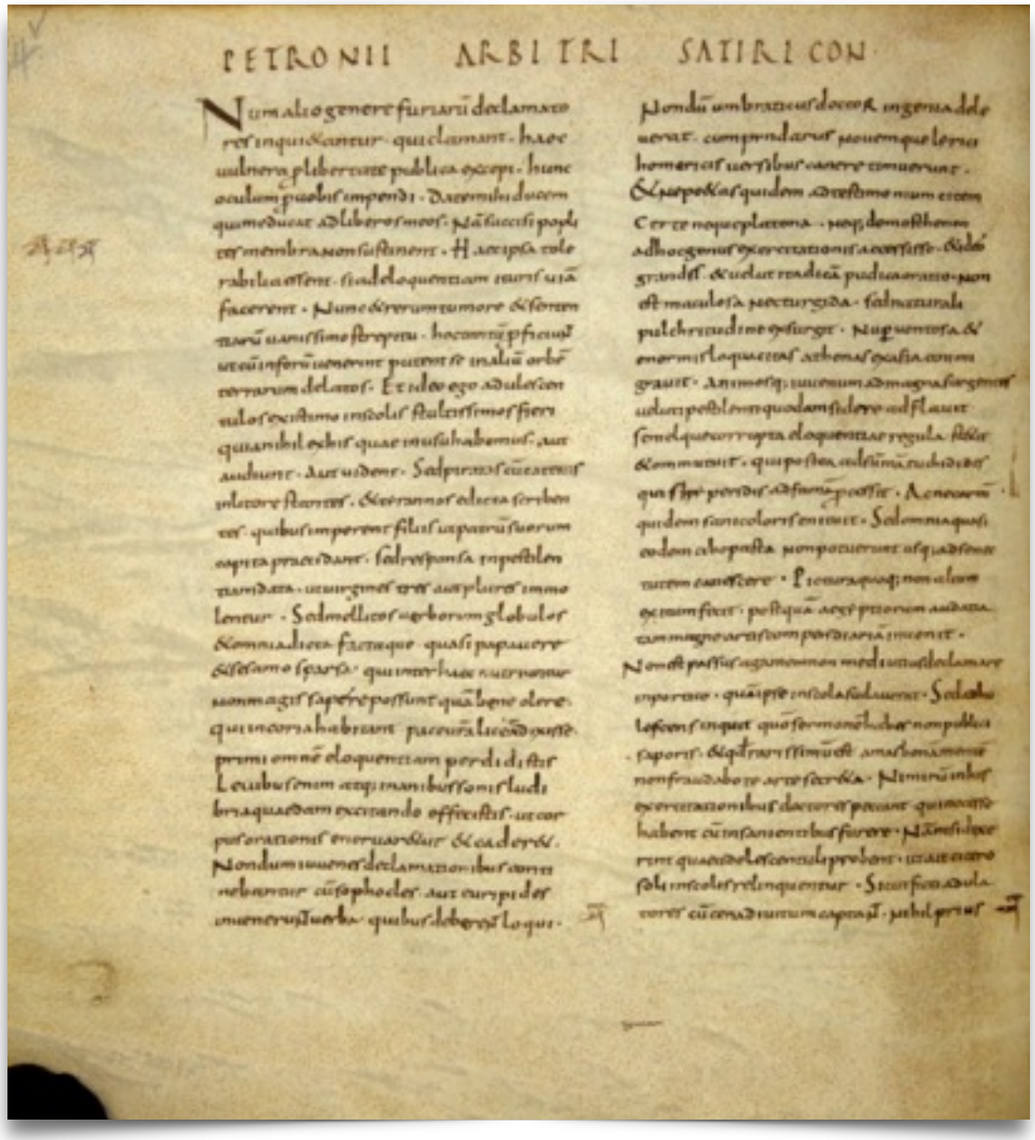
Connected Fragments

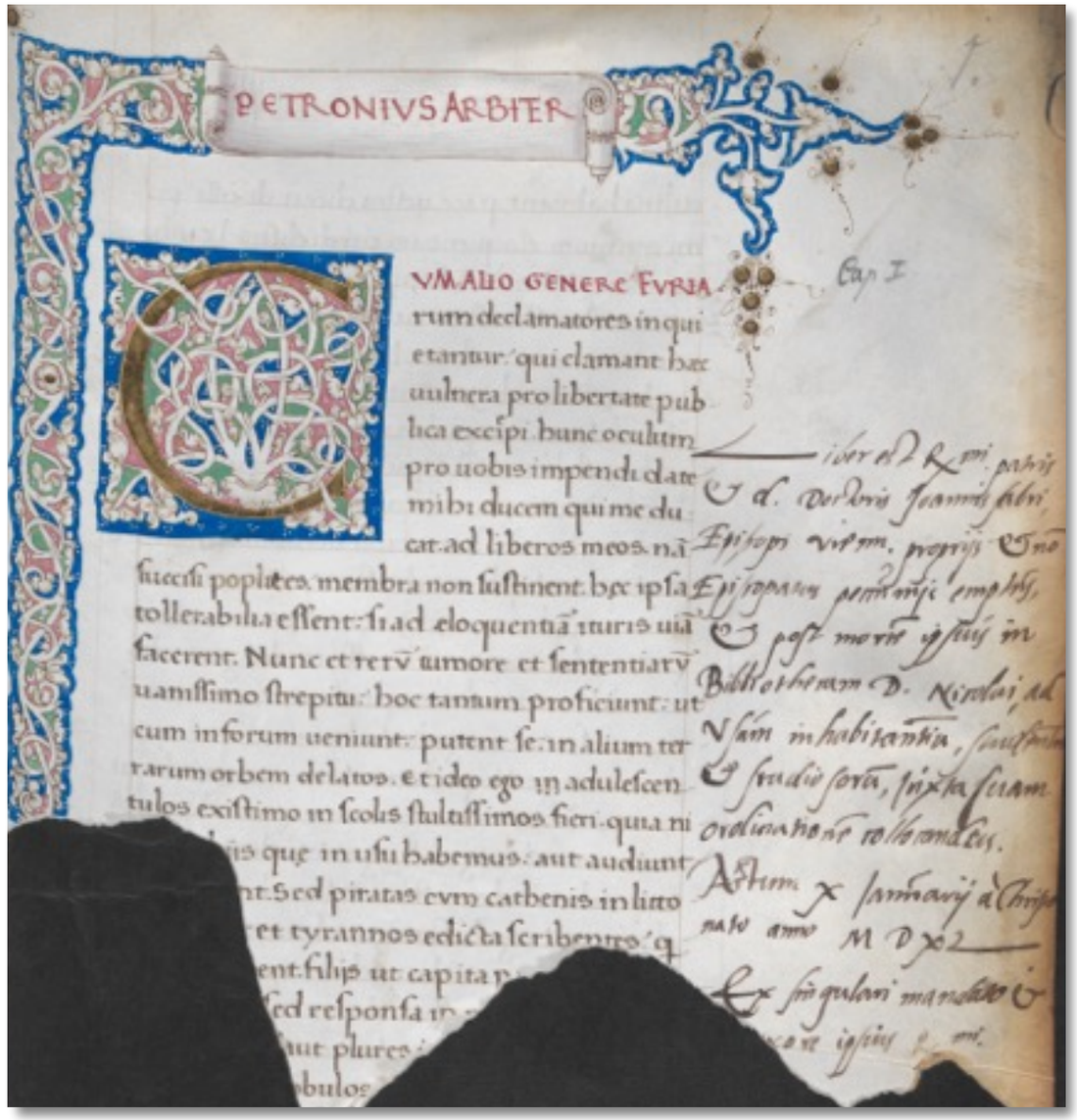
Text transmitted by the manuscripts					
Content	Chapters	Words	O	L	H
School Episode	1–5	700	yes	yes	
Brothel Episode	6–8	300	yes	yes	
Rape of Giton	9–9.5	100	only 9.5	yes	
Quarrel	9.6–11	300		yes	
Market Episode	12–15	700	poem 14.2	yes	
Quartilla's Entrance	16–19.1	500	yes	yes	
Orgy	19.2–26.6	1100	fr.	yes	
Bridge to *Cena*	26.7–27.1A	100			yes
Beginning *Cena*	27.1B–37.5	1800		yes *	yes
Cena	37.6–78	10500	only 55.4–6	fr.	yes
Ascyltos takes Giton	79–80	500	poem 80.9	yes	
Encolpius' Solitude	81–82	350	only 81.1–2	yes	
Meeting Eumolpus	83–84	350	fr.	yes	
Boy of Pergamon	85–87	600		yes	
Eumolpus about Literature	88	300	yes	yes	
Halosis Troiae	89	450		yes	
Reaction of Pedestrians	90	150		yes	
Reconnection with Giton	91–94	300	fr.	yes	
Eumolpus vs. *deversitor*	95–96	300	yes	yes	
Ascyltos' Search	97–99	650		yes	

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Manuscripts





Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Editions

SATYRICON. 33

dentibus oleum infuderat; & pueri, deterſis pauliſper oculis, redierant ad miniſterium, quum 93 intrans cym-

nus inſulſtre, primum Architriclino Vinum adlatum, non vero ut biberet, ſed ut Guſtaret tantum: *Vt autem Guſtauit Architriclinus aquam vinum factam*. ait D. Ioannes c. 11. Ubi indubie ejus indicatum fuiſſe officium animadvertent Viri Docti, Indeque jam & Inſcriptiones Lapidum clarere:

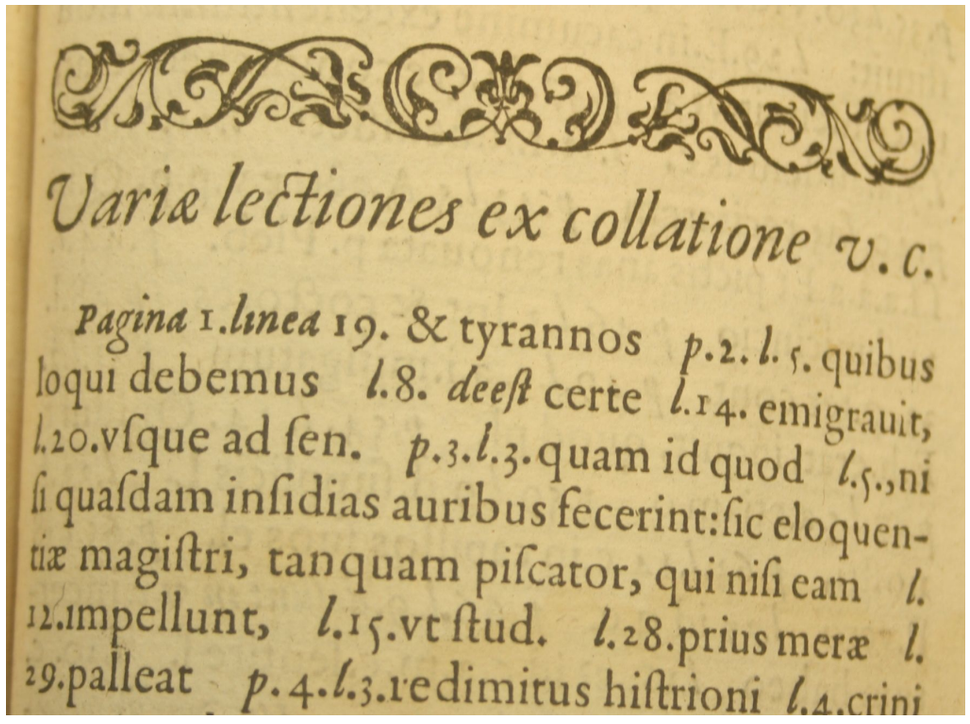
A. LAGUNA. TRICLINIARCHA, *Conſalnu.*

93. *Intrans Cymbaliſtria, & concrepans æra.* Mulieres, quæ cymbala quatiebant, pro analogia vocamus CYMBARISTRIAS, vidimuſque eas ſæpius Romæ in ſcalpturis Bacchanalium ſacrorum, è quibus hæc damus.

Hinc autem adparet quomodo cymbala colliderentur. quod intelligens Amlarius Fortunatus ſcripeum reliquit lib. III. c. III.

cymbala invicem tangantur ut ſonent: ideo à quibuſdam labiis noſtris comparata ſunt. Dioſcorides lib. IV. teſtatur cymbala ejus figuræ fuiſſe, qua

C

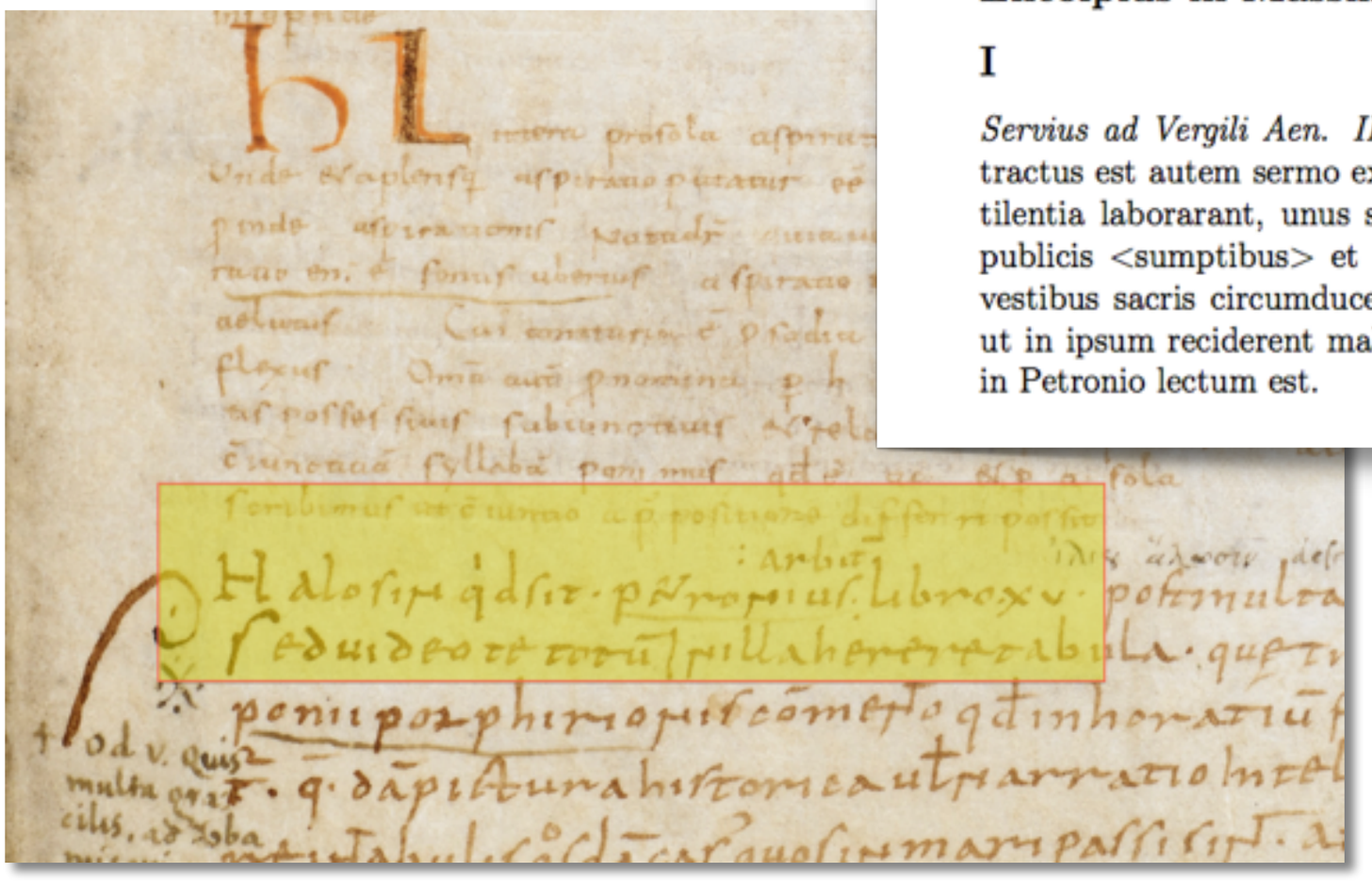


Variæ lectiones ex collatione V. C.

Pagina 1. *linea* 19. & tyrannos *p.* 2. *l.* 5. quibus loqui debemus *l.* 8. *deeſt* certe *l.* 14. emigrauit, *l.* 20. vſque ad ſen. *p.* 3. *l.* 3. quam id quod *l.* 5. ni ſi quaſdam inſidias auribus fecerint: ſic eloquentiæ magiſtri, tanquam piſcator, qui niſi eam *l.* 12. impellunt, *l.* 15. vt ſtud. *l.* 28. prius meræ *l.* 29. palleat *p.* 4. *l.* 3. redimitus hiſtrioni *l.* 4. crini

Thomas Köntges
NLP of Historical Languages
June 2016

References



Fragmenta Petroniana

Encolpius in Massilia

I

Servius ad Vergili Aen. III 57: auri sacra fames] sacra id est exsecrabilis. tractus est autem sermo ex more Gallorum. nam Massilienses quotiens pestilentia laborarant, unus se ex pauperibus offerebat, alendus anno integro publicis <sumptibus> et purioribus cibis. hic postea ornatus verbenis et vestibus sacris circumducebatur per totam civitatem cum exsecrationibus, ut in ipsum reciderent mala totius civitatis, et sic proiciebatur. hoc autem in Petronio lectum est.

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Self-References

Fragmenta de Petronio Approbata

Conspectus (FrPA I)

FrPA I (= 81.3b–5)

81.3b | effugi iudicium, harenae imposui, hospitem occidi, ut inter <tot> audaciae nomina mendicus, exul, in deversorio Graecae urbis iacerem desertus?... *L*
4 adulescens omni libidine impurus et sua quoque confessione dignus exilio,
stupro liber, stupro ingenuus, cuius anni ad tesseram venierunt, quem tam-
5 quam puellam conduxit etiam qui virum putavit. quid ille alter? qui 5
[tamquam] die togae virilis stolam sumpsit, qui ne vir esset a matre per-
suasus est, qui opus muliebre in ergastulo fecit, qui postquam conturbavit
et libidinis suae solum vertit, reliquit veteris amicitiae nomen et, pro pudor,
tamquam mulier secutuleia unius noctis tactu omnia vendidit.

The sanctury of Priapus (FrPA XXI–XXII)

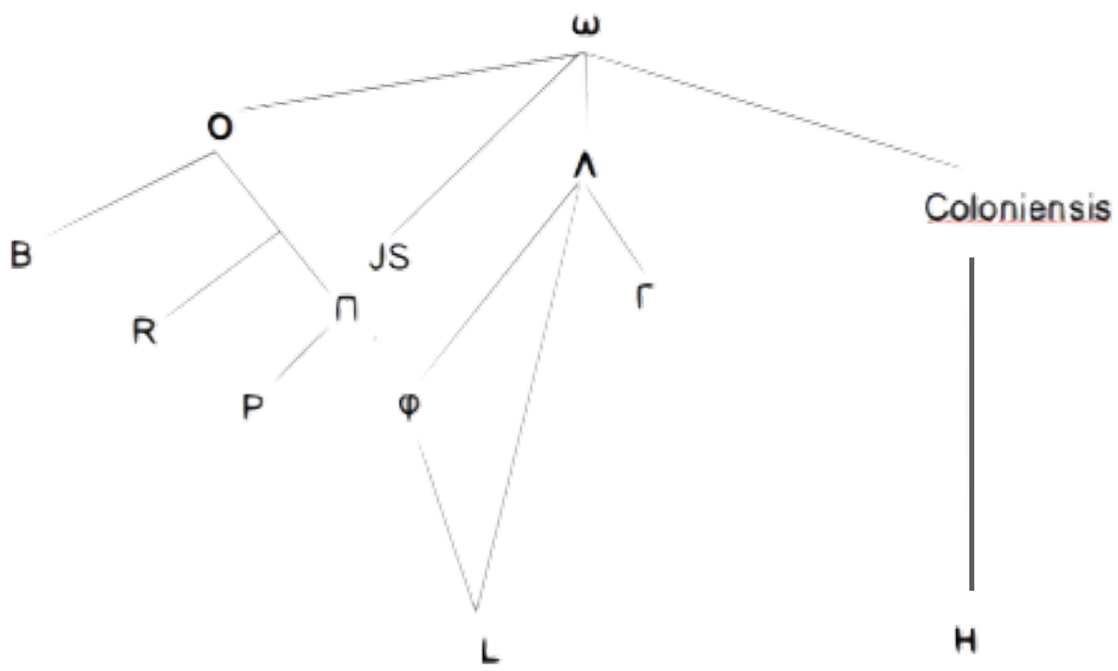
FrPA XXI (= 16.3)

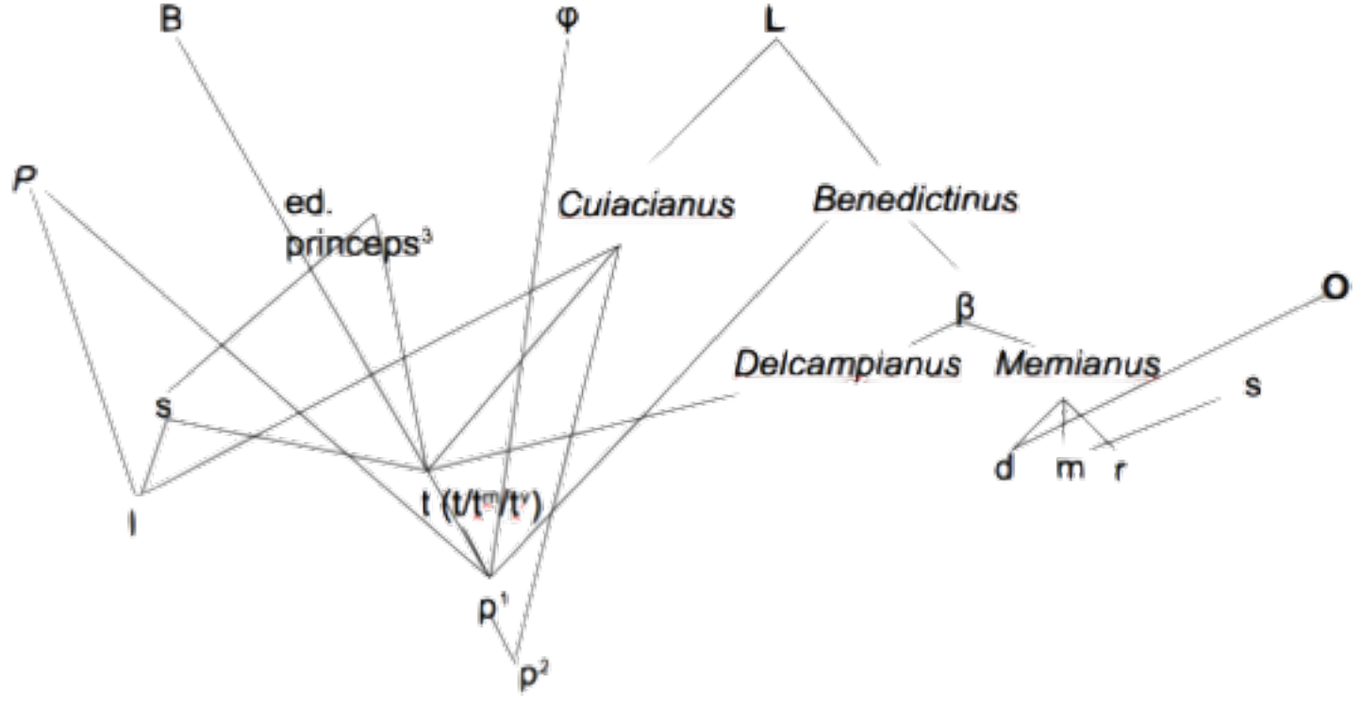
16.3 |'ego sum ancilla Quartillae, cuius vos sacrum ante cryptam turbastis. *LO*

Thomas Köntges
NLP of Historical Languages
June 2016

Reconstructions

- Bücheler 1862, editio maior (14 manuscripts)
- Müller 2009, Satyricon reliquiae (24 manuscripts)

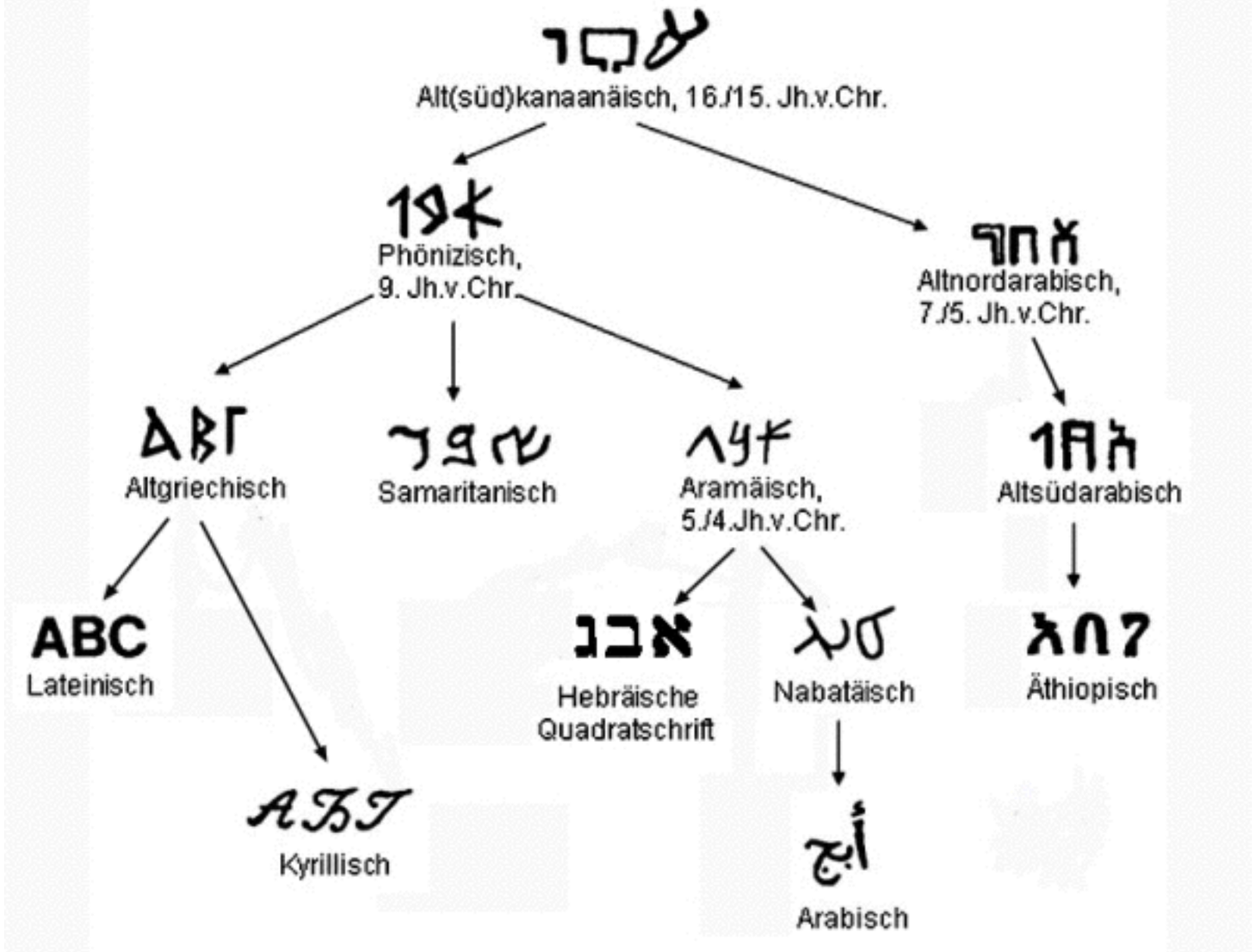




Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

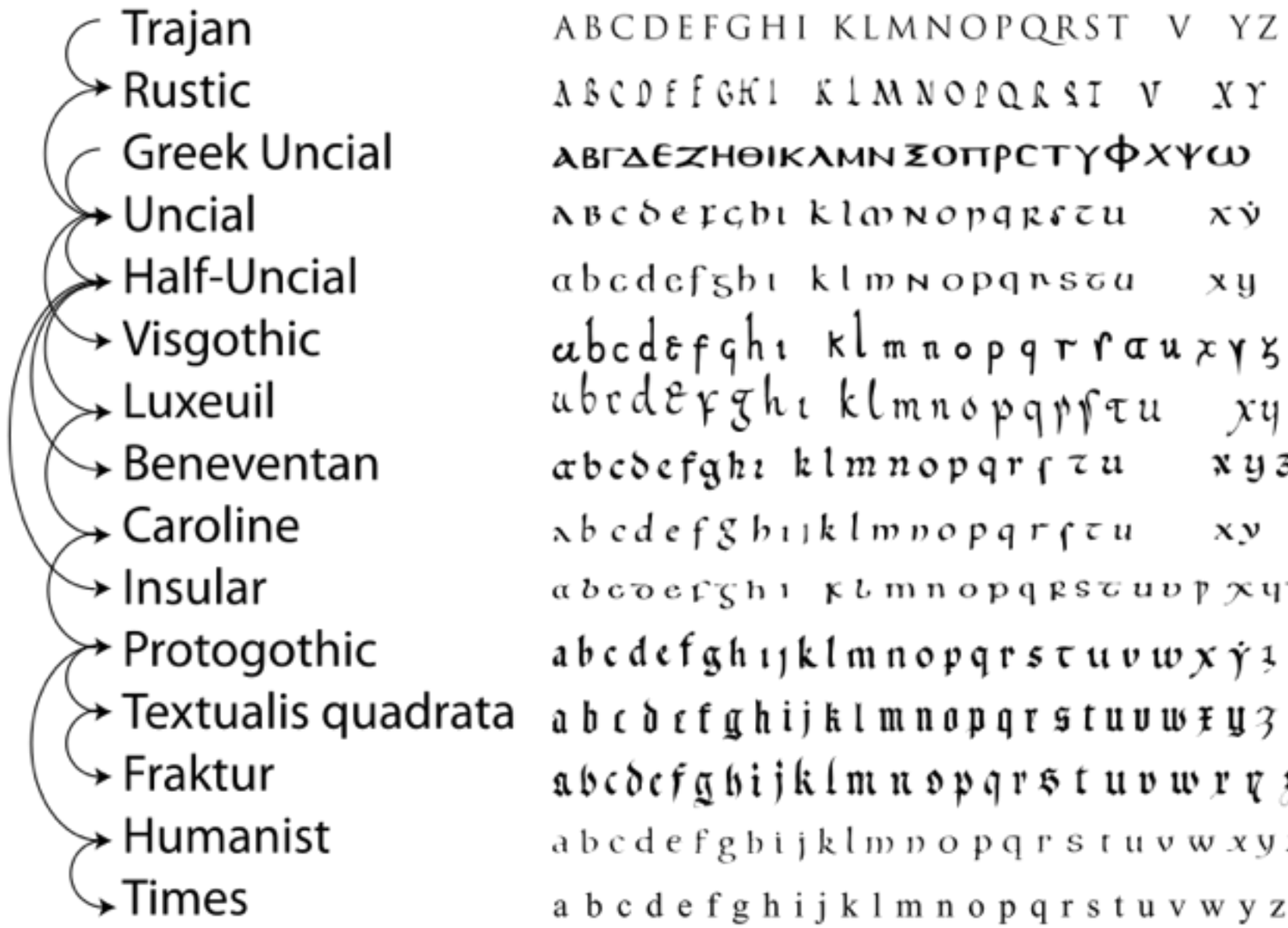
Code Change



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Code Change



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Examples

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Combination of Analyses

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

CTS/CITE Architecture

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

canonical text services

- stable identifier for canonical texts
- traditional citation != canonical citation
- sentence level !=/= canonical citation (verse, prose, prosimetric texts, religious texts…)
- what the digamma, is a canonical citation then?

CTS/CITE URNs

CTS URN

Namespace	Specific	CTS Namespace	TextGroup	Work	Version	Exemplar	Citation	Subreference
urn	:cts	:greekLit	:tlg0012	.tlg001	.msA	.thosJeff	:1.2	@οὐλομέν

urn:cite:CITENAMEPSACE:COLLECTION.OBJECTID

CITE URN → `urn:cite:hmt:msA.12r`

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Homer multitext project

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

HMT

Digital services from the Homer Multitext project
HMT: home | HMT Digital: home | browse MSS | scholia

HMT Digital: Scholia Reader

8 Scholia for 1.1

Manuscript urn:cts:greekLit:tlg0012.tlg001.msA:1.1

Homeric epic *Iliad* A

§1

1 Μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος

(urn:cts:greekLit:tlg0012.tlg001.msA:1.1)

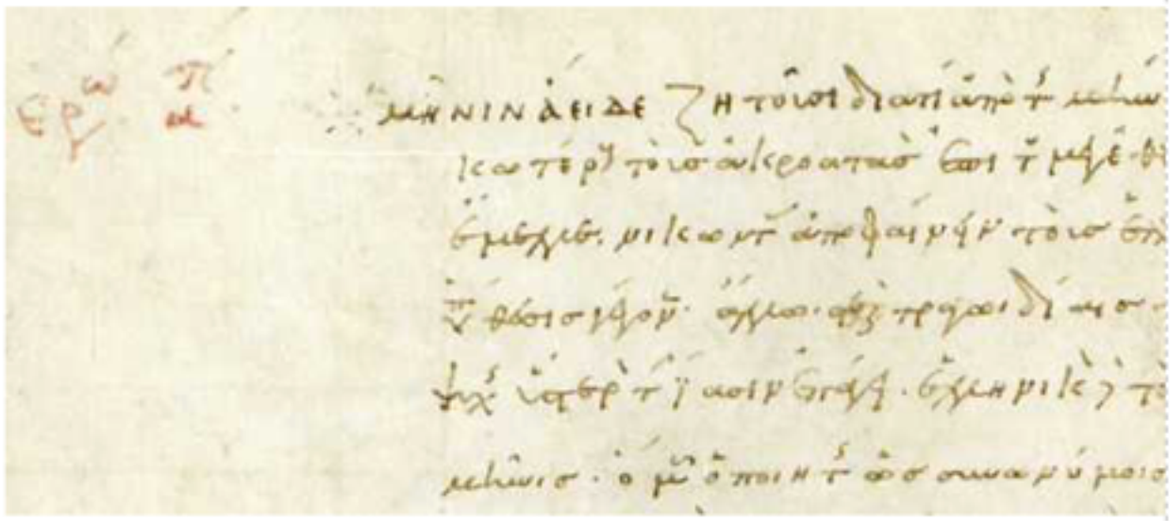
Scholia to the Iliad *Main marginal scholia* msA

§1

§1

μῆνιν ἄειδε

ζητοῦσι δια τί ἀπὸ τῆς μήνιδος ἤρξατο οὕτως δυσφήμου ὀνόματος· δια δύο ταῦτα· πρῶτον μὲν ἵν' ἐκ τοῦ πάθους ἀπκαταῤῥεύσῃ τὸ τοιοῦτο μόριον τῆς ψυχῆς καὶ προσεκτικωτέρους τοὺς ἀκροατὰς ἐπι τοῦ μεγέθους ποιήσῃ καὶ προσεθίζῃ φέρειν γενναίως ἡμᾶς τὰ πάθη. μέλλων πολλέμους ἀπαγγέλλειν· δεύτερον. ἵνα τὰ ἐγκώμια τῶν Ἑλλήνων παθανώτερα ποιήσῃ· ἐπεὶ δὲ ἔμελλε, νικωντας ἀποφαίνειν τοὺς Ἕλληνας, εἰκότως οὐ κατατρέχει ἀξιοπιστότερον ἐκ τοῦ μὴ παντα χαρίζεσθαι τῷ ἐκείνων ἐπαίνῳ· ἤρξατο μὲν ἀπὸ μήνιδος



Thomas Köntges
NLP of Historical Languages
June 2016

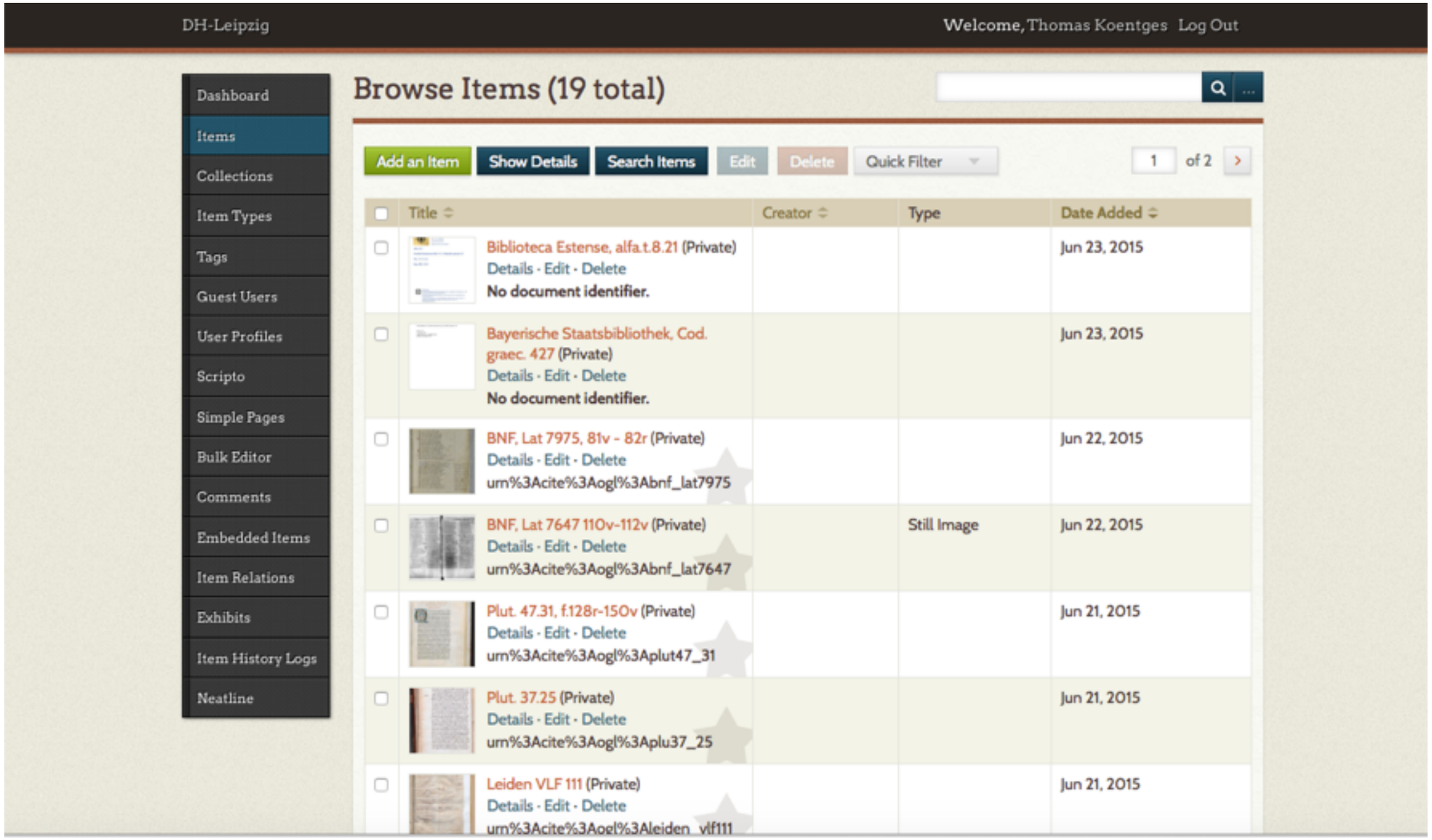
Digital Humanities
UNIVERSITÄT LEIPZIG

Manage Citations

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Manage the Content



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

CTS/CITE architecture:
find the logical structure of
a text and make it citable.

Textual Transmission and Machine Learning / Data Mining

Thomas Köntges
NLP of Historical Languages
June 2016

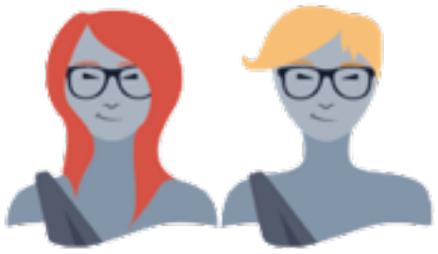
Digital Humanities
UNIVERSITÄT LEIPZIG

arXe.type Workflow

e-codices

SLUB
Wir führen Wissen.

Learners/Annotators



PERSEUS DIGITAL LIBRARY

Open Philology
UNIVERSITÄT LEIPZIG

Digitisation / Transcription

Transmission/ Reconstruction

Reading / Commenting

WEKA
The University of Waikato

Machine-Learning

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Optical Character Recognition / Transcription

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

OCR

booktif/0016/01000e.bin.png

ſtularemus miſit in faciem Aſcylti tunicá, & libe

ſtulatemus miſit in faciemu Aſelti tunicã, & libo

booktif/0016/01000f.bin.png

ratos querela iuſſit pallium deponere quod ſo-

tatos querela iuſſit pallium deponete quod ſo-

booktif/0016/010010.bin.png

lum litem faciebat, & recuperato vt putabamus

lumu fitem ffaciebat, & recupetato 1t putabamus

booktif/0016/010011.bin.png

theſauro in diuerſorium præcipites abimus, præ-

theſauro tn diuetſorium ptrcipites àbimus, præ-

booktif/0016/010012.bin.png

cluſisq; foribus ridere acumé non minus concio

cluſisqp foribus ridere acumẽ non minus concio

Thomas Köntges
NLP of Historical Languages
June 2016

OCR

Ὁ μὲν οὖν Αὔγουστος ἀπεβίω ὡς εἴρηται, τὴν PI545
δὲ μοναρχίαν ὁ Τιβέριος διεδέξατο. ὃς εὐπατρίδης C
μὲν ἦν καὶ πεπαίδευτο, τὴν δὲ γνώμην ἦν ποικιλώ- WII171
τατος, ἐναντίους τῇ προαιρέσει τοὺς λόγους ποιού-
μενος. ὧν γὰρ ἐβούλετο τἀναντία ἔλεγεν, ἄλλα μὲν
κεύθων ἐνὶ φρεσίν, ἄλλα δὲ λέγων· καὶ ὀργίζεσθαι
προσποιούμενος ἐν οἷς οὐκ ὠργίζετο, καὶ ἐν οἷς ἐθυ-
μοῦτο σχηματιζόμενος ἐπιείκειαν· καὶ ὡς οἰκειότατον
ἑώρα τὸν ἔχθιστον, καὶ ὡς ἀλλοτριωτάτῳ προσεφέ-
ρετο τῷ φιλτάτῳ. καὶ οὐκ ἠξίου τοῖς ἄλλοις δῆλον
εἶναί οἱ τὸ φρόνημα, τοῦτο προσήκειν τῷ αὐταρχοῦντι
φρονῶν. καὶ εἴτε τις ἠναντιοῦτο οἷς ἔλεγεν εἴτε μὴν D
καὶ συνήνει, μεμίσητο.

Τέως δ' οὖν εἰς τὰ στρατόπεδα καὶ εἰς τὰ ἔθνη
πάντα ὡς αὐτοκράτωρ αὐτίκα ἐπέστειλε, μὴ λέγων
αὐτοκράτωρ εἶναι· ψηφισθὲν γὰρ αὐτῷ καὶ τοῦτο
μετὰ τῶν ἄλλων ὀνομάτων οὐκ ἐδέξατο. καὶ τὰ τῆς
ἀρχῆς διοικῶν ἅπαντα, μηδὲν αὐτῆς δεῖσθαι ἔλεγε,
καὶ ταύτης ἐξίστασθαι ἐκομψεύετο καὶ διὰ τὴν ἡλι-
κίαν, ἓξ γὰρ καὶ πεντήκοντα ἐτῶν ἦν, καὶ δι' ἀμ-
βλυωπίαν· πλεῖστον γὰρ ἐν σκότει βλέπων ἐλάχιστα
τὴν ἡμέραν ἑώρα. εἶτα κοινωνοὺς ἤτει τῆς ἀρχῆς
καὶ συνάρχοντας, οὐδὲν τούτων ποιῆσαι μέλλων, ἀλλ' PI546

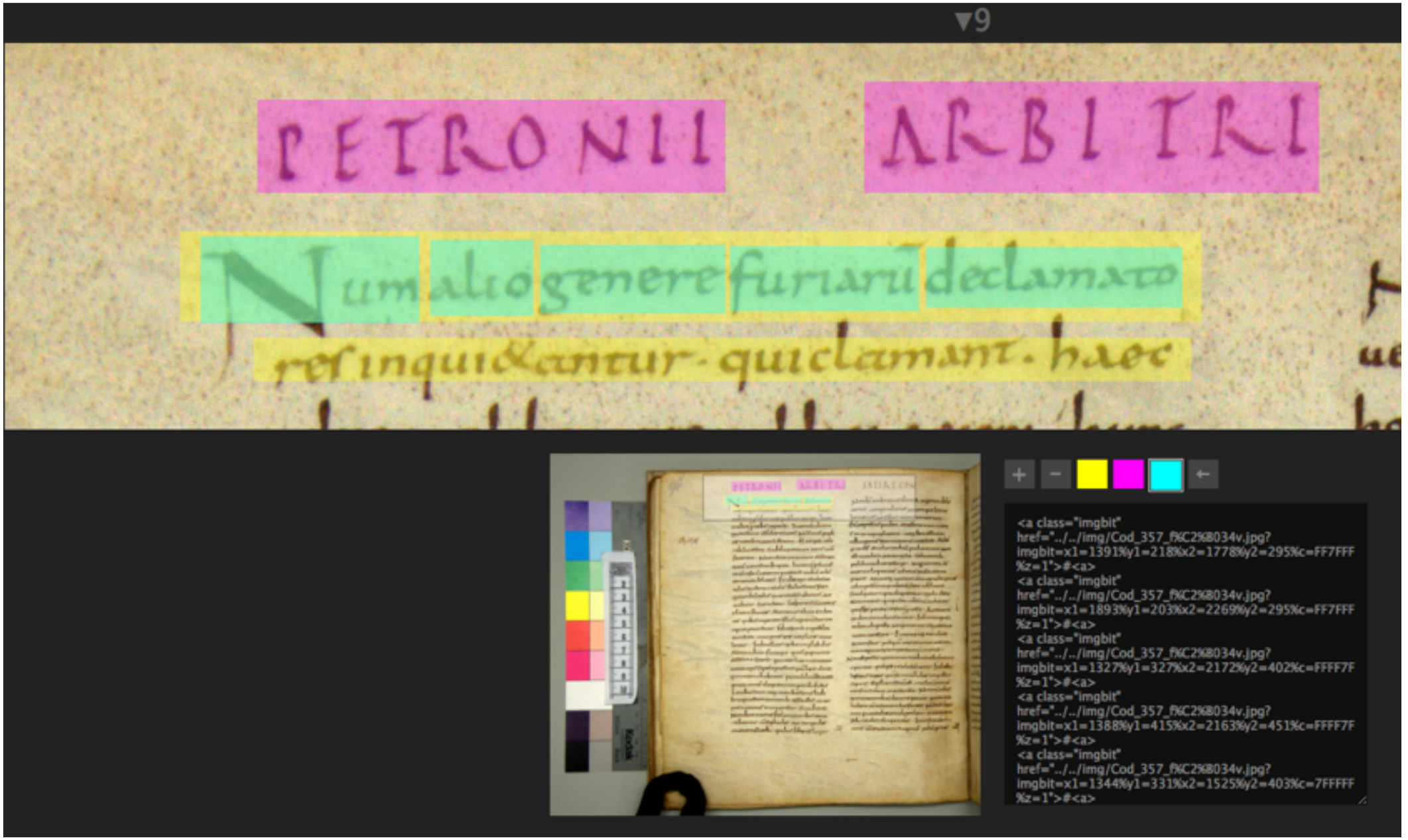
Cap. 1. Dionis Historiae Romanae l. 57, c. 1—13.

. PI515
D μὲν οὖν 7ύγουστος ἀπεβίω ὡς εἴρηται, τὴν,
δὲ μοναρχίαν ὁ Ιιβέριος διεδέξατο. ὅς εὐπατρίδης VW II171
μὲν ἦν καὶ πεπαίδευτο, τὴν δὲ γνώμην ἦν ποικιλώ-
τατος, ἐναντίους τῇ προαιρέσει τοὺς λόγους ποιού-
5 μενος. ν γὰρ ἐβούλετο τάναντία ἔλεγεν, ἄλλα μὲν
κεύθων ἐνὶ φρεσίν, ἄλλα δὲ λέγων· καὶ ὀργίζεσθαι
προσποιούμενος ἐν οἷς οὐκ ὠργίζετο, καὶ ἐν οἷς ἐθυ-
μούτο σχηματιζόμενος ἐπείκειαν· καὶ ὡς οἰκειότατον
ἔώρα τὸν ἔχθιστυPa,PPMé ὡς ἀλλοτριωτάτῳ προσεφέ-
10 ρετο τῷ φιλτάτῳ. καὶ ούa ἠξίου τοῖς ἄλλοις δῆλον
εἶναί οἱ τὸ φρόνημα, τοῦτο προσήκειν τῷ αὐταρχοῦντι
φρονῶν. καὶ εἴτε τις ἠναντιοῦτο οἷς ἔλεγεν εἴτεμὴνI)
καὶ συνήνει, μεμίόητο.
ἑαως δ' οὐν εἰς τὰ στρατόπεδα καὶ εἰς τὰ ἔθνη
15 πάντα ὡς αὐτοκράτωρ αὐτίκα ἐπέστειλε, μὴ λέγων
αὐτοκράτωρ εἶναι' ψηφισθὲν γὰρ αὐτῷ καὶ τοῦτο
μετὰ τῶν ἄλλων ὀνομάτων οὐκ ἐδέξατο. καὶ τὰ τῆς
ἀρχῆς διοικῶν ἅπαντα, μηδὲν αὐτῆς δεῖσθαι ἔλεγε,
καὶ ταύτης ἐξίστασθαι ἐκομψεύετο καὶ διὰ τὴν ἡλι-
2ο κίαν, ἔΗ γὰρ καὶ πεντήκοντα ἐτῶν ἦν, καὶ δι' ἀμ-
βλυωπίαν· πλειστον γὰρ ἐν σκότει βλέπων ἐλάχιστα
τὴν ἡμέραν ἔώρα. εἶτα κοινωνοὺς ἤτετ τῆς ἀρχῆς
καὶ συνάρχοντας, οὐδὲν τούτων ποιῆσαι μέλλων, ἀλλ' PI516
Cap. 1. Dionis Historiae Romanae 1. 57, c. 1-13.
eoNARA8 1II,

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Transcription

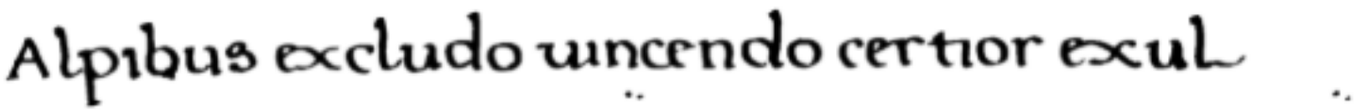


Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

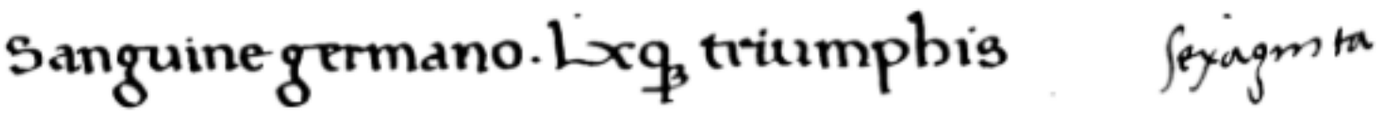
Transcription

book/0001/010001.bin.png



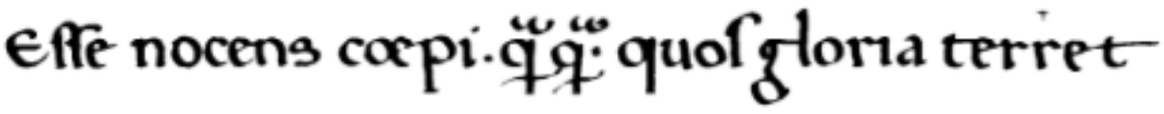
Alpibus excludo uincendo certior exul

book/0001/010002.bin.png



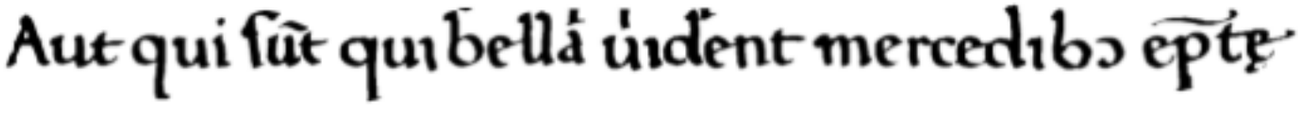
Sanguine germano. Lxq3 triumphis

book/0001/010003.bin.png



esse nocens coepi. q^{w}q^{w} quos gloria terret

book/0001/010004.bin.png



Aut qui sũt qui bellá úident mercedib9 epte

Aut qui sũt qui bellá úident mercedib^{9} epte

Thomas Köntges
NLP of Historical Languages
June 2016

Weka

- Weka 3.7.11: open source data mining software in Java
- a collection of machine-learning algorithms for data-mining tasks
- although Weka is a powerful tool, classifying the manuscripts will be a challenge

Weka

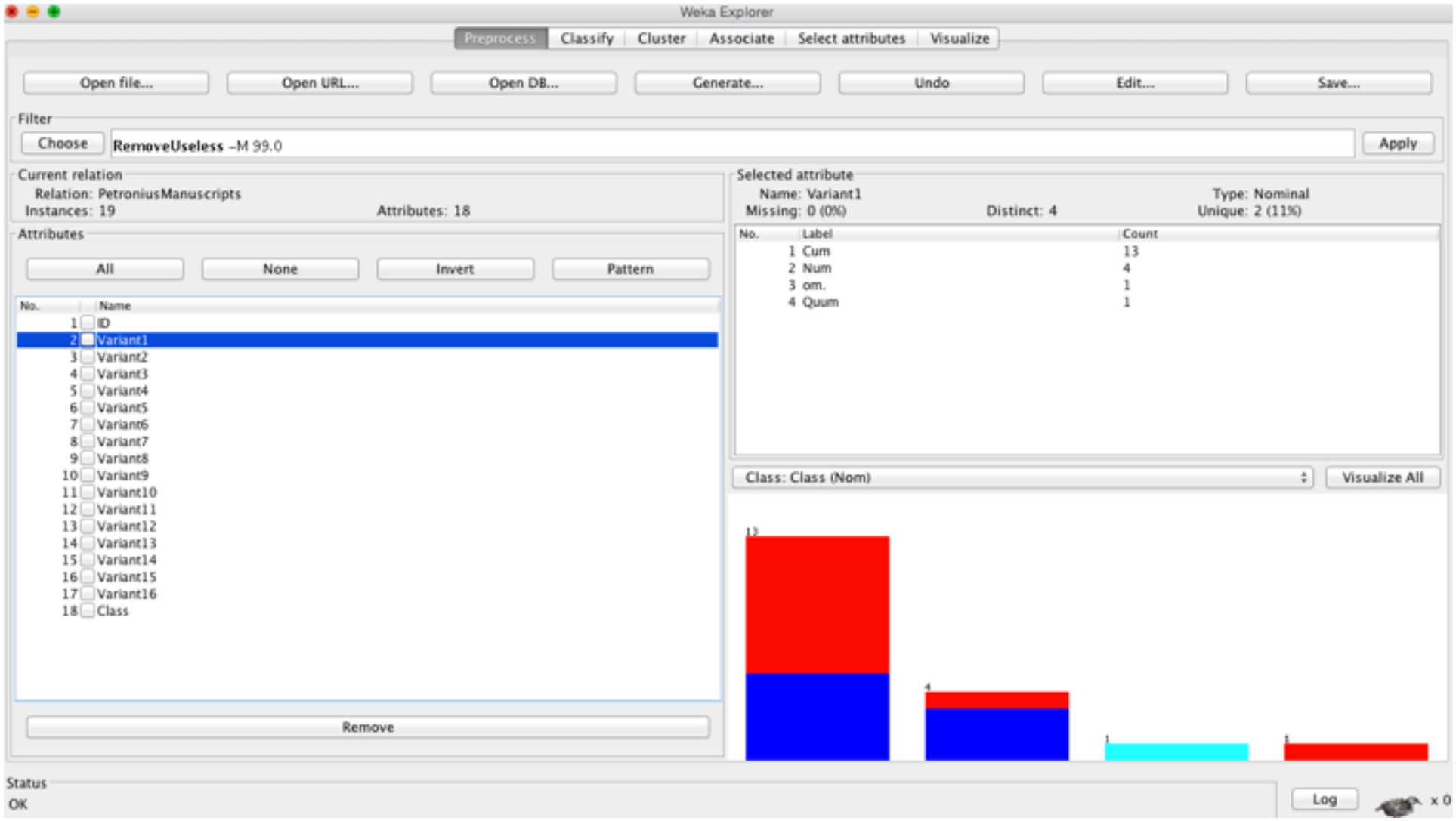
@relation PetroniusManuscripts

@attribute ID numeric
@attribute Variant1 {Cum,Num,om.,Quum}
@attribute Variant2 {Nec,Non,om.}
@attribute Variant3 {Essent,om.}
@attribute Variant4 {ituris,nuris,om.,turis}
@attribute Variant5 {facent,facerent,om.}
@attribute Variant6 {rerum,runici,verborum,om.}
@attribute Variant7 {et,om.}
@attribute Variant8 {strepita,strepitum,om.}
@attribute Variant9 {terrarum_orbem,orbem_terrarum,om.}
@attribute Variant10 {qui,quia,quorum,om.}
@attribute Variant11 {his,iis}
@attribute Variant12 {et,om.,sed}
@attribute Variant13 {et,om.,sed}
@attribute Variant14 {pestilentia,pestilentiam,om.}
@attribute Variant15 {omnem,omnium,om.}
@attribute Variant16 {quidem,om.,solum}
@attribute Class {L,O}

@data

1,Num,Non,Essent,ituris,facerent,rerum,et,strepitum,orbem_terrarum,quia,his,et,sed,pestilentiam,omnium,quidem,O
2,Num,Non,Essent,ituris,facerent,rerum,et,strepitum,terrarum_orbem,quia,iis,et,sed,pestilentiam,omnium,quidem,L
3,Num,Non,Essent,ituris,facerent,rerum,et,strepitum,terrarum_orbem,quia,iis,sed,sed,pestilentiam,omnium,quidem,L

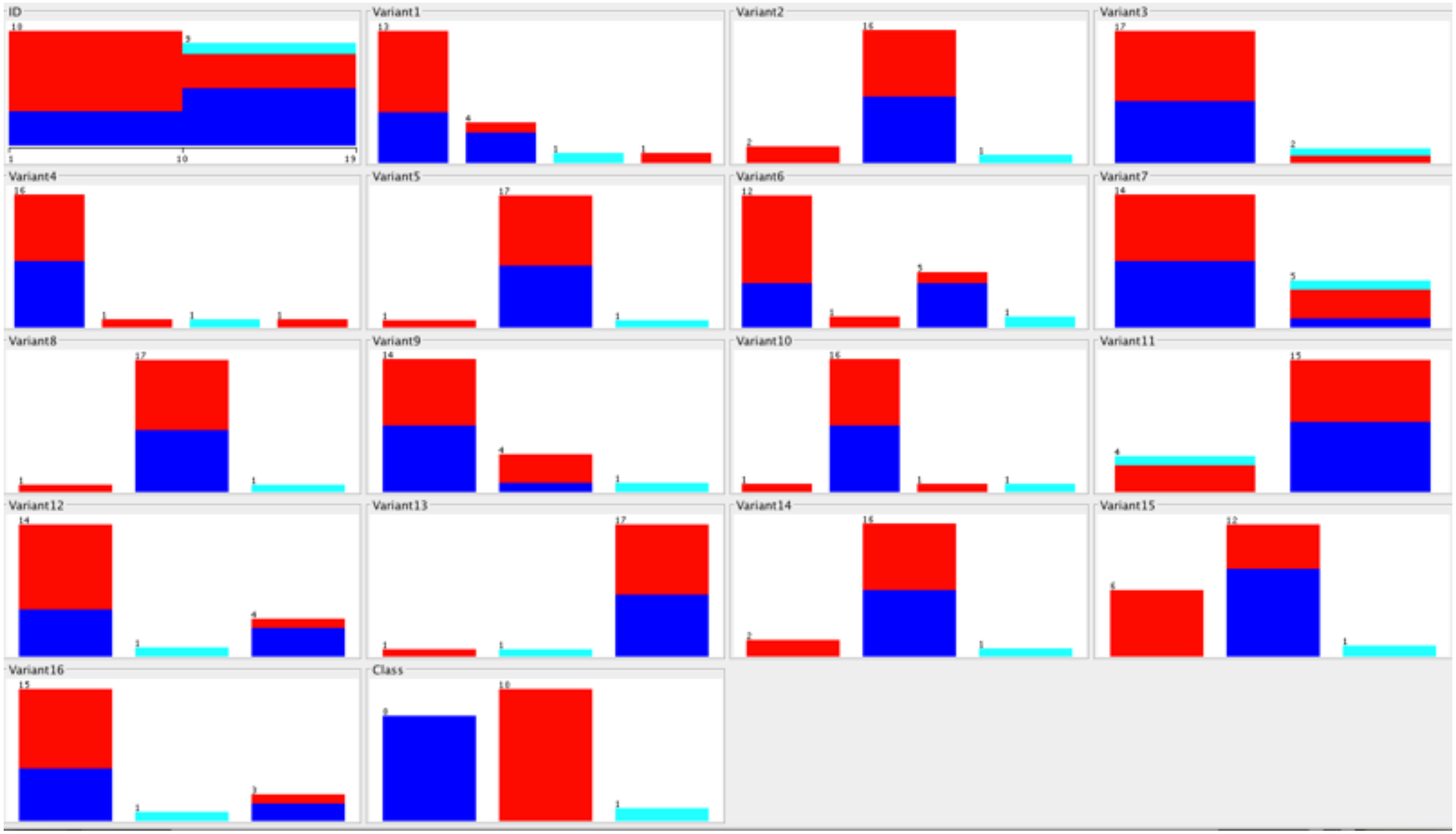
Weka



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

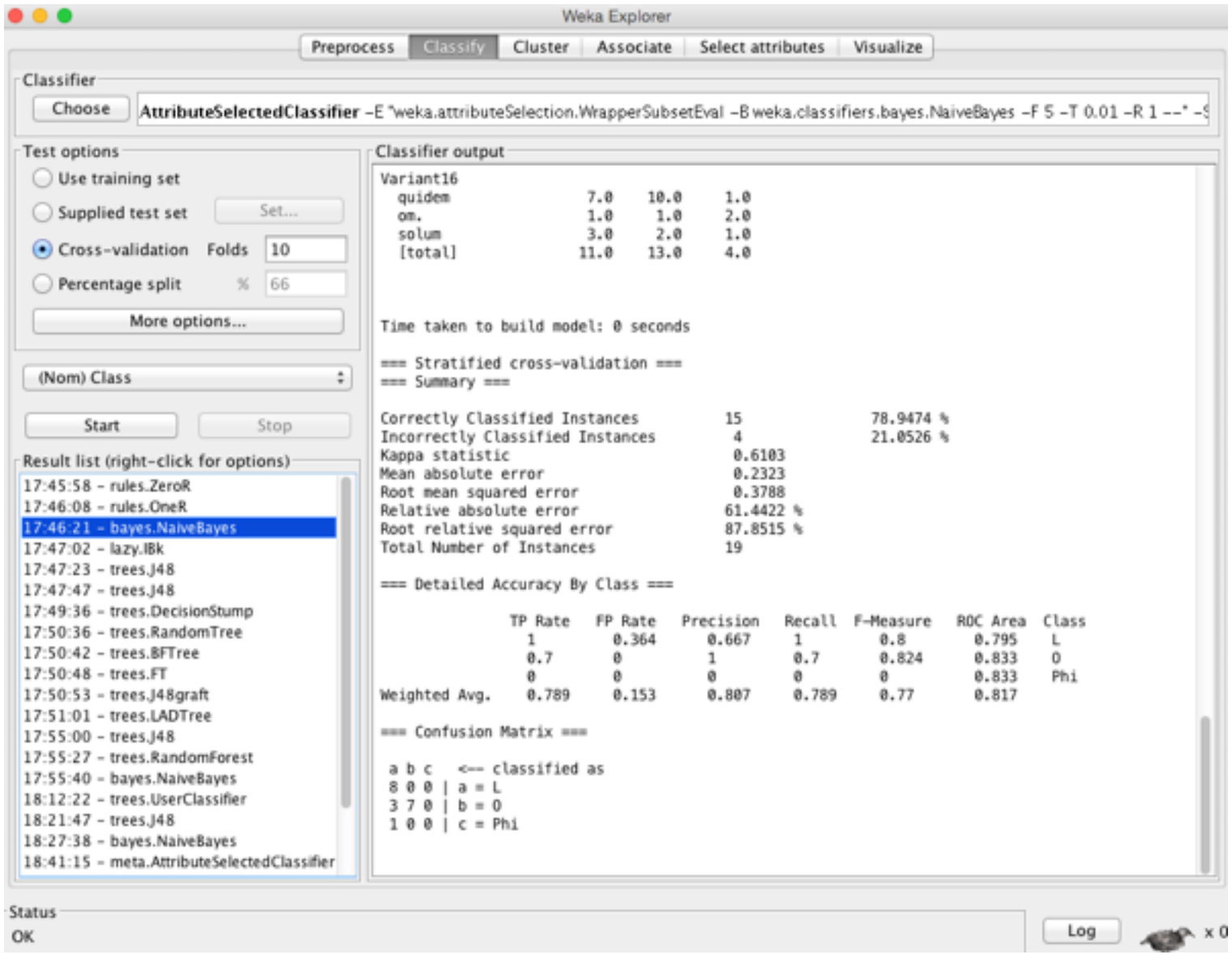
Weka



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Weka



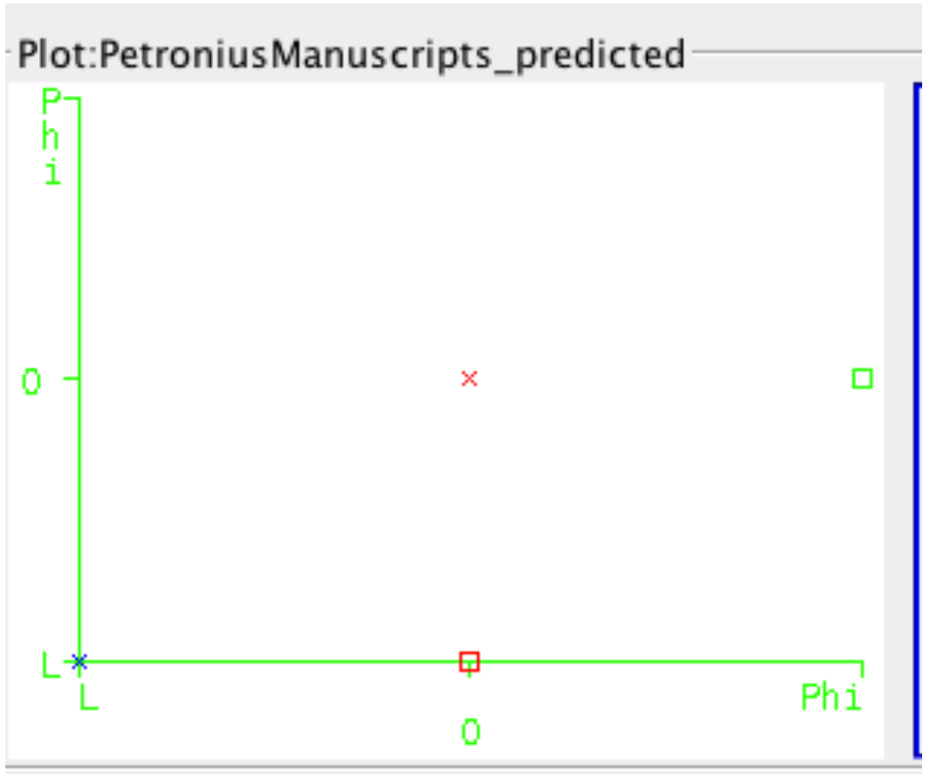
Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

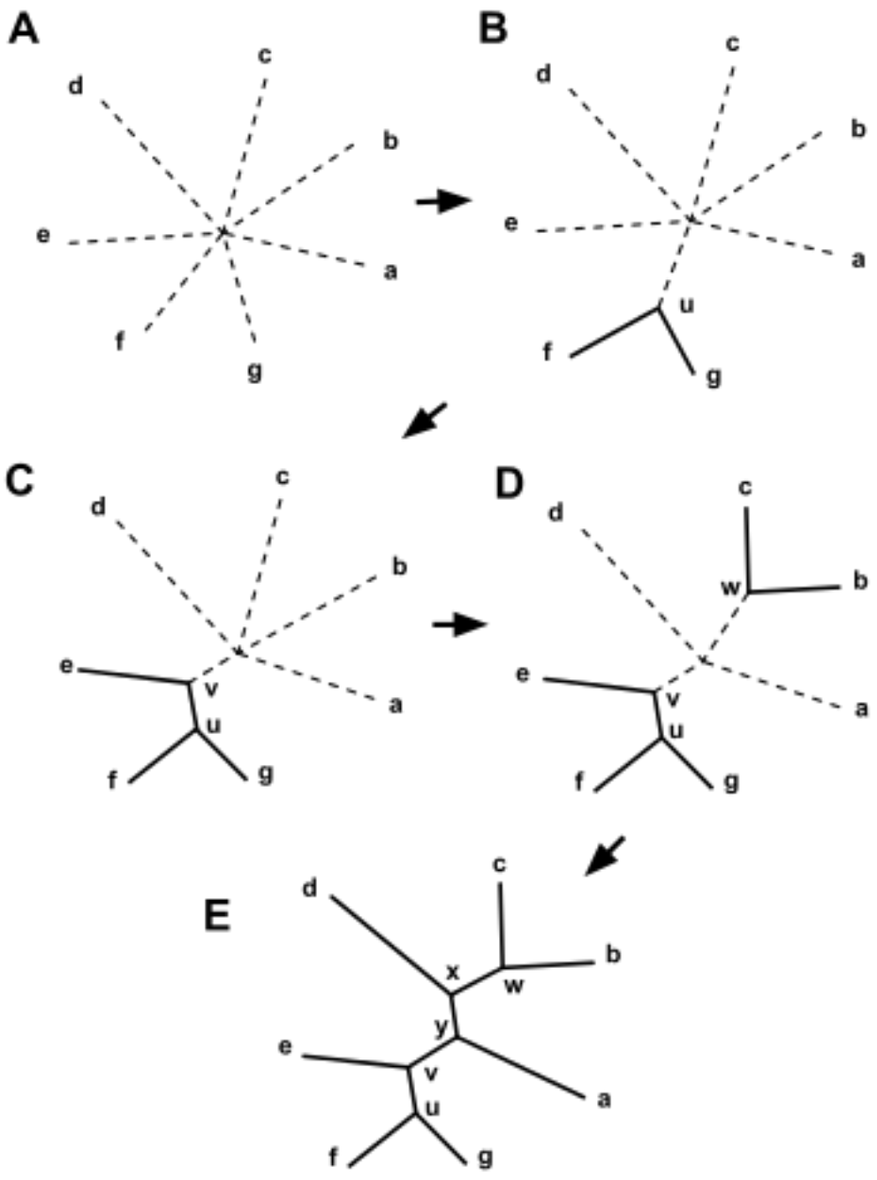
Weka

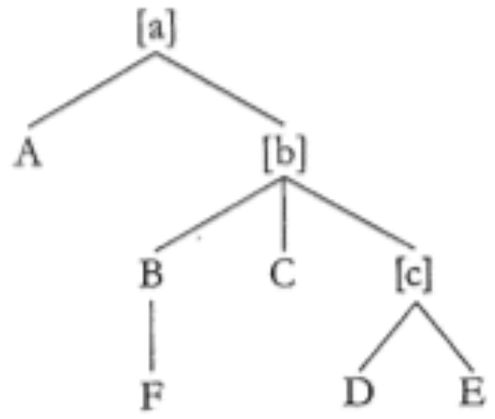
δ = codex a Poggio Florentino anno 1420 repertus, nunc deperditus, ex quo pendent codices novicii qui omnes saeculo XV scripti sunt

- α
  - A = codicis Parisini lat. 7989 (Orellii *C*, Beck *Trag. pars prior*) olim Traguriensis excerpta vulgaria, olim Cipici, cf. H
  - I = codex Indianensis Notre Dame 58
  - F = codex Leidensis Vossianus 265 O.81 (Beck *Vb*)
  - K = codex Vaticanus lat. 1671
- ζ
  - J = codex Florentinus Laurentianus 37,25 (Beck *F1*)
  - V = codex Vindobonensis 179, olim 235 (Endlicheri 218; Beck *V1*))
  - W = codex Vindobonensis 3198 (Endlicheri 108), olim Ioannis Sambuci (Beck *V2*, BÜCHELER[1] *S*)
  - C = codex Vaticanus Urbinas 670 (Beck *Vat.*)
  - D = codex Florentinus Laurentianus 47,31 (Beck *F2*)
  - G = codex Guelferbytanus extravag. 299
  - Q = codex Vaticanus lat. 3403



Thomas Köntges
NLP of Historical Languages
June 2016





Suppose we have species W, X, Y and Z, with the following sequences (in reality, much longer sequences would be used):

W	AAAAAAAA
X	GGAAAAAA
Y	CCTTTTAA
Z	CCTTCCAA

The distance matrix would be

	W	X	Y	Z
W	–	2	6	6
X	–	–	6	6
Y	–	–	–	2

and the tree inferred would be:

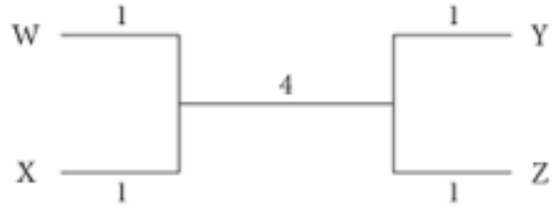


Figure 1. Hypothetical example of a distance matrix analysis.

Stemmaweb (T. Andrews)

Stemmaweb - a collection of tools for analysis of collated texts

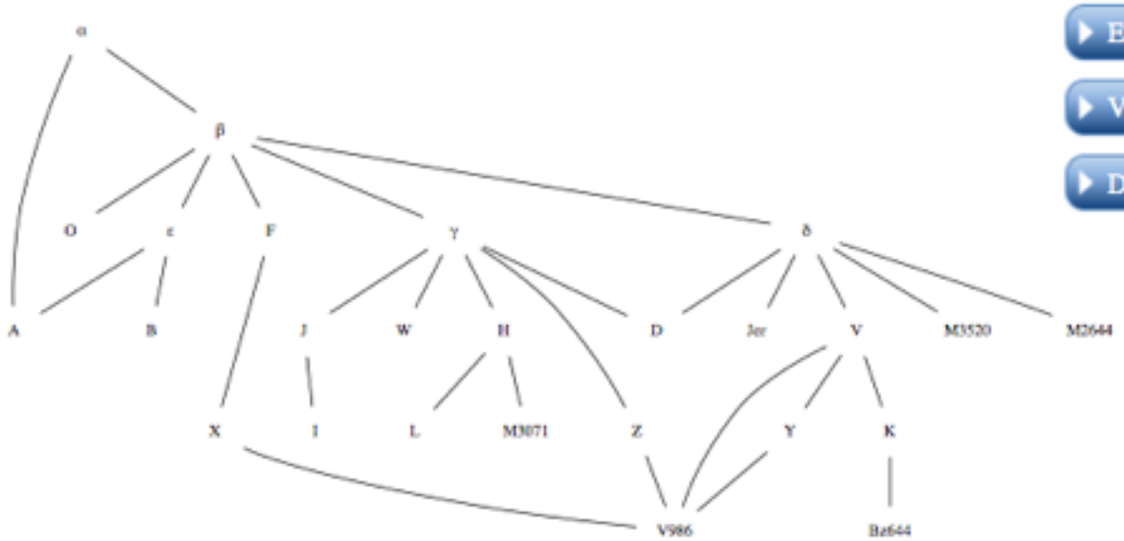
Text directory

Public text traditions (read-only)

- Sapientia
- Parzival 249-255
- Hanc concordi
- Quaestiones ad Antiochum ducem (partially)
- *Chronicle of Matthew*
- Notre besoin artificial
- Parzival artificial
- Florilegium Coislinianum B

Text Chronicle of Matthew

- is owned by no one
- is public
- has Armenian as its primary language
- has witnesses: F, J, C, G, Jer, V, A, O, K, D, Y, E, W, L, Z, H, I, B, X



Examine variants against this stemma

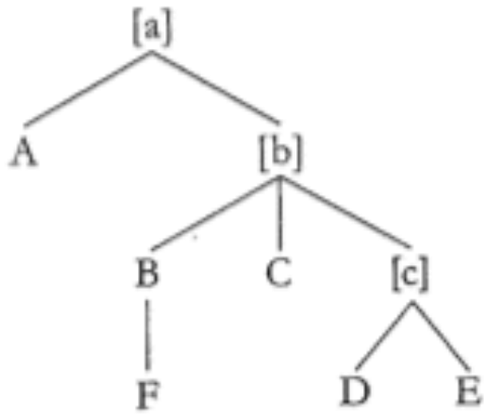
View collation and relationships

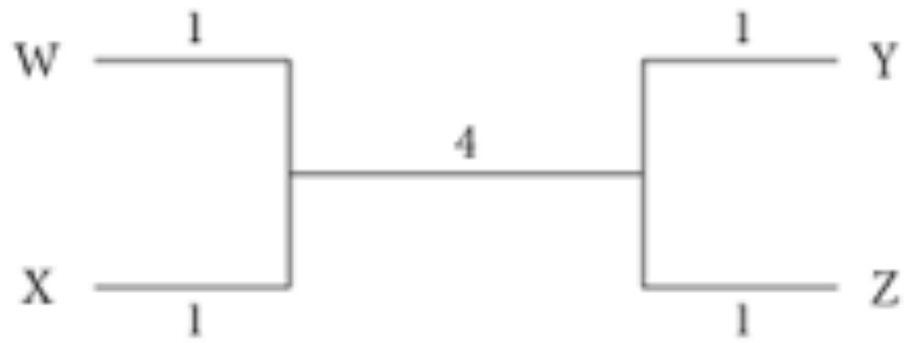
Download tradition

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Scoring Variants





- mathematical "scoring" with the premise that we have full information (can be weighted)

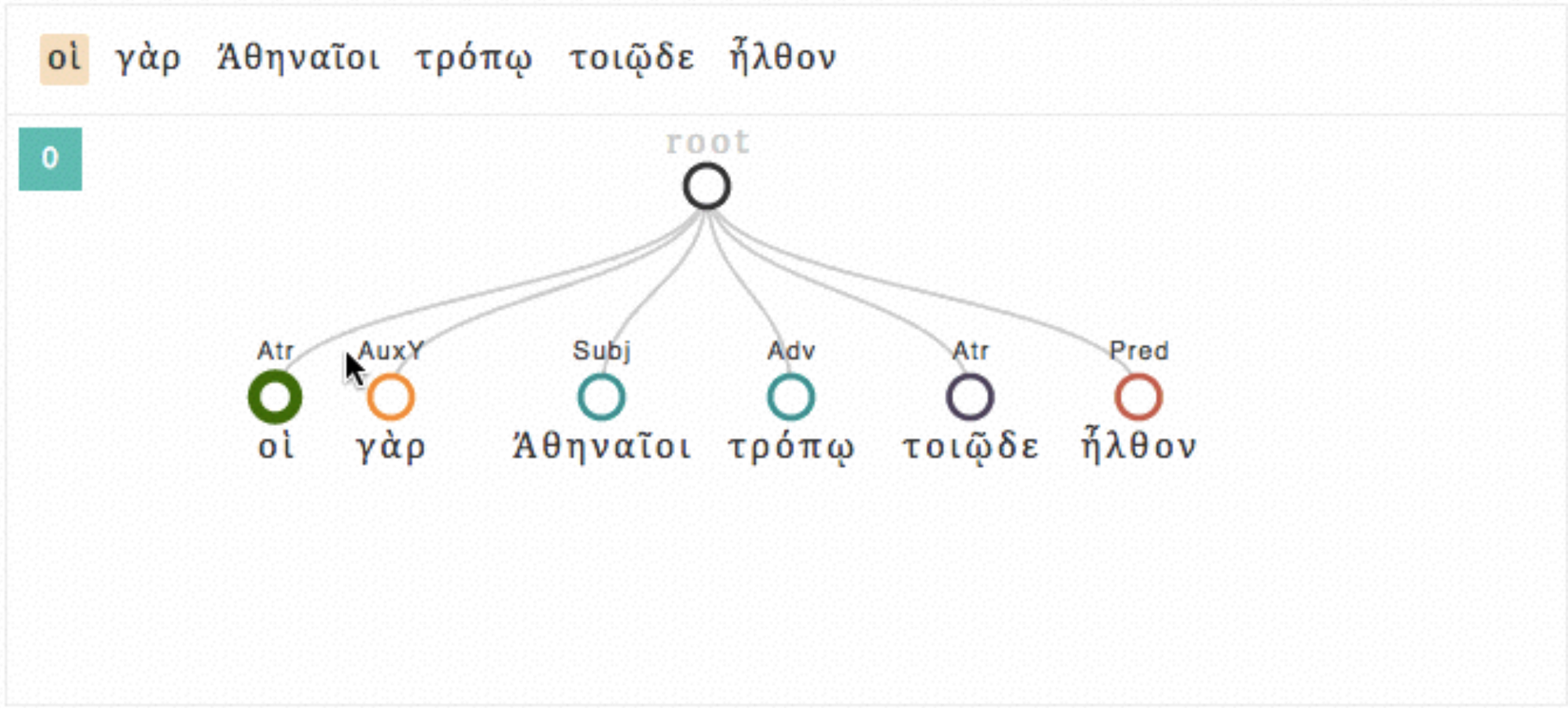
- philological "scoring": correct/false grammar; meaning; similarity; context -> explanation of the variant, sets the weighting

Tree Banking
&
Alignment

Thomas Köntges
NLP of Historical Languages
June 2016

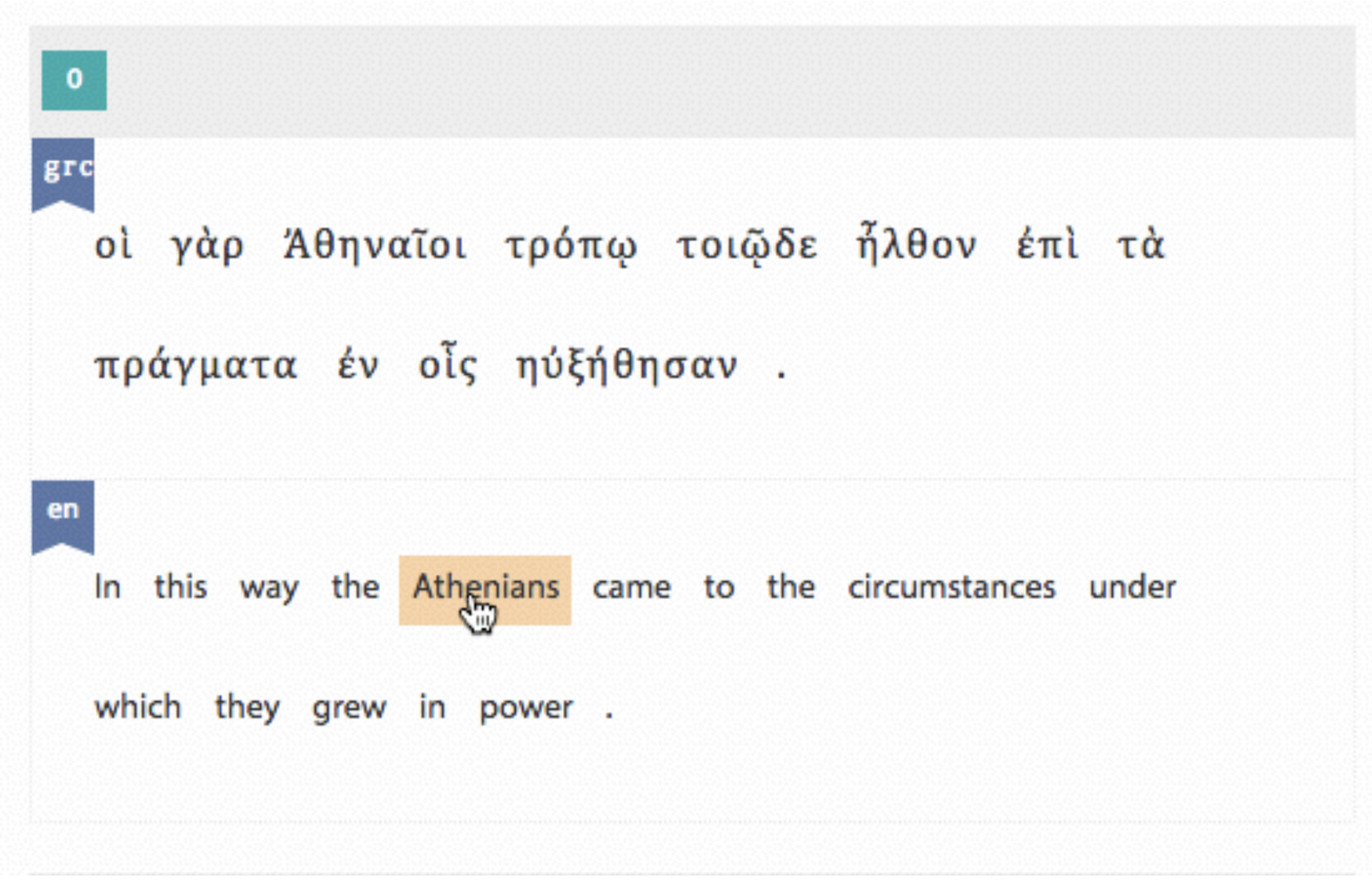
Digital Humanities
UNIVERSITÄT LEIPZIG

Tree Banking



Thomas Köntges
NLP of Historical Languages
June 2016

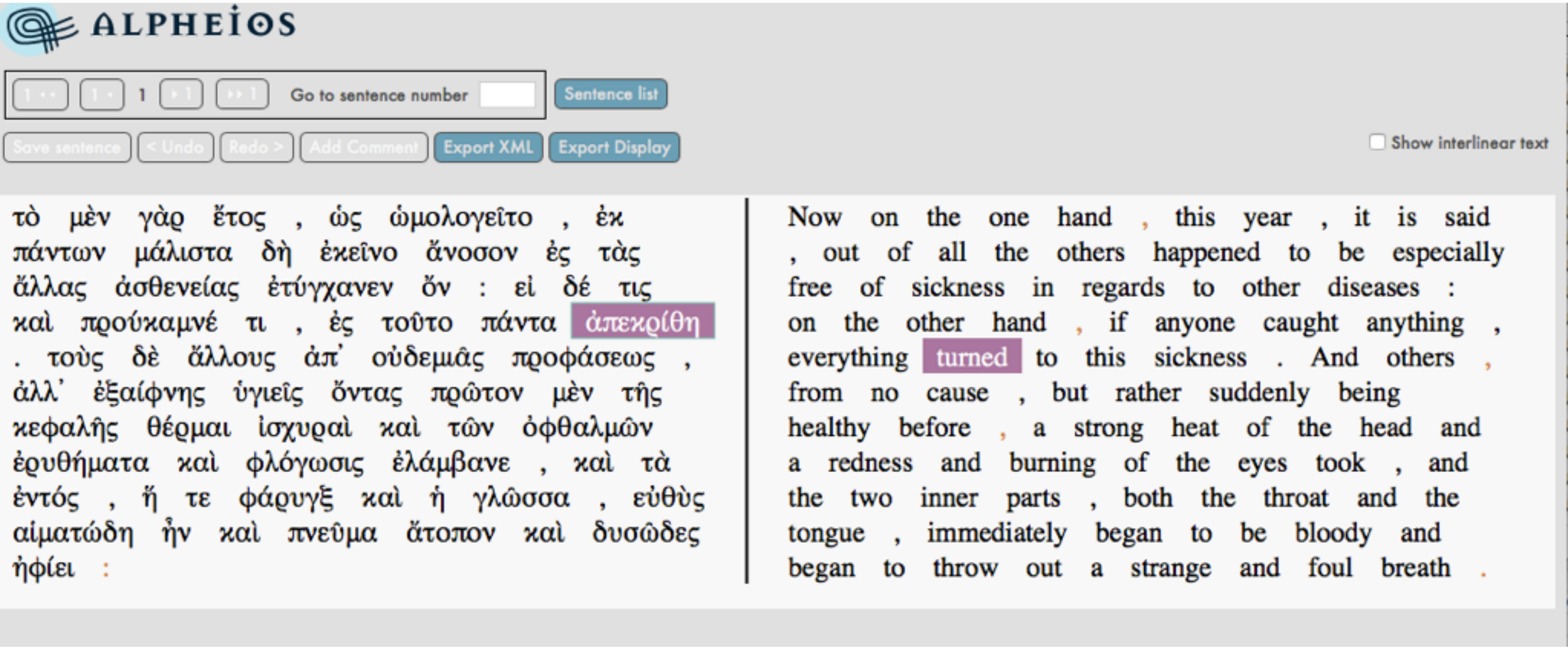
Translation Alignment



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

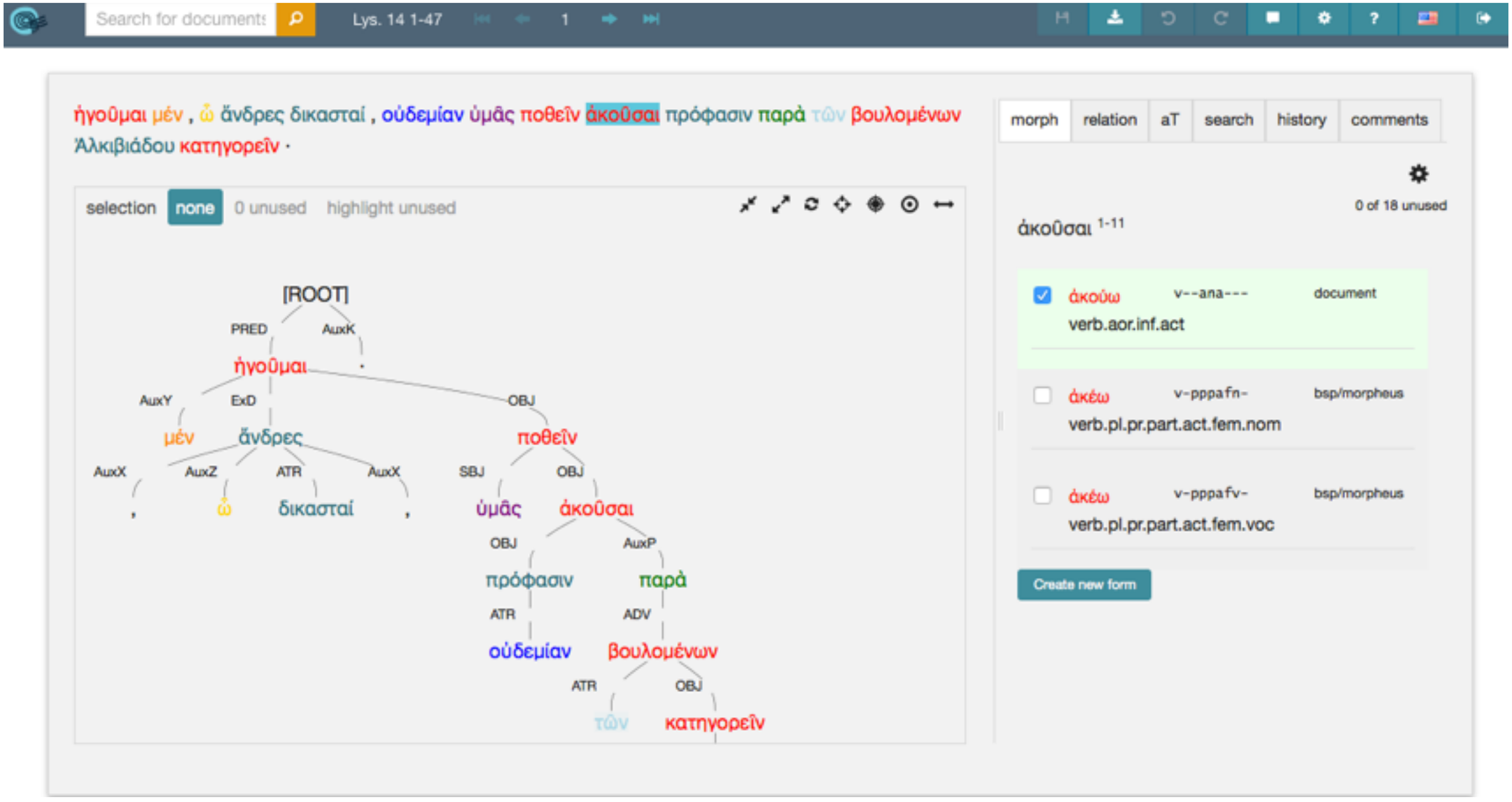
Gradually More Complex



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Gradually More Complex



Thomas Köntges
NLP of Historical Languages
June 2016

Morphological Parsers

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Morpheus API

```
– <dict>
    <hdwd xml:lang="lat">servo</hdwd>
    <pofs order="1">verb</pofs>
  </dict>
– <infl>
  – <term xml:lang="lat">
      <stem>serv</stem>
      <suff>a_</suff>
    </term>
    <pofs order="1">verb</pofs>
    <mood>imperative</mood>
    <num>singular</num>
    <pers>2nd</pers>
    <tense>present</tense>
    <voice>active</voice>
    <stemtype>conj1</stemtype>
    <derivtype>are_vb</derivtype>
</infl>
```

Topic Modelling

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Bag-Of-Words

- Simplification of texts
- Zellig Harris (1954)
- Word order and/or grammar does not matter

Topic Modelling

- Collective knowledge continues to grow; it becomes more difficult to find what one is looking for

- Topic Models are more than a search&link-approach

- Zooming in and out is possible with TMs

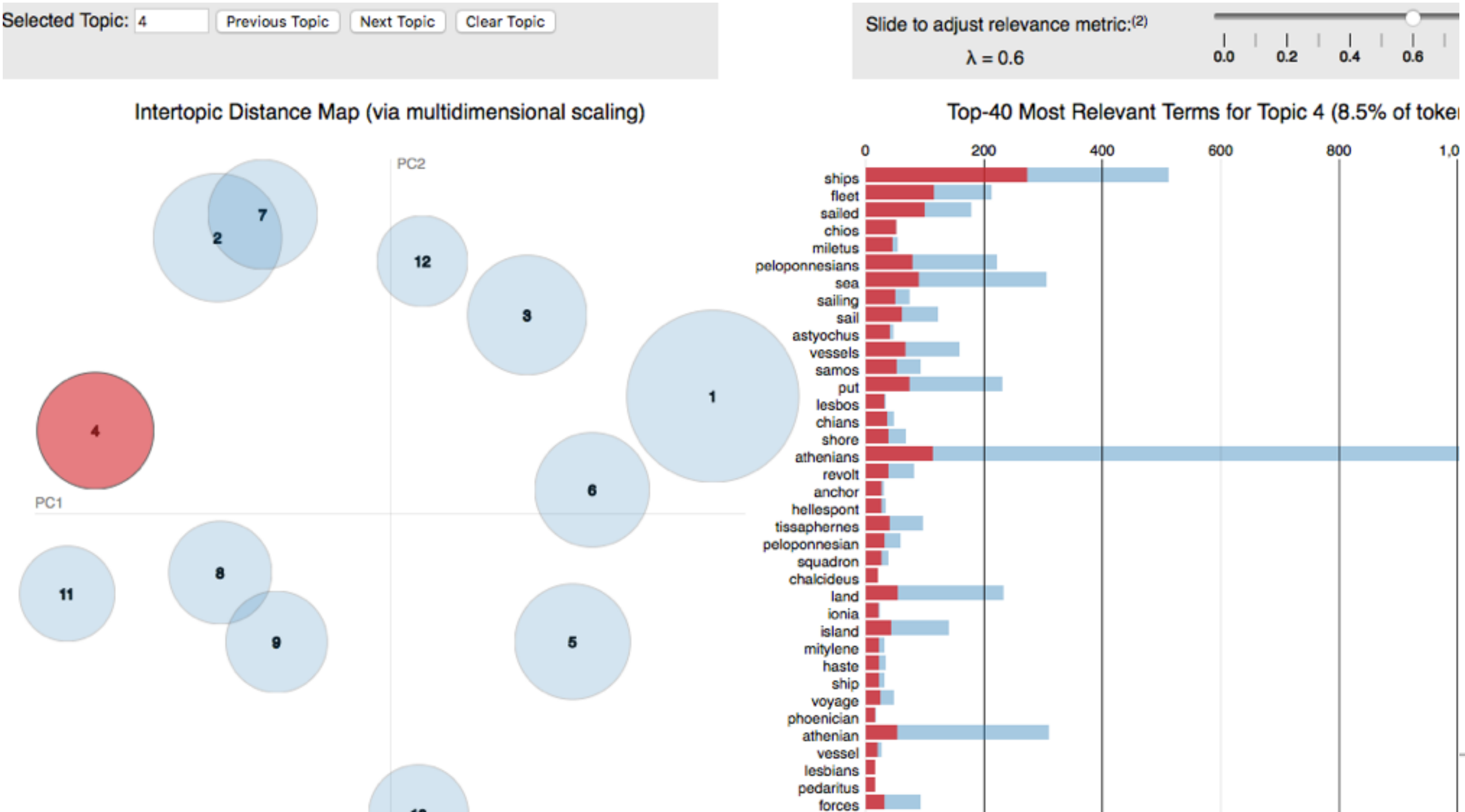
- Topic Model algorithms do not require prior annotation or labelling of the texts

- Topic Models discover the hidden thematic structure in large archives of documents

Topic Modelling

- A method to find clusters of words in large bodies of text
- Those clusters are called topics
- A topic is a recurring pattern of co-occurring words
- Topic models are probabilistic models that are often based on the number of topics in the corpus being assumed and fixed

Topic Modelling



More (better) Data and
SIMPLE Algorithms
Often BEAT
COMPLEX Analytics Models

“We don’t have better algorithms. We just have more data.”

– Peter Norvig (Google)

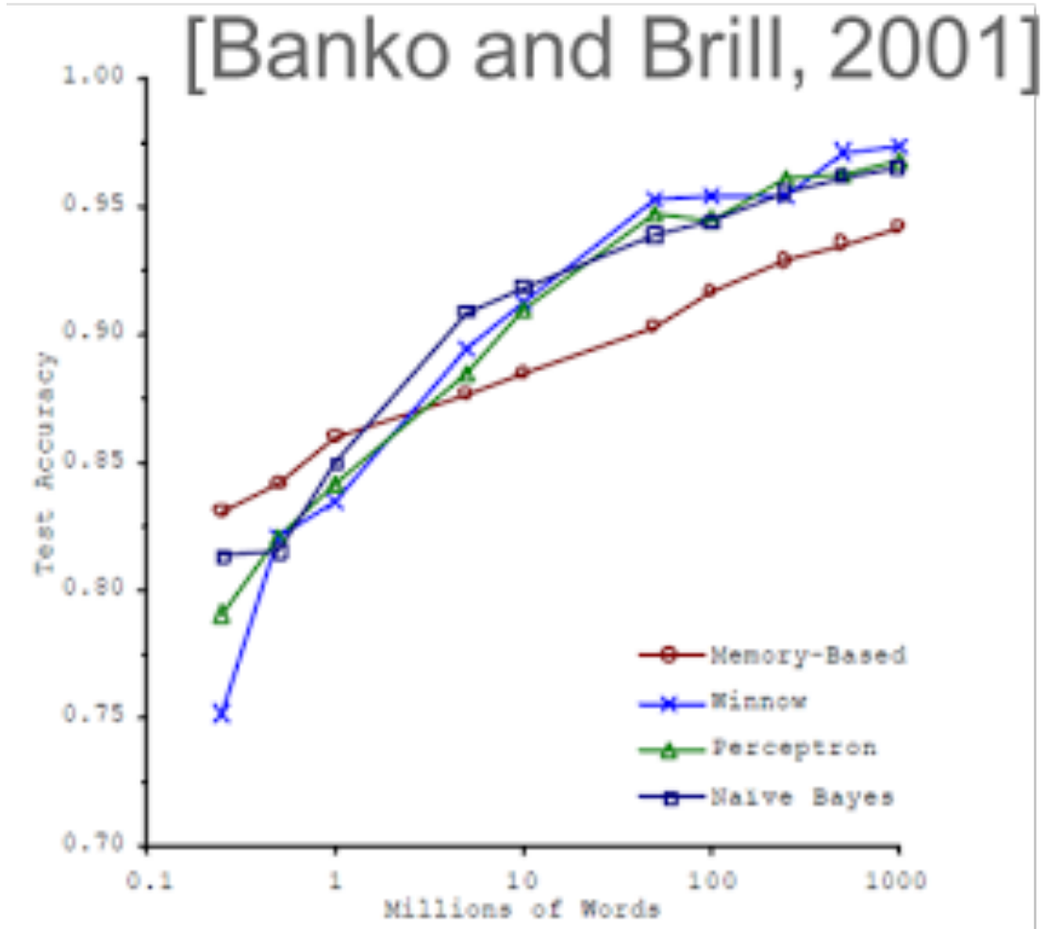


Figure 1. Learning Curves for Confusion Set Disambiguation

Latent-Dirichlet-Allocation (LDA)

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Statistical Inference

Deducing properties of
an underlying distribution
by analysis of data.

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

LDA

Assumes that one can find a Dirichlet Distribution by analysing the words in a corpus.

Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Dirichlet Distribution

Is a probability distribution
over all possible
multinominal distributions

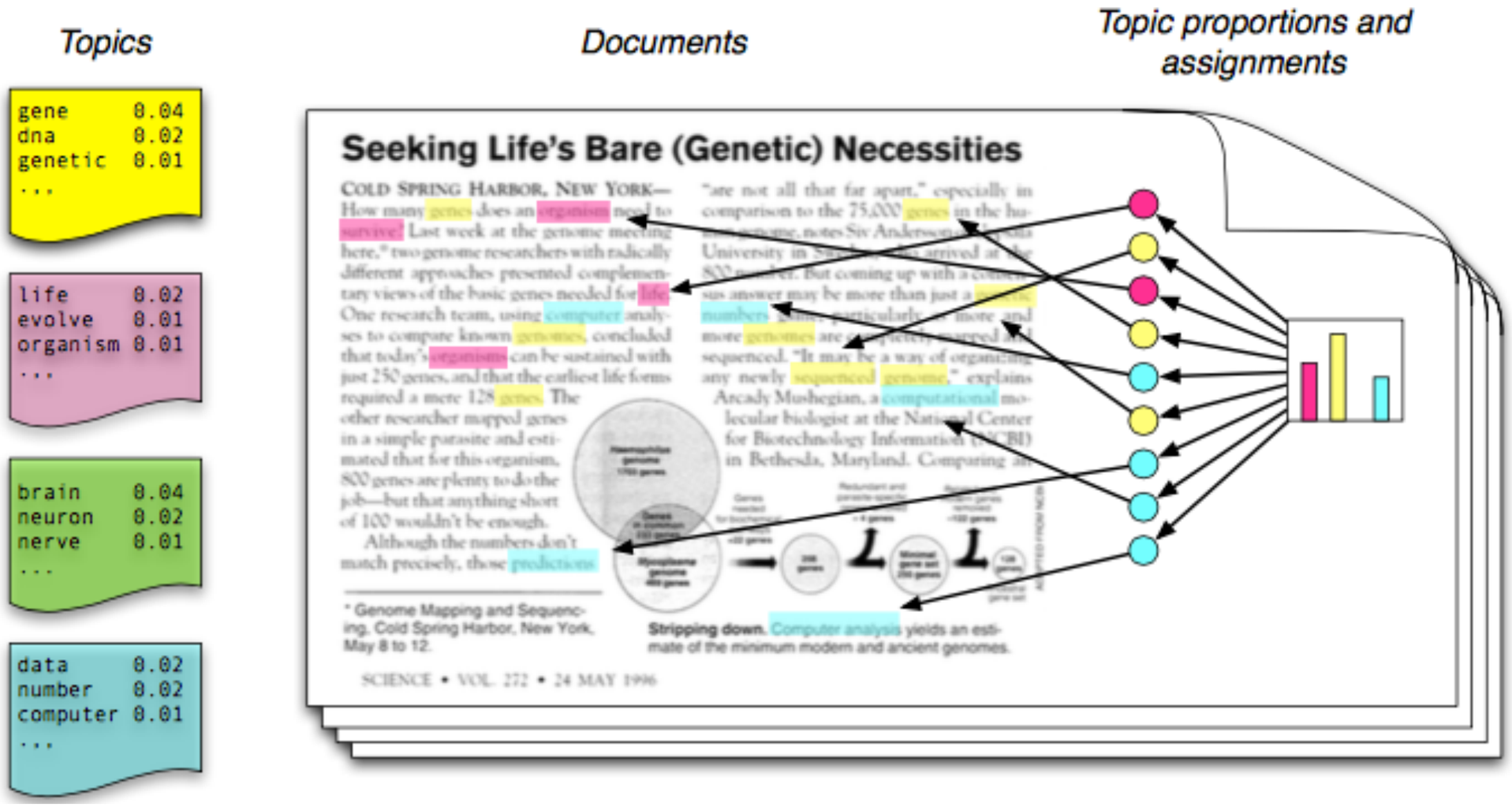
Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Dirichlet Distribution of Words in Corpus

1. For each document draw a topic distribution
2. For each word in the document
   a. Draw a specific topic
   b. Draw a word

Dirichlet Distribution of Words in Corpus



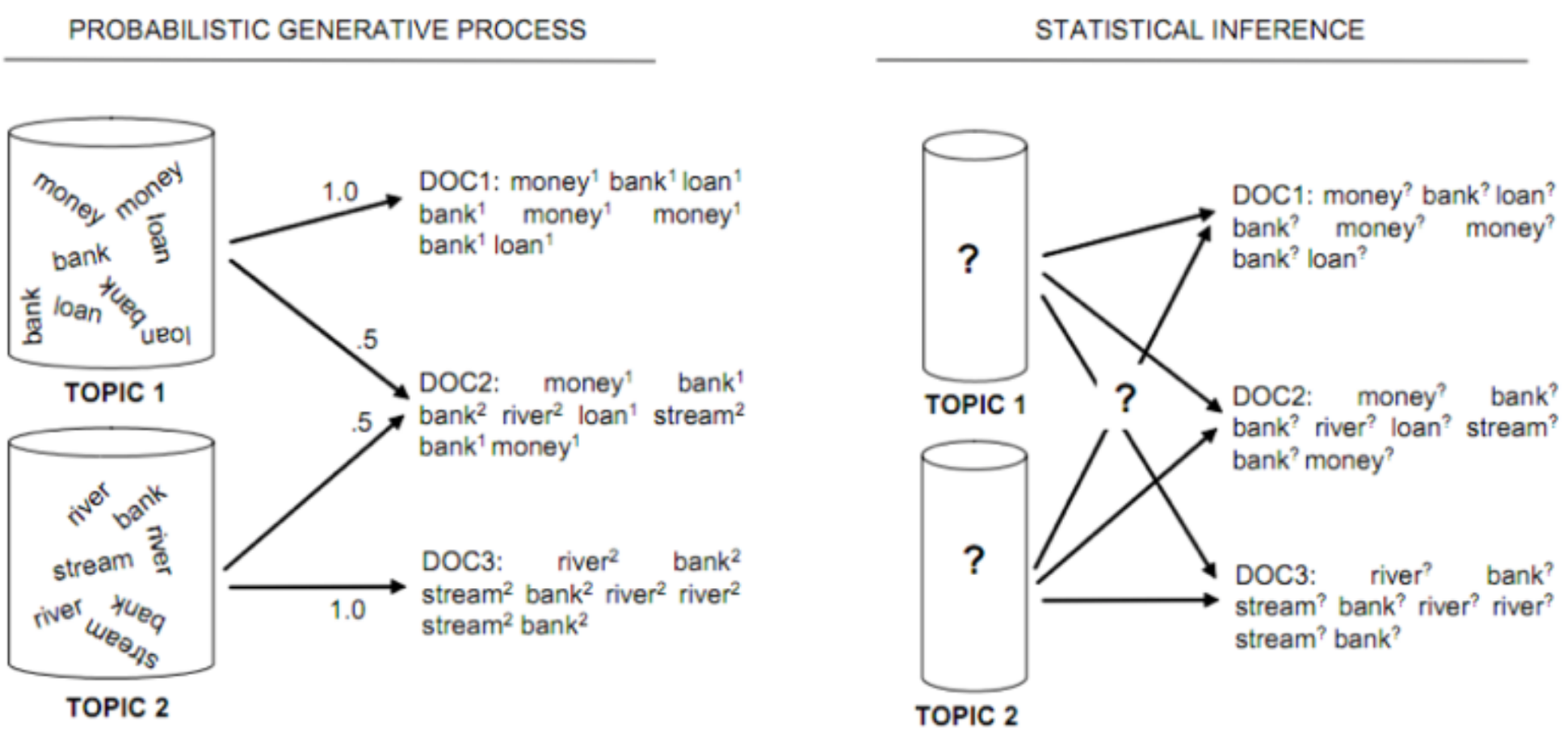
Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Latent Dirichlet Allocation

- Simplification of how the documents in a dataset were created using Dirichlet Distribution
- Documents are Bag-of-words (order of the words in each document is irrelevant)
- Each corpus has words from a number of known topics
- We do not know which words belong to which topic
- “panel” in document A and “panel” in document B are not the same “word”

Latent Dirichlet Allocation



Thomas Köntges
NLP of Historical Languages
June 2016

Digital Humanities
UNIVERSITÄT LEIPZIG

Latent Dirichlet Allocation

- The central goal of topic modelling is to automatically discover the topics from a collection of documents.

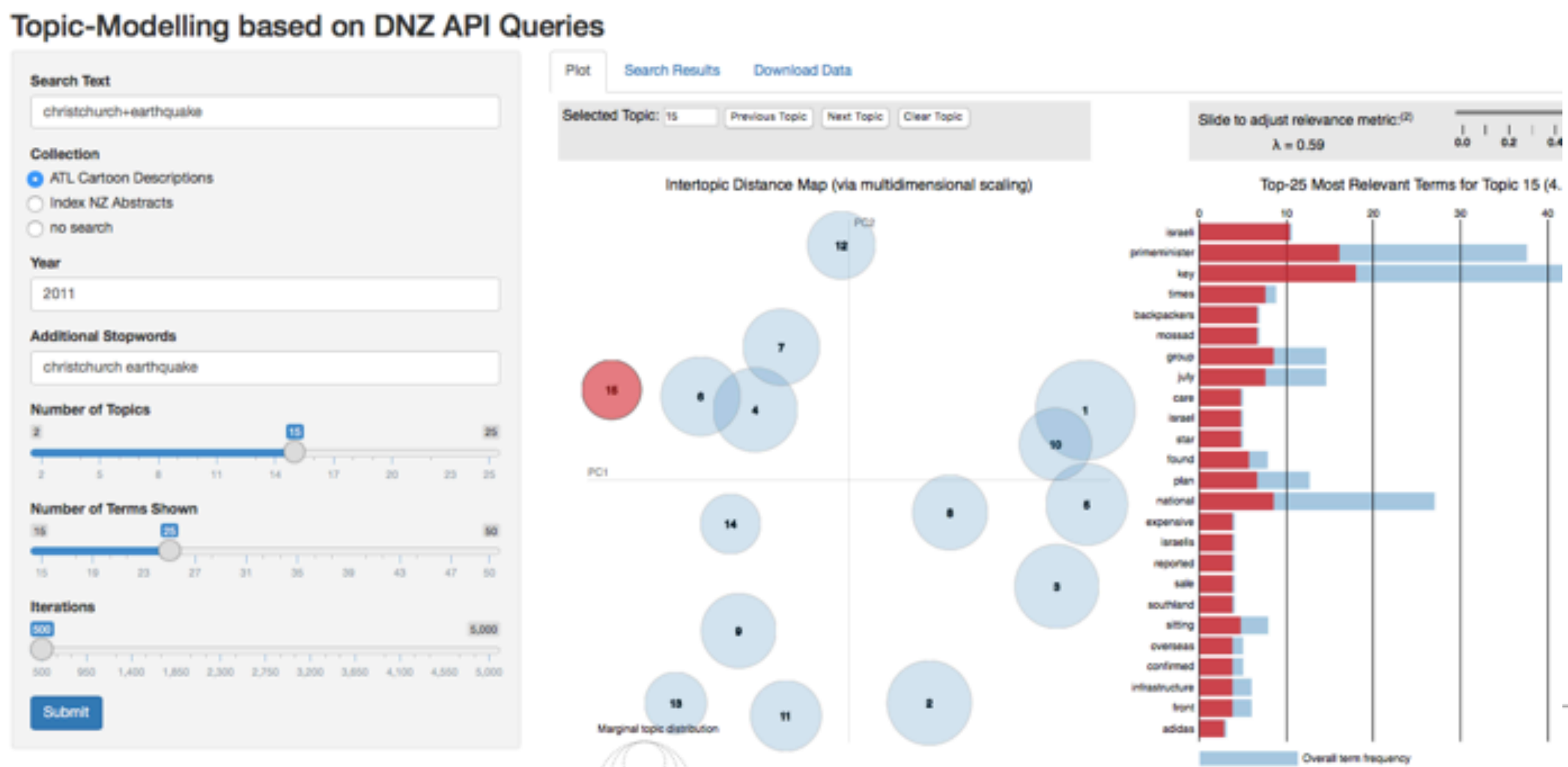
- The central inferential problem for LDA is determining the posterior distribution of the latent variables given the document.

Success and Results of LDA rely on apriori-set Variables

- Number of topics
- Number of iterations
- Normalisation of the data
- Stopwords

A Practical Example

- https://github.com/ThomasK81/TopicModellingR



What does each variable do?

- Search Text
- Collection
- Year
- Additional Stopwords
- Number of Topics
- Numbers of Terms Shown
- Iterations

The Risk of Normalisation

- http://thomask81.github.io/Greek_vis/#topic=4&lambda=1&term=
- ArabicMorph
- ArabicTranslated
- ArabicTM

Topic Modelling is not an end, but a means to an end.

Sight Reading Finder based on Topic Modelling Thucydides (Tufts, GRK 0103)

- Topic-Modelling Thucydides

- Morphological normalisation of Thucydides
  -> Result: Three different versions of the same text

- Passage-similarity matrix based on TM

- Put it into shiny
  -> Individual App for setting exam questions and enabling preparation.
  https://thomask81.shinyapps.io/sightreading_app/

Questions?