Sunoikisis Digital Classics
Fall 2020: session 4

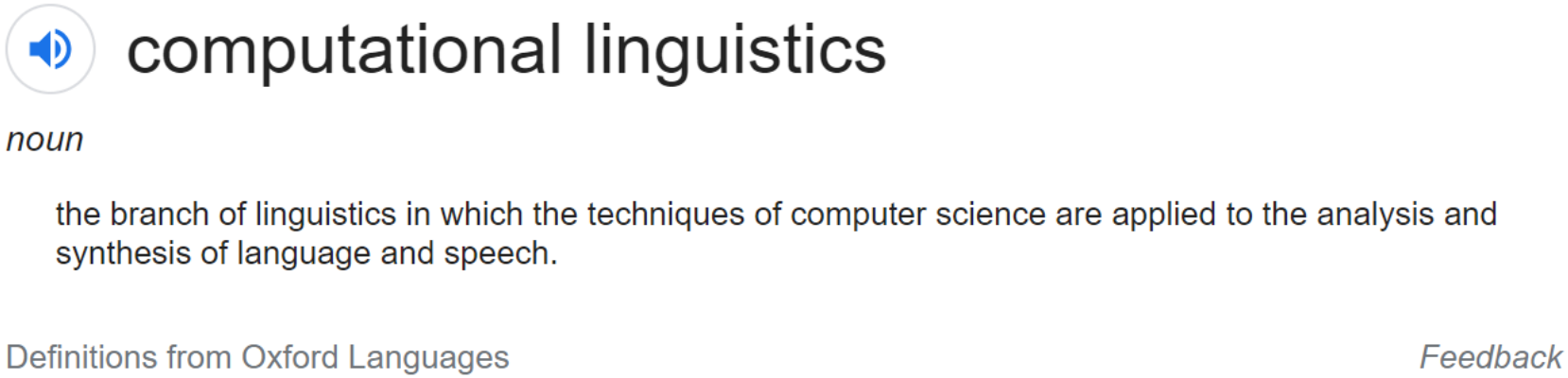
Overview of Computational Linguistics

Alek Keersmaekers (KU Leuven)
Marton Ribary (Surrey)
Thea Sommerschield (Oxford)

Classical Computational Linguistics: an introduction

What is computational linguistics?

- In a broad sense: any linguistic work with a computational methodology
- Computational analysis of linguistic data (i.e. corpus linguistics)
- Automated processing and/or annotation of linguistic data (i.e. natural language processing)



Corpus compilation (1)

- Most of the work in classical computational linguistics: focused on the compilation of corpus resources
- Full-text databases (see also *https://wiki.digitalclassicist.org/Greek_and_Latin_texts_in_digital_form*):
  - *TLG*, *Perseus* (literary texts)
  - *Packard Humanities Institute* (inscriptions)
  - *Duke Databank of Documentary Papyri* (papyri)
- Automatically lemmatized + morphologically annotated texts: e.g. *Perseus under Philologic*, *Diorisis, LatinISE*
- Manually lemmatized + morphologically and syntactically annotated texts (treebanks): see next slide
- Has led to several corpus-based studies (even though the classical languages are still lagging behind: see Jenset & McGillivray 2017)

Corpus compilation (2)

Greek project	Tokens	Texts
AGDT	560K	Archaic + Classical poetry/prose
Gorman	324K	Classical + post-classical prose
Pedalion	300K	Classical + post-classical poetry/prose
PROIEL	270K	Herodotus, NT, Sphrantzes
Harrington	18K	Lucian, Septuagint, Life of Aesop
Aphthonius	7K	Aphthonius
Sematia	6K	Papyri

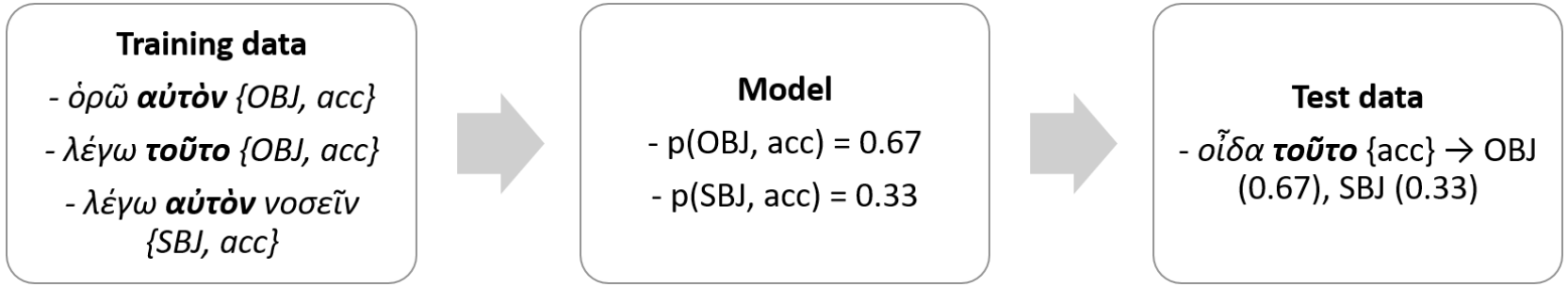
Latin project	Tokens	Texts
Index Thomisticus	450K	Thomas Aquinas
PROIEL	225K	Classical + post-classical prose
LLCT	200K	Medieval charter texts
Harrington	120K	Classical prose/poetry
ALDT	53K	Classical prose/poetry

NLP and Classical languages

- … in contrast, natural language processing of classical languages is still in its infancy (large amounts of data required)
- Natural language can be processed in several ways:
  - Morphological processing and lemmatization (McGillivray & Kilgariff 2013; Celano, Crane & Majidi 2016)
  - Syntactic parsing (e.g. Mambrini & Passarotti 2012; Ponti & Passarotti 2016)
  - Named Entity Recognition (e.g. Erdmann et al. 2016; Palladino, Karimi & Mathiak 2020)
  - Distributional lexical semantics (e.g. Rodda, Senaldi, and Lenci 2016; Bamman and Burns 2020)
  - …
- See the 3 case studies in this session for more examples
- See also the Classical Language Toolkit (http://cltk.org/, see also https://github.com/SunoikisisDC/SunoikisisDC-2017-2018/wiki/The-Classical-Language-Toolkit-(CLTK) and https://www.digitalclassicist.org/wip/wip2018-09pb.html)

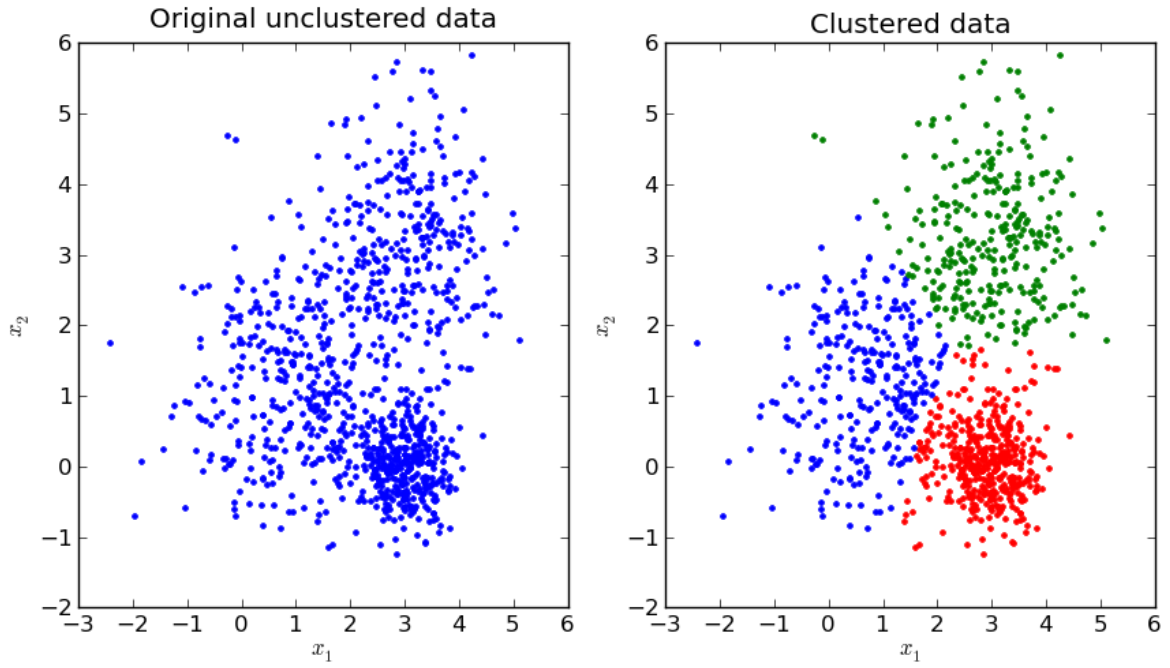
Machine learning: main concepts (1)

- Older computational work (for English): rule-based
- Most methods nowadays: machine learning
- **Supervised** machine learning:
  - To predict a certain class (e.g. OBJ vs. SBJ)
  - … one starts from a large corpus annotated for a number of **features** (e.g. case), the **training data**
  - The computer 'learns' patterns from the data and formalizes them in a mathematical **model**
  - This model can be used to predict class labels for new, unseen data, the **test data**



Machine learning: main concepts (2)

- **Unsupervised** machine learning (e.g. clustering): no pre-defined class labels
- Instead, the computer divides the data into groups with internally similar members



Classical computational linguistics: main challenges

- Typological characteristics of Greek and Latin (most NLP methods tailored to English): e.g. highly inflectional, free word order, large rate of ellipsis
- Large genre and diachronic variation
- Data sparseness

References

- Bamman, D. & Burns, P. J. 2020. Latin BERT: A Contextual Language Model for Classical Philology. *arXiv Preprint arXiv:2009.10053*.
- Celano, G. G. A., Crane, G. & Majidi, S. 2016. Part of Speech Tagging for Ancient Greek. *Open Linguistics* 2(1).
- Erdmann, A., Brown, C., Joseph, B., … Marneffe, M.-C. de. 2016. Challenges and Solutions for Latin Named Entity Recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*. Osaka: The COLING 2016 Organizing Committee, pp. 85–93.
- Jenset, G. B. & McGillivray, B. 2017. *Quantitative historical linguistics: a corpus framework*. Oxford: Oxford University Press.
- Mambrini, F. & Passarotti, M. C. 2012. Will a parser overtake Achilles? First experiments on parsing the ancient Greek dependency treebank. In *Eleventh International Workshop on Treebanks and Linguistic Theories*. Edições Colibri, pp. 133–144.
- McGillivray, B. & Kilgarriff, A. 2013. Tools for historical corpus research, and a corpus of Latin. In Durrell, P., Scheible, M., Whitt, S. & Bennett, R. J. (eds.), *New Methods in Historical Corpus Linguistics*. pp. 247–257.
- Palladino, C., Karimi, F. & Mathiak, B. 2020. NER on Ancient Greek with minimal annotation. In *Digital Humanities 2020*. Ottawa: DH2020.
- Ponti, E. M. & Passarotti, M. 2016. Differentia compositionem facit. A Slower-Paced and Reliable Parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA), pp. 683–688.
- Rodda, M. A., Senaldi, M. S. & Lenci, A. 2016. Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. In *CLiC-it/EVALITA*.

Text pre-processing and semantic clustering

Marton Ribary (Surrey)

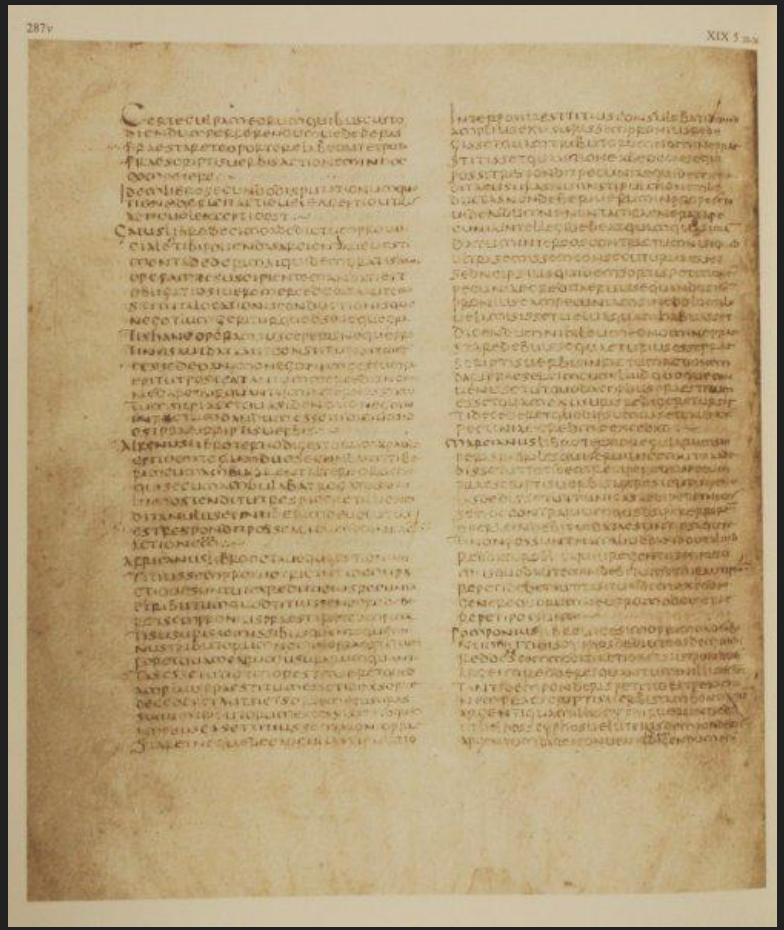
Sunoikisis Digital Classics
Fall 2020 Session 4. Computational Linguistics
Thursday 29 October 2020

Article

Ribary, M and McGillivray, B.,
“A Corpus Approach to Roman Law Based on Justinian’s Digest.”
Informatics **7** (2020), 44.
DOI: https://doi.org/10.3390/informatics7040044

Demo

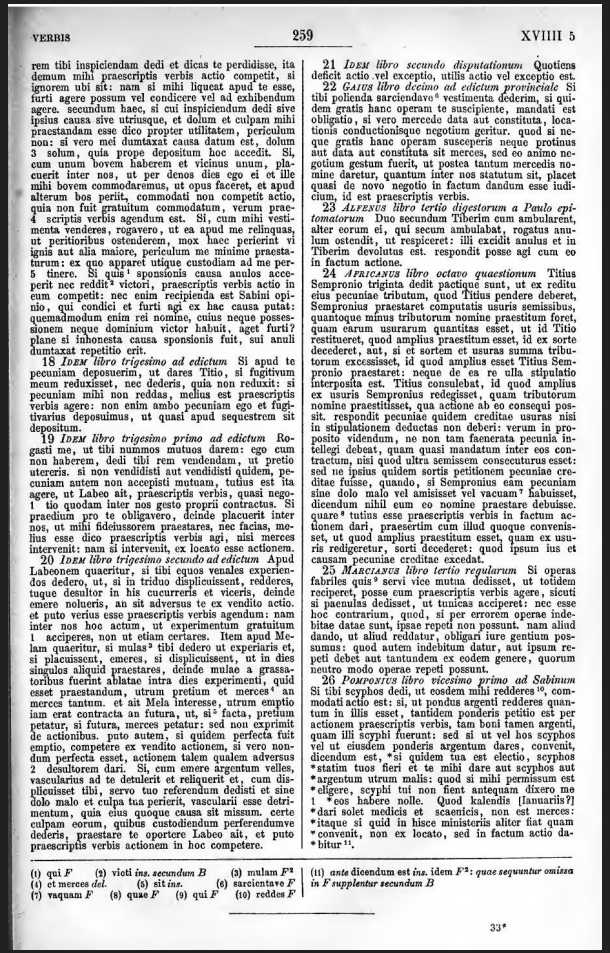
https://colab.research.google.com/drive/1fBsLYLeuh4raQ1CGfvlWAZe2ftKRzHXR?usp=sharing



Manuscript

- Digest - historical sourcebook of Roman law in 432 thematic sections
  - 9132 passages in 21055 text units
  - (almost) exclusive source for Roman law
- compiled by Tribonian under the order of Justinian I
- and published in 533 CE
- oldest (and complete) MS from 554 CE

Littera Florentina 287v with texts for D.19,5,20-26



VERBIS 259 XVIIII 5

rem tibi inspiciendam dedi et dicas te perdidisse, ita demum mihi praescriptis verbis actio competit, si ignorem ubi sit: nam si mihi liqueat apud te esse, furti agere possum vel condicere vel ad exhibendum agere. secundum haec, si cui inspiciendum dedi sive ipsius causa sive utriusque, et dolum et culpam mihi praestandam esse dico propter utilitatem, periculum non: si vero mei dumtaxat causa datum est, dolum 3 solum, quia prope depositum hoc accedit. Si, cum unum bovem haberem et vicinus unum, placuerit inter nos, ut per denos dies ego ei et ille mihi bovem commodaremus, ut opus faceret, et apud alterum bos periit, commodati non competit actio, quia non fuit gratuitum commodatum, verum prae- 4 scriptis verbis agendum est. Si, cum mihi vestimenta venderes, rogavero, ut ea apud me relinquas, ut peritioribus ostenderem, mox haec perierint vi ignis aut alia maiore, periculum me minime praestaturum: ex quo apparet utique custodiam ad me per- 5 tinere. Si quis [1] sponsionis causa anulos acceperit nec reddit [2] victori, praescriptis verbis actio in eum competit: nec enim recipienda est Sabini opinio, qui condici et furti agi ex hac causa putat: quemadmodum enim rei nomine, cuius neque possessionem neque dominium victor habuit, aget furti? plane si inhonesta causa sponsionis fuit, sui anuli dumtaxat repetitio erit.

18 *Idem libro trigesimo ad edictum* Si apud te pecuniam deposuerim, ut dares Titio, si fugitivum meum reduxisset, nec dederis, quia non reduxit: si pecuniam mihi non reddas, melius est praescriptis verbis agere: non enim ambo pecuniam ego et fugitivarius deposuimus, ut quasi apud sequestrem sit depositum.

19 *Idem libro trigesimo primo ad edictum* Rogasti me, ut tibi nummos mutuos darem: ego cum non haberem, dedi tibi rem vendendam, ut pretio utereris. si non vendidisti aut vendidisti quidem, pecuniam autem non accepisti mutuam, tutius est ita agere, ut Labeo ait, praescriptis verbis, quasi nego- 1 tio quodam inter nos gesto proprii contractus. Si praedium pro te obligavero, deinde placuerit inter nos, ut mihi fideiussorem praestares, nec facias, melius esse dico praescriptis verbis agi, nisi mercES intervenit: nam si intervenit, ex locato esse actionem.

20 *Idem libro trigesimo secundo ad edictum* Apud Labeonem quaeritur, si tibi equos venales experiendos dedero, ut, si in triduo displicuissent, redderes, tuque desultor in his cucurreris et viceris, deinde emere nolueris, an sit adversus te ex vendito actio. et puto verius esse praescriptis verbis agendum: nam inter nos hoc actum, ut experimentum gratuitum 1 acciperes, non ut etiam certares. Item apud Melam quaeritur, si mulas [3] tibi dedero ut experiaris et, si placuissent, emeres, si displicuissent, ut in dies singulos aliquid praestares, deinde mulae a grassatoribus fuerint ablatae intra dies experimenti, quid esset praestandum, utrum pretium et mercES [4] an merces tantum. et ait Mela interesse, utrum emptio iam erat contracta an futura, ut, si [5] facta, pretium petatur, si futura, merces petatur: sed non exprimit de actionibus. puto autem, si quidem perfecta fuit emptio, competere ex vendito actionem, si vero nondum perfecta esset, actionem talem qualem adversus 2 desultorem dari. Si, cum emere argentum velles, vascularius ad te detulerit et reliquerit et, cum displicuisset tibi, servo tuo referendum dedisti et sine dolo malo et culpa tua perierit, vascularii esse detrimentum, quia eius quoque causa sit missum. certe culpam eorum, quibus custodiendum perferendumve dederis, praestare te oportere Labeo ait, et puto praescriptis verbis actionem in hoc competere.

21 *Idem libro secundo disputationum* Quotiens deficit actio vel exceptio, utilis actio vel exceptio est.

22 *Gaius libro decimo ad edictum provinciale* Si tibi polienda sarciendave [6] vestimenta dederim, si quidem gratis hanc operam te suscipiente, mandati est obligatio, si vero mercede data aut constituta, locationis conductionisque negotium geritur. quod si neque gratis hanc operam susceperis neque protinus aut data aut constituta sit merces, sed eo animo negotium gestum fuerit, ut postea tantum mercedis nomine daretur, quantum inter nos statutum sit, placet quasi de novo negotio in factum dandum esse iudicium, id est praescriptis verbis.

23 *Alfenus libro tertio digestorum a Paulo epitomatorum* Duo secundum Tiberim cum ambularent, alter eorum ei, qui secum ambulabat, rogatus anulum ostendit, ut respiceret: illi excidit anulus et in Tiberim devolutus est. respondit posse agi cum eo in factum actione.

24 *Africanus libro octavo quaestionum* Titius Sempronio triginta dedit pactique sunt, ut ex reditu eius pecuniae tributum, quod Titius pendere deberet, Sempronius praestaret computatis usuris semissibus, quantoque minus tributorum nomine praestitum foret, quam earum usurarum quantitas esset, ut id Titio restitueret, quod amplius praestitum esset, id ex sorte decederet, aut, si et sortem et usuras summa tributorum excessisset, id quod amplius esset Titius Sempronio praestaret: neque de ea re ulla stipulatio interposita est. Titius consulebat, id quod amplius ex usuris Sempronius redegisset, quam tributorum nomine praestitisset, qua actione ab eo consequi possit. respondit pecuniae quidem creditae usuras nisi in stipulationem deductas non deberi: verum in proposito videndum, ne non tam faenerata pecunia intellegi debeat, quam quasi mandatum inter eos contractum, nisi quod ultra semissem consecuturus esset: sed ne ipsius quidem sortis petitionem pecuniae creditae fuisse, quando, si Sempronius eam pecuniam sine dolo malo vel amisisset vel vacuam [7] habuisset, dicendum nihil eum eo nomine praestare debuisse. quare [8] tutius esse praescriptis verbis in factum actionem dari, praesertim cum illud quoque convenisset, ut quod amplius praestitum esset, quam ex usuris redigeretur, sorti decederet: quod ipsum ius et causam pecuniae creditae excedat.

25 *Marcianus libro tertio regularum* Si operas fabriles quis [9] servi vice mutua dedisset, ut totidem reciperet, posse cum praescriptis verbis agere, dicuti si paenulas dedisset, ut tunicas acciperet: nec esse hoc contrarium, quod, si per errorem operae indebitae datae sunt, ipsae repeti non possunt. nam aliud dando, ut aliud reddatur, obligari iure gentium possumus: quod autem indebitum datur, aut ipsum repeti debet aut tantundem ex eodem genere, quorum neutro modo operae repeti possunt.

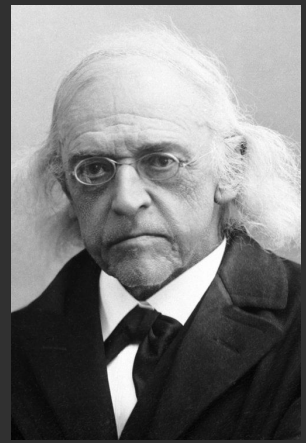
26 *Pomponius libro vicesimo primo ad Sabinum* Si tibi scyphos dedi, ut eosdem mihi redderes [10], commodati actio est: si, ut pondus argenti redderes quantum in illis esset, tantidem ponderis petitio est per actionem praescriptis verbis, tam boni tamen argenti, quam illi scyphi fuerunt: sed si ut vel hos scyphos vel ut eiusdem ponderis argentum dares, convenit, dicendum est, *si quidem tua est electio, scyphos *statim tuos fieri et te mihi dare aut scyphos aut *argentum utrum malis: quod si mihi permissum est *eligere, scyphi tui non fient antequam dixero me 1 *eos habere nolle. Quod kalendis [Ianuariis?] *dari solet medicis et scaenicis, non est merces: *itaque si quid in hisce ministeriis aliter fiat quam *conveni, non ex locato, sed in factum actio da- *bitur [11].

(1) qui *F* (2) victi *ins. secundum B* (3) mulam F^2
(4) et merces *del.* (5) sit *ins.* (6) sarcientave *F*
(7) vaquam *F* (8) quae *F* (9) qui *F* (10) reddes *F*
(11) *ante* dicendum est *ins.* idem F^2: *quae sequuntur omissa in F supplentur secundum B*

33*

Printed edition

- Largely based on the *littera Florentina*
- Edited by Theodor Mommsen



Theodor Mommsen (1817-1903)

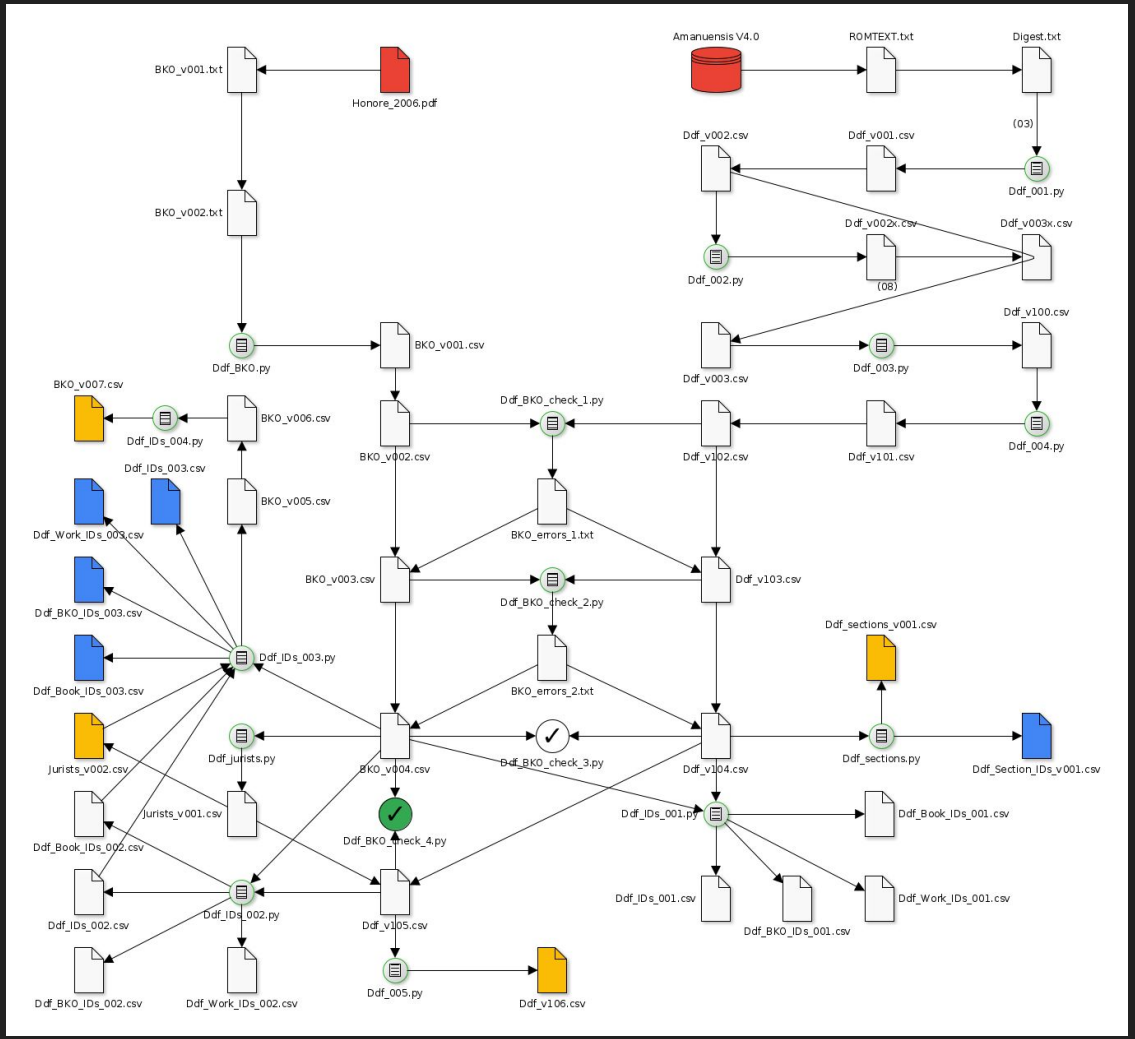
D. 19,5,17,3 - D. 19,5,26 from Mommsen's edition of the Digest (1889)

Digital edition

- ROMTEXT database (Linz) since the 1970s
- Amanuensis interface developed by Peter Riedlberger

Amanuensis V5.0 (2020) ©Peter Riedlberger

http://www.riedlberger.de/08amanuensis.html



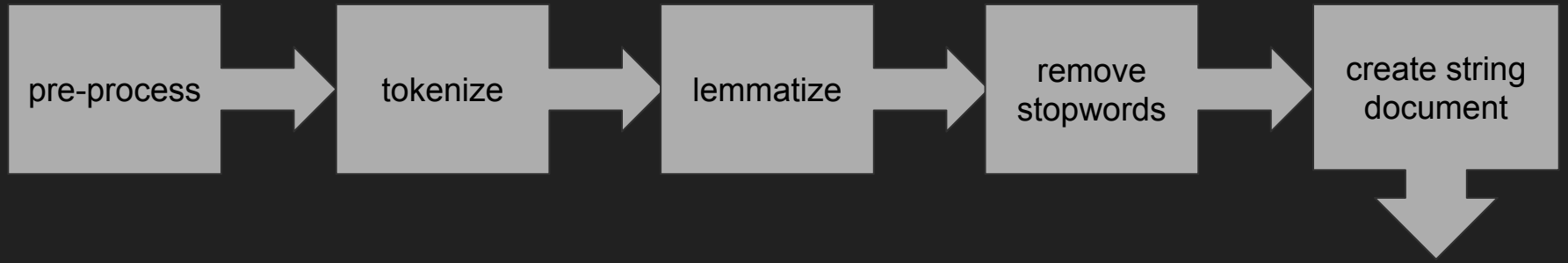
Database

Processing flowchart from the GitLab repository
https://gitlab.eps.surrey.ac.uk/mr0048/pydigest

Ribary, M., “A Relational Database of Roman Law Based on Justinian’s Digest.” *Journal of Open Humanities Data* 6(1), p.5. DOI: http://doi.org/10.5334/johd.17

Ribary, M., “A relational database of Roman law based on Justinian’s Digest.” *figshare*. Dataset. https://doi.org/10.6084/m9.figshare.12333290.v1

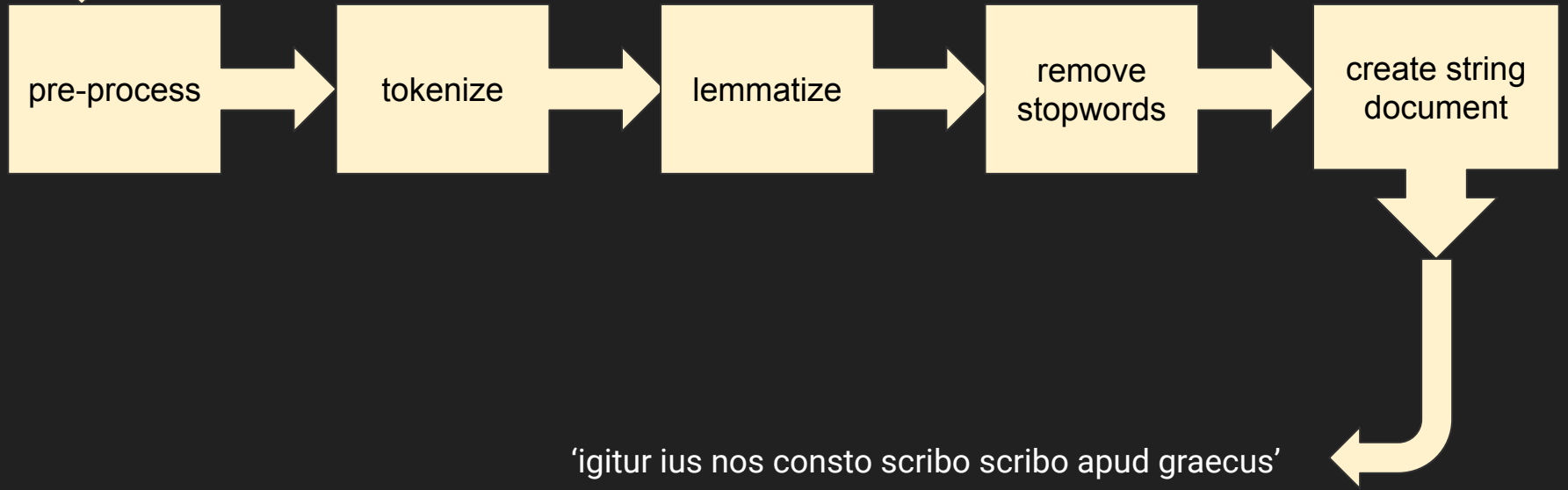
Getting the text ready



Demo available

'Hoc igitur ius nostrum constat aut ex scripto aut sine scripto, ut apud Graecos: τῶν νόμων οἱ μὲν ἔγγραφοι, οἱ δὲ ἄγραφοι .'

(D.1.1.6.1)



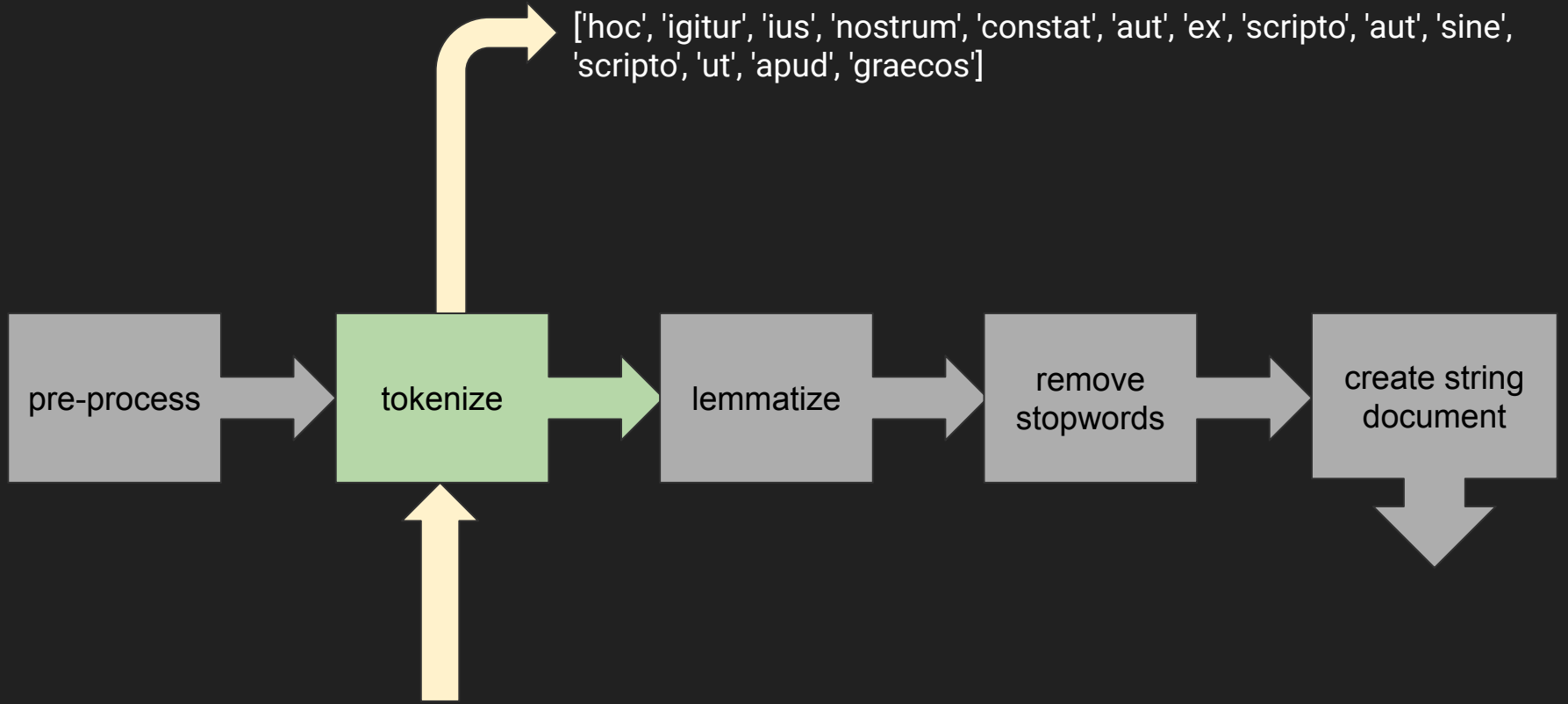
Demo available

'Hoc igitur ius nostrum constat aut ex scripto aut sine scripto, ut apud Graecos: τῶν νόμων οἱ μὲν ἔγγραφοι, οἱ δὲ ἄγραφοι .'

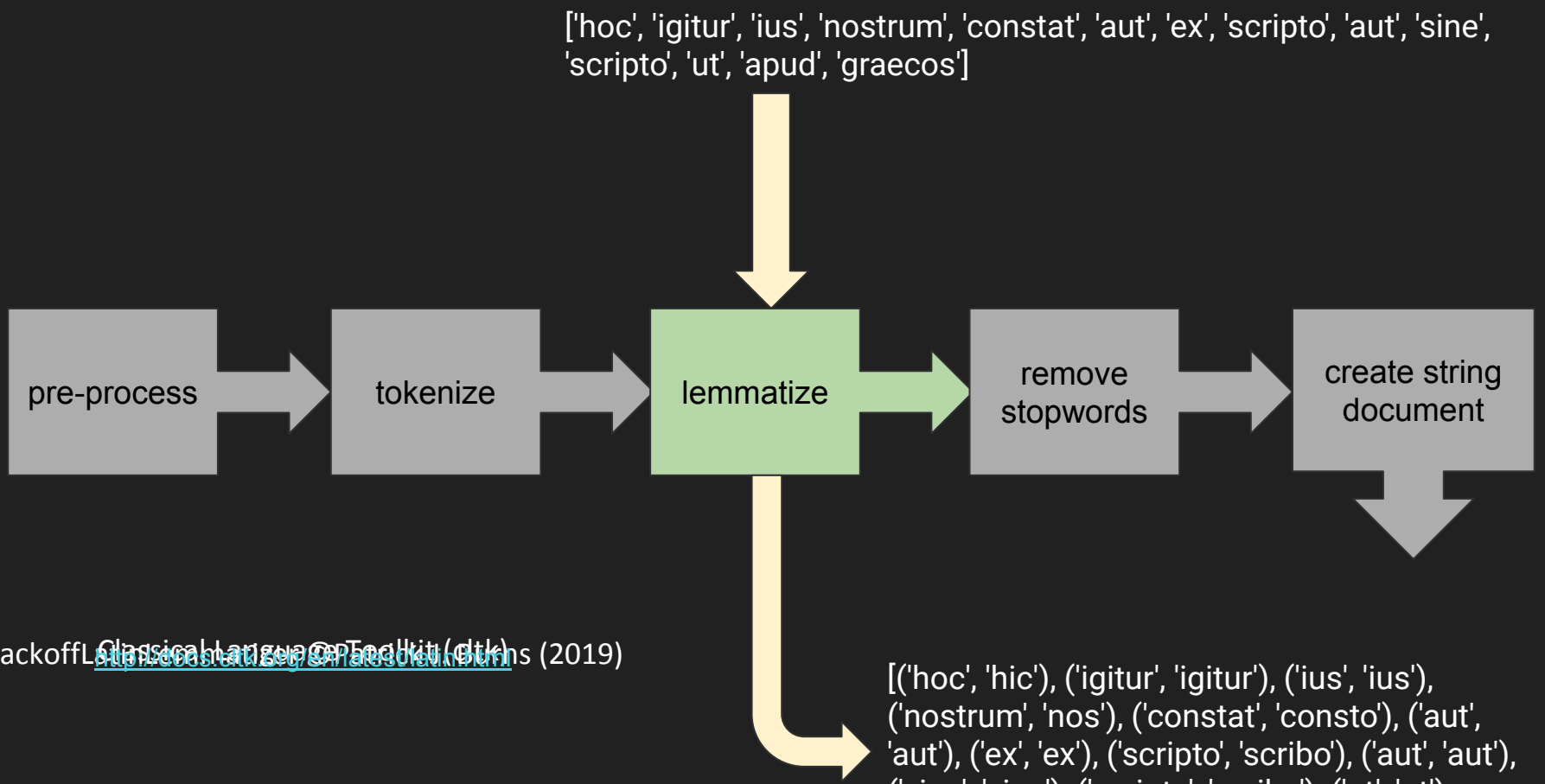
(D.1.1.6.1)

'hoc igitur ius nostrum constat aut ex scripto aut sine scripto ut apud graecos'

Demo available



Demo available

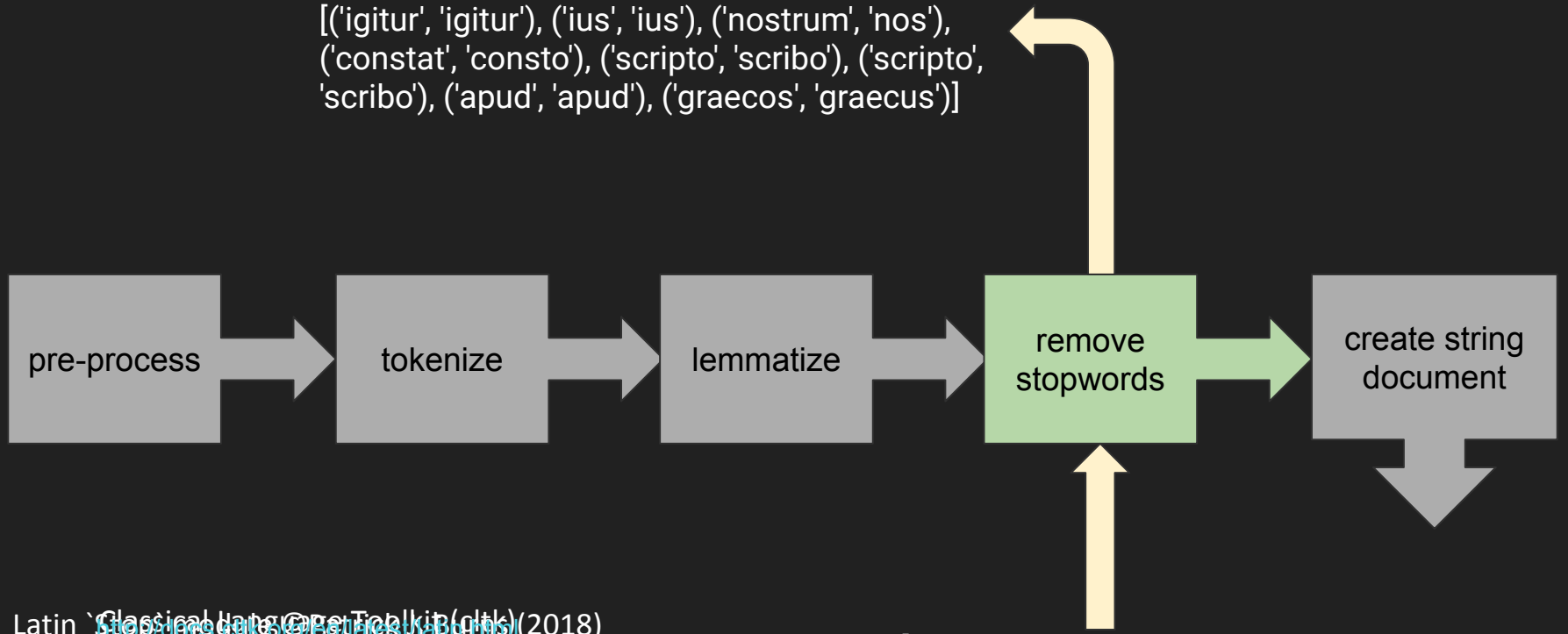


BackoffLatinLemmatizer (Patrick J. Burns (2019)

Classical Language Toolkit (CLTK)

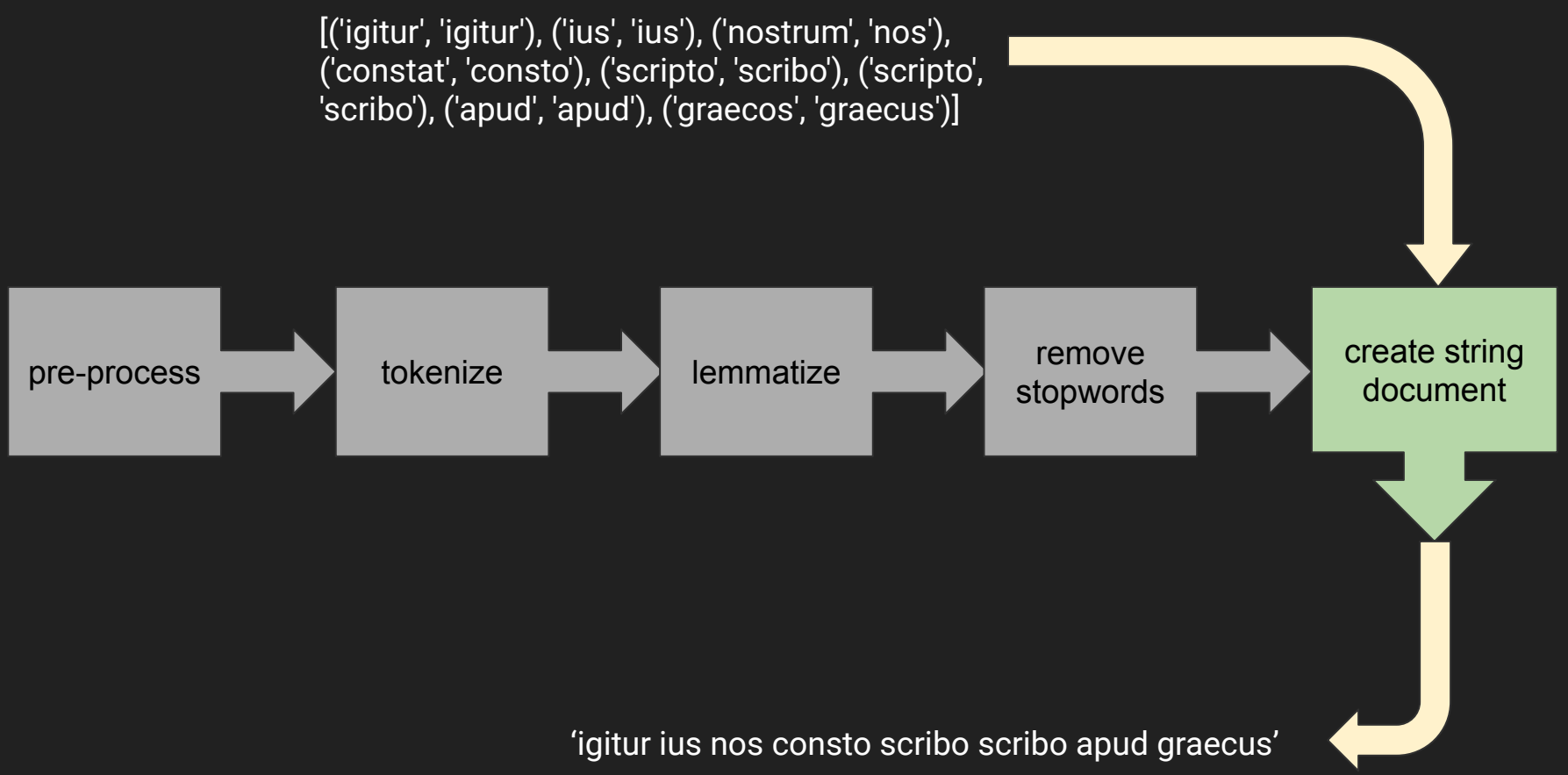
http://docs.cltk.org/en/latest/latin.html

Demo available



Latin `Classical Language Toolkit (cltk)` (2018)

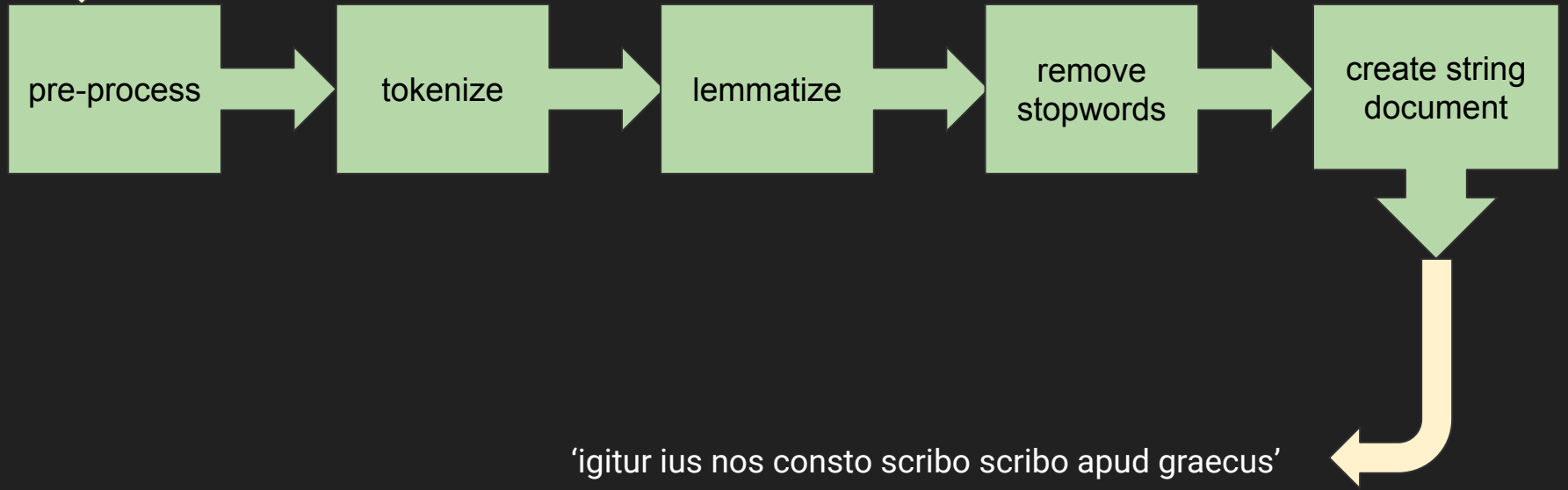
Demo available



Demo available

'Hoc igitur ius nostrum constat aut ex scripto aut sine scripto, ut apud Graecos: τῶν νόμων οἱ μὲν ἔγγραφοι, οἱ δὲ ἄγραφοι .'

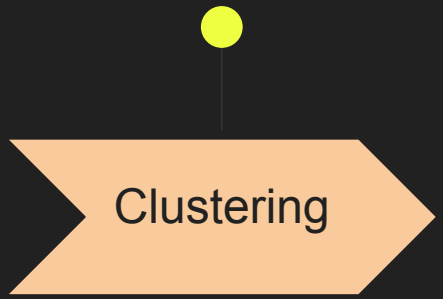
(D.1.1.6.1)



‘igitur ius nos consto scribo scribo apud graecus’

Demo available

Discover the structure of Roman law from the bottom up



Assumptions:

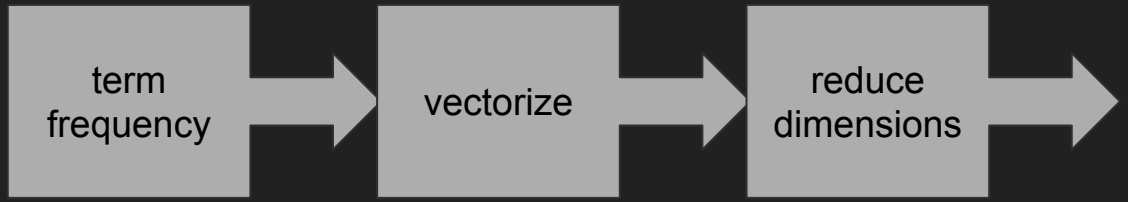
1. Sections are thematically homogeneous
2. The Digest is a comprehensive overview of what Roman law is

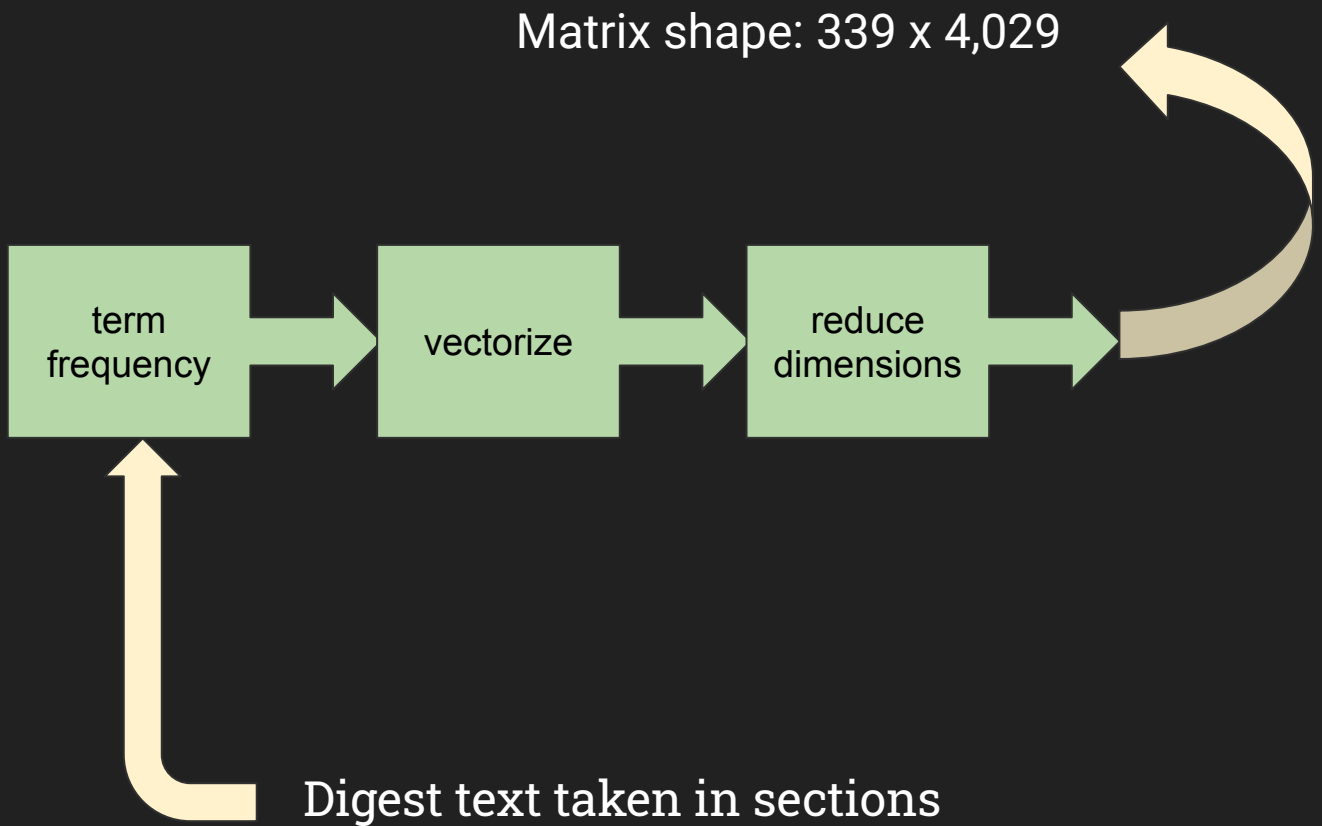
Textual data of sections can be transformed to numbers by tfidf

Hierarchical clustering of sections brings out the bottom-up structure of Roman law

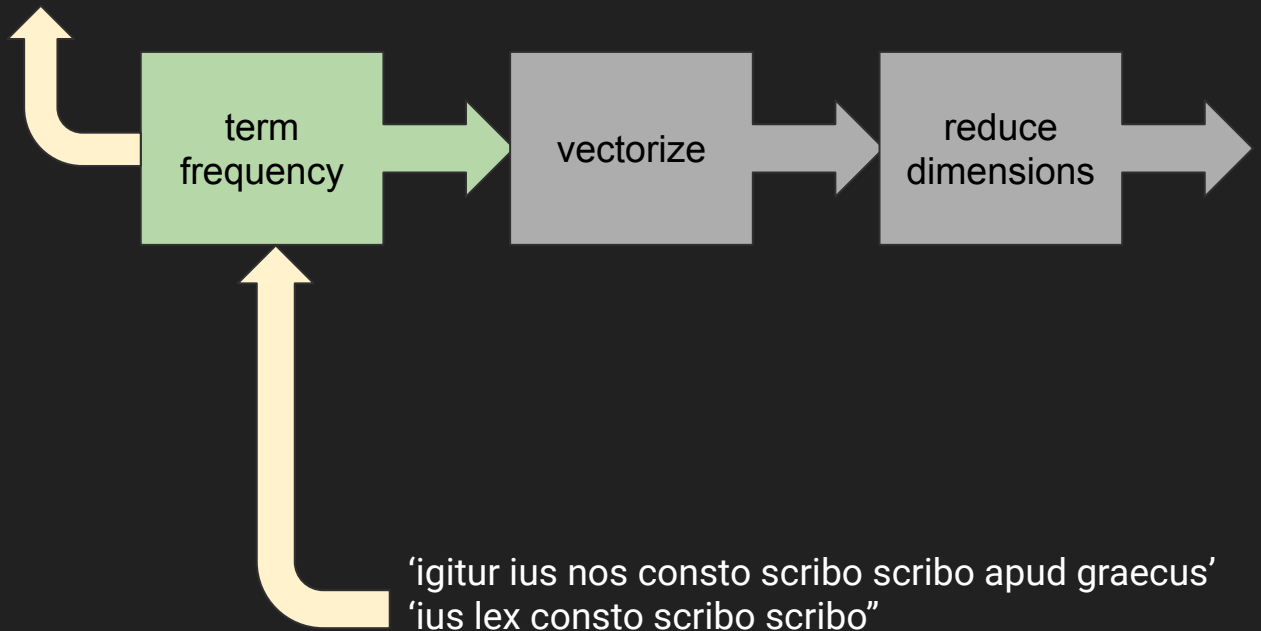
Turn text to numbers

Term frequency - Inverse document frequency (tfidf)

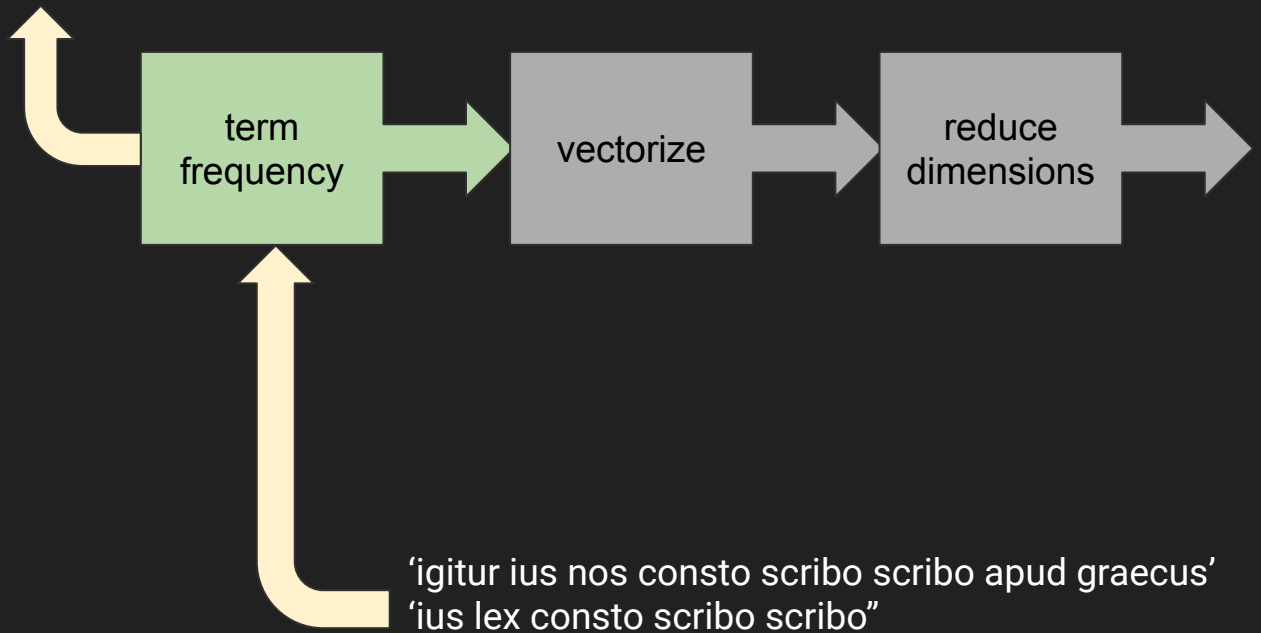



Matrix shape: 339 x 4,029
term frequency
vectorize
reduce dimensions
Digest text taken in sections

	igitur	ius	nos	consto	scribo	apud	graecus	lex	romanus
1	1	1	1	1	2	1	1	0	0
2	0	1	0	1	2	0	0	1	0
3	1	0	0	1	0	1	0	1	1



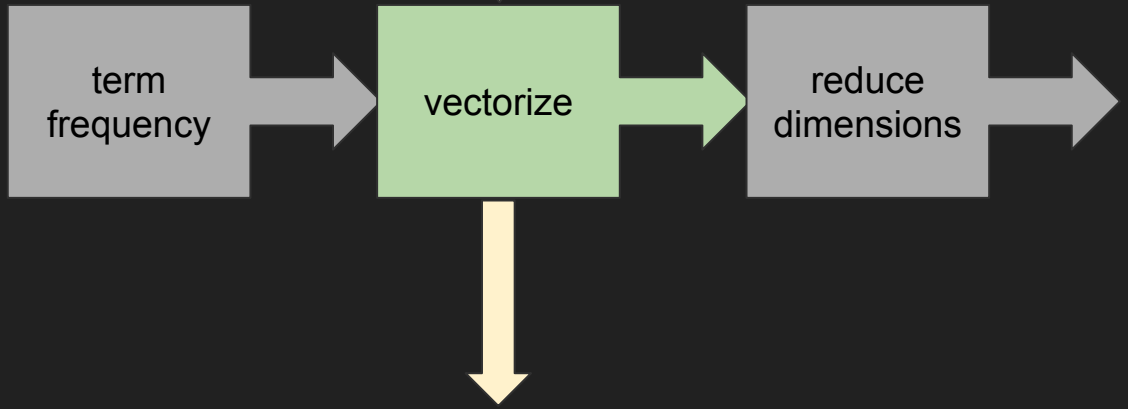
1	1	1	1	2	1	1	0	0
0	1	0	1	2	0	0	1	0
1	0	0	1	0	1	0	1	1



1	1	1	1	2	1	1	0	0
0	1	0	1	2	0	0	1	0
1	0	0	1	0	1	0	1	1

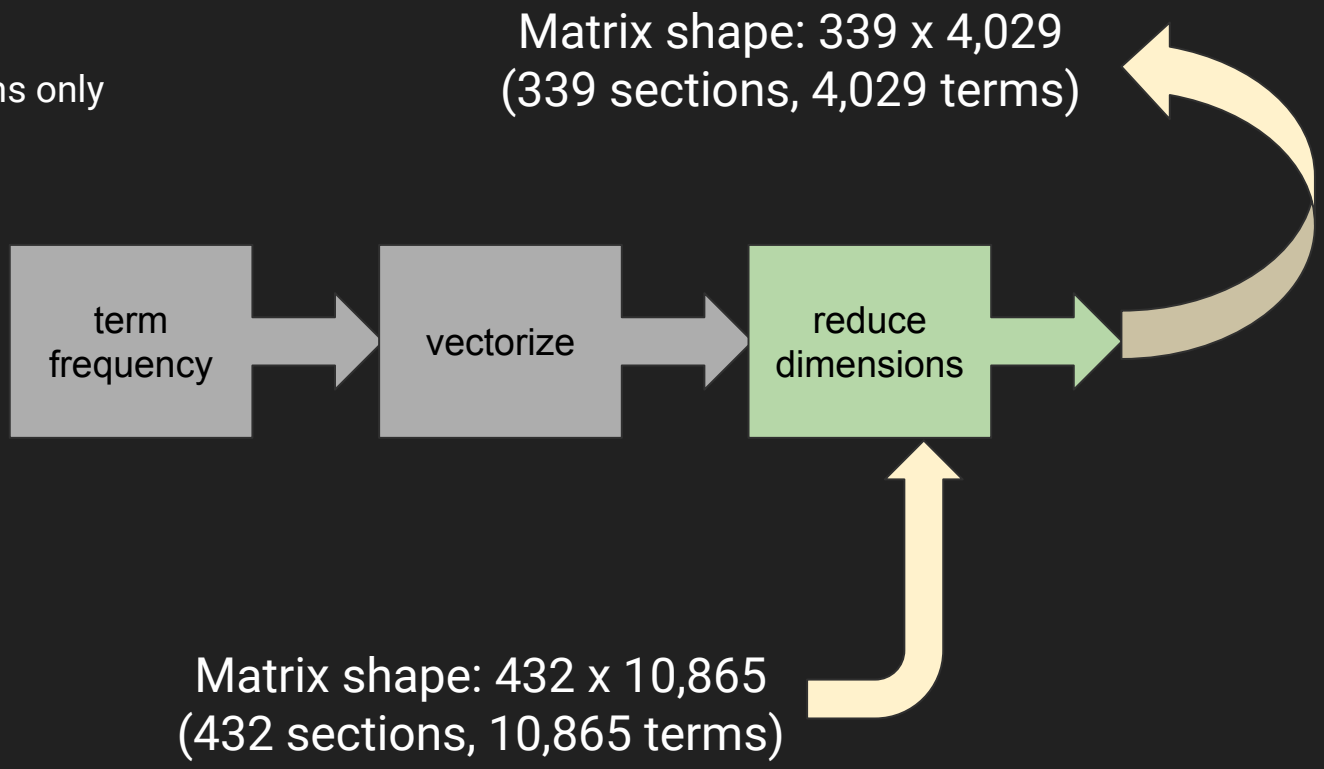
Term frequency - Inverse document frequency (tfidf)

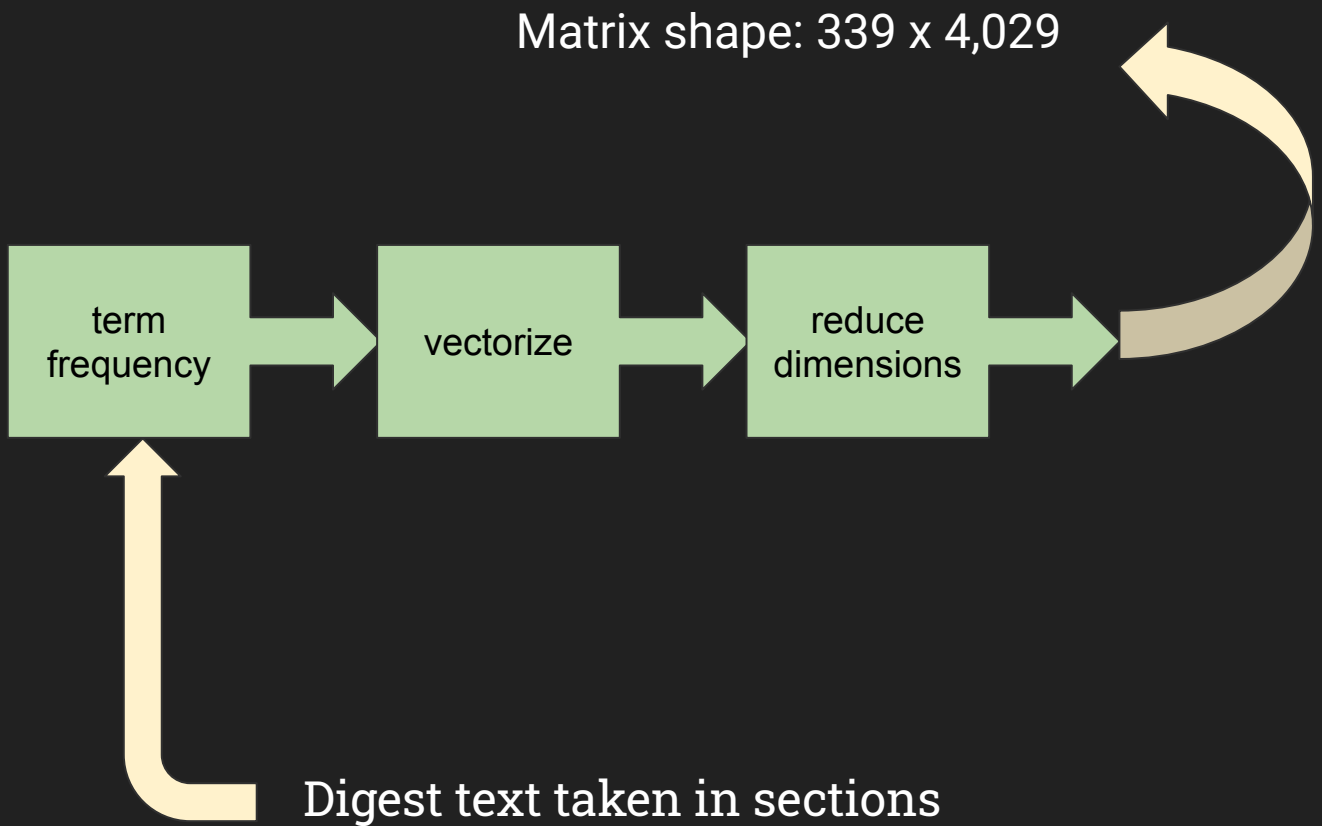
all sections with word weighting

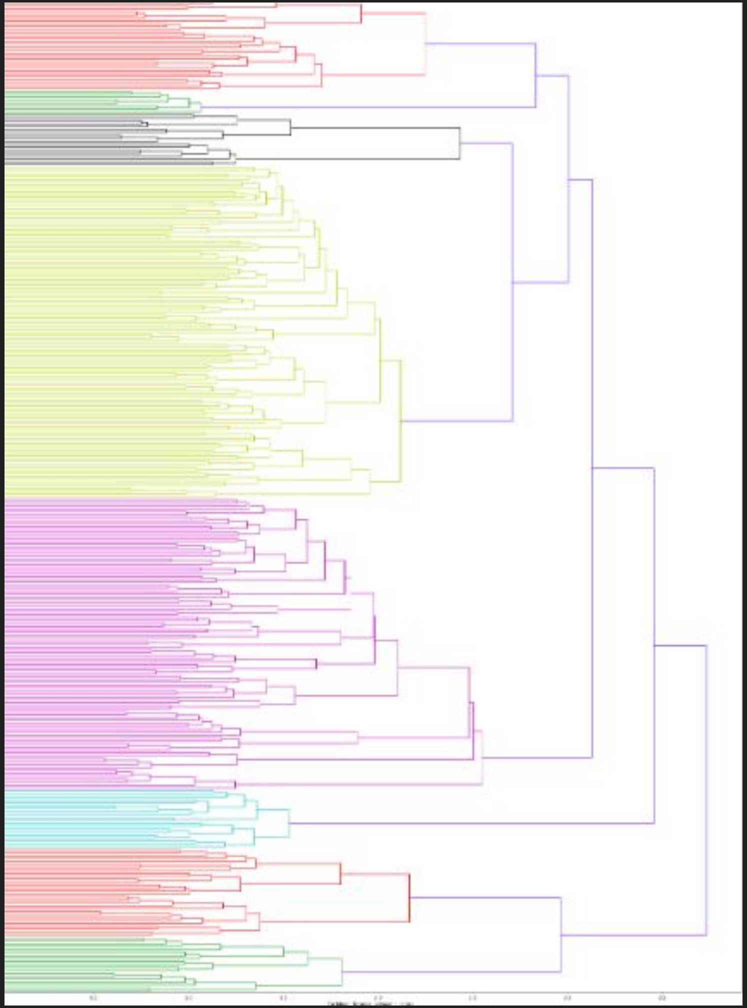


Matrix shape: 432 x 10,865
(432 sections, 10,865 terms)

- Remove short sections (339 remains)
- Select top 50 terms only




Matrix shape: 339 x 4,029
term frequency
vectorize
reduce dimensions
Digest text taken in sections

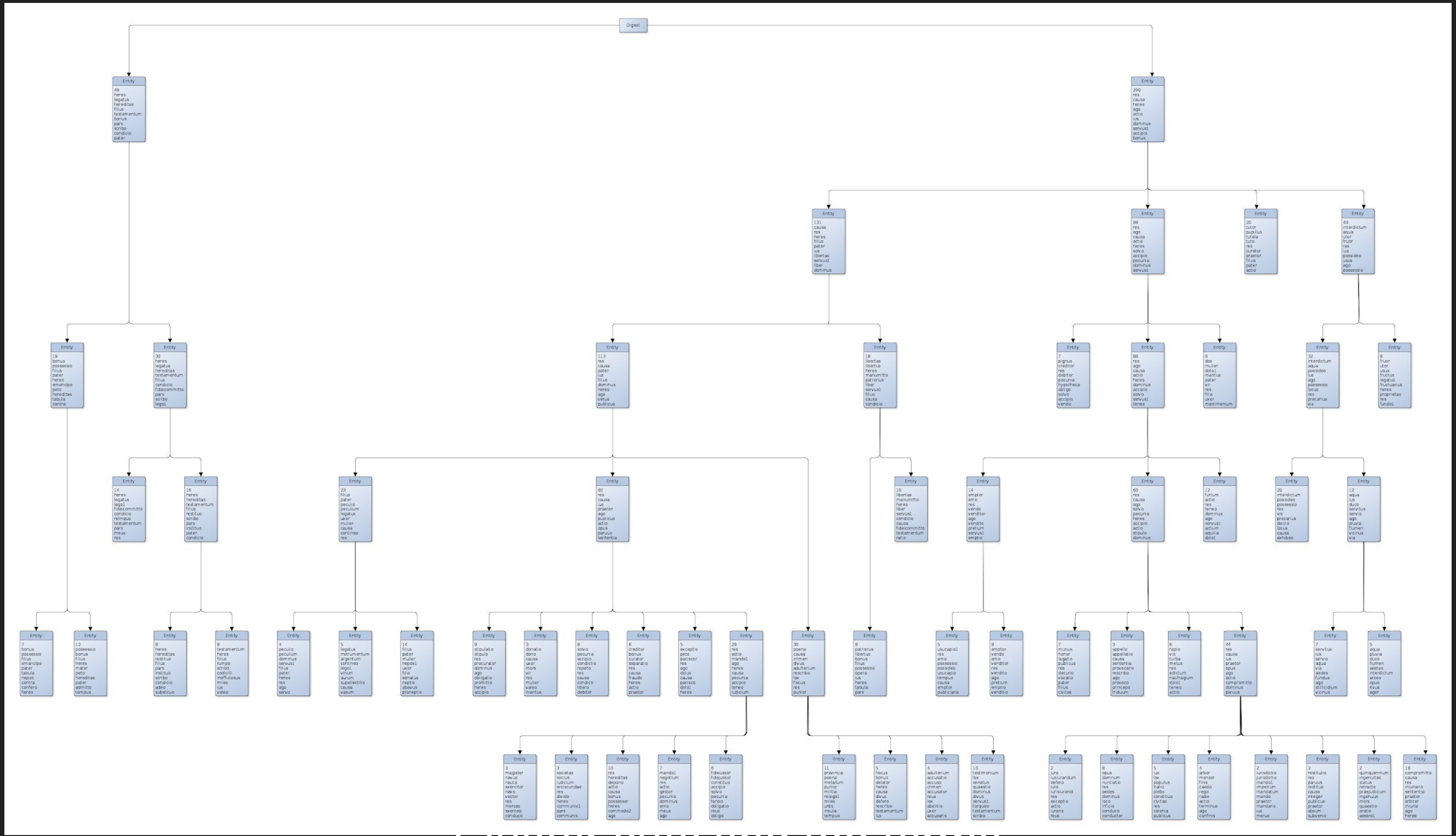


Hierarchical clustering on the tfidf matrix

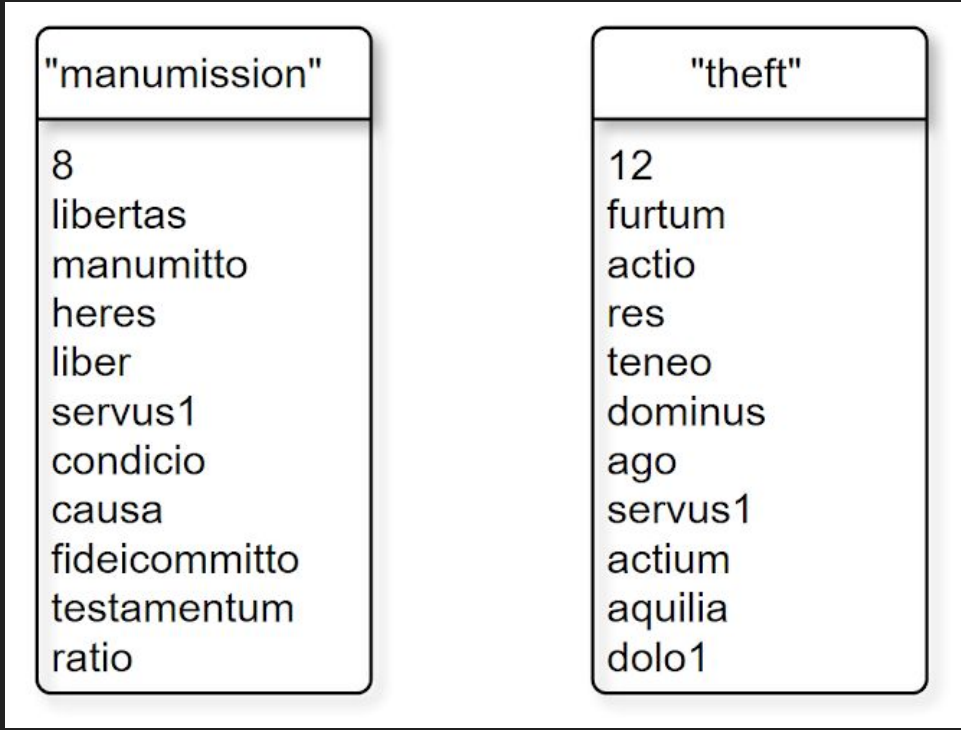
Threshold (Euclidean distance)	Number of clusters at threshold
3.5	2
3.0	5
2.5	10
2.0	17
1.75	31
1.50	55
1.375	80

Dendrogram of hierarchical clustering

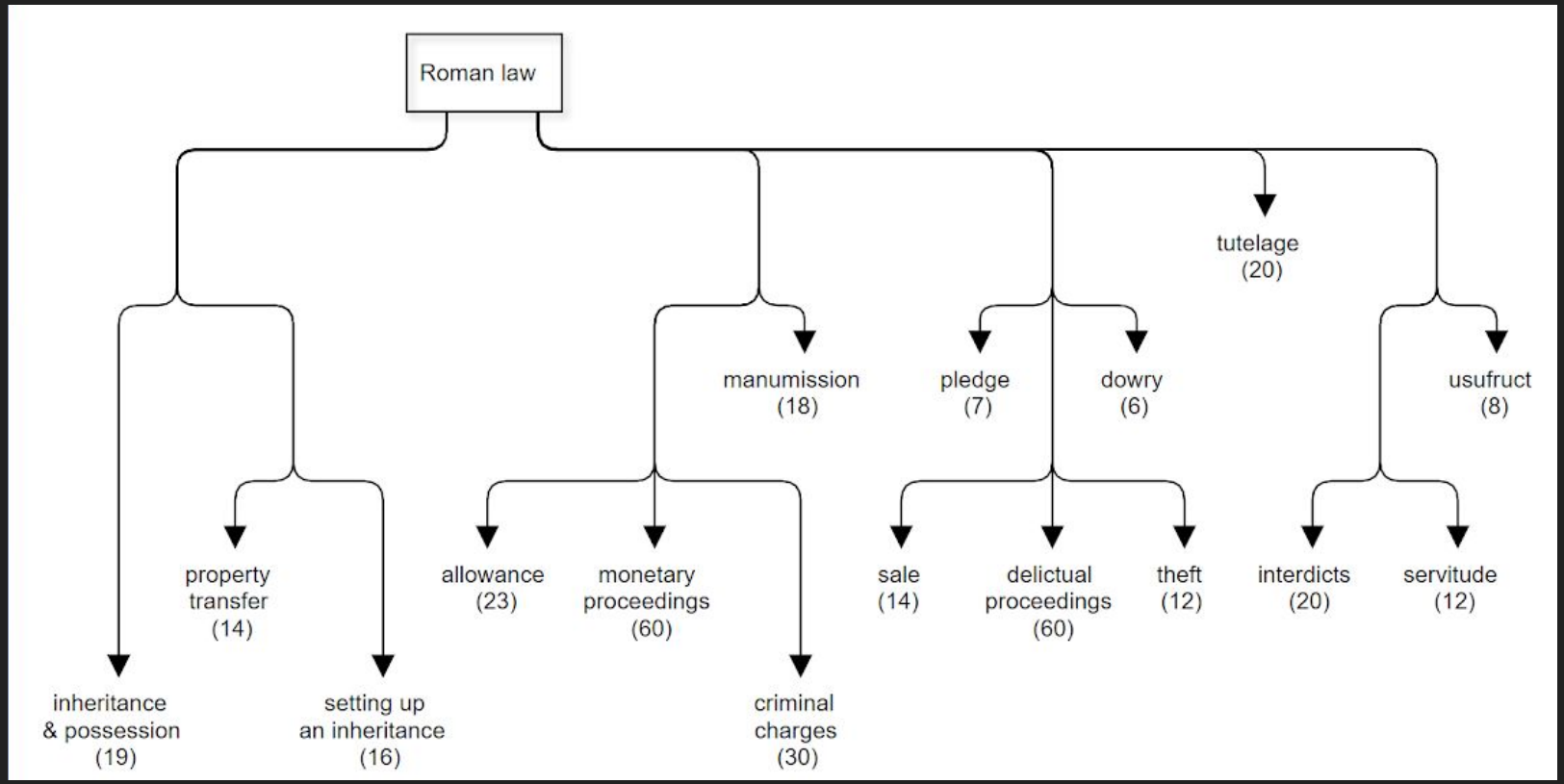
https://gitlab.eps.[illegible]master/NLP_docu



https://gitlab.eps.surrey.ac.uk/mr0048/pydigest/-/blob/master/NLP_docu
/mentation.md#4-inspect-the-output-of-iterative-clustering



Cluster examples



A simplified empirical conceptual tree-map of Roman law

References

Burns, P. J., "Building a text analysis pipeline for classical languages," in Digital classical philology: Ancient Greek and Latin in the digital revolution, edited by Monica Berti, Berlin: Walter de Gruyter, 2019, 159-176.

Burns, P. J., "Constructing stoplists for historical languages," Digital Classics Online 4:2 (2018): 4-20.

Burns, P. J., "Multiplex lemmatization with the Classical Language Toolkit," presented at the First LiLa Workshop: Linguistic Resources & NLP Tools for Latin on 3 June 2019.

Klingenberg, G., "Die ROMTEXT-Datenbank," Informatica e diritto 4 (1995): 223-232.

McGillivray, B. and Kilgarriff, A., "Tools for historical corpus research, and a corpus of Latin," on the SketchEngine website. [Accessed on 15 June 2020]

Mommsen, T. and Krüger, P., eds. Corpus Iuris Civlis. Vol 1: Institutiones. Digesta. 5th edition. Berlin: Weidmann, 1889.

Ribary, M. and McGillivray, B. "A Corpus Approach to Roman Law Based on Justinian's Digest." Informatics 7 (2020), 44. Available: https://doi.org/10.3390/informatics7040044

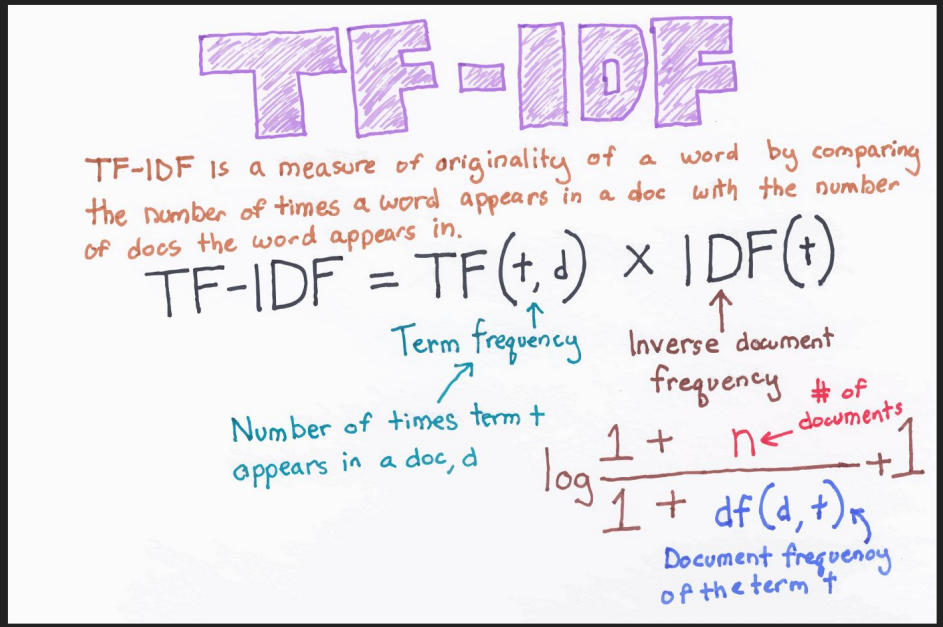
Ribary, M., "A Relational Database of Roman Law Based on Justinian's Digest." Journal of Open Humanities Data 6(1), p.5. DOI: http://doi.org/10.5334/johd.17

Ribary, M., "A relational database of Roman law based on Justinian's Digest." figshare. Dataset. https://doi.org/10.6084/m9.figshare.12333290.v1

Ribary, M. pyDigest, A GitLab Repository of Scripts, Files and Documentation. Available online: https://gitlab.eps.surrey.ac.uk/mr0048/pydigest

Riedlberger, P. and Rosenbaum, G., eds. Amanuensis V5.0. München: 2020.

Tfidf and dimension reduction



- Term frequency - Inverse document frequency

  => 432 x 10,865

- Dimension reduction
  - Remove short sections (339 remains)
  - Select top 50 terms only

  => 339 x 4,029

©Yassine Hamdaoui (2019)

https://towardsdatascience.com/tf-idf-term-frequency-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

Sunoikisis Digital Classics
Session 4. Computational Linguistics

29 October 2020 Leipzig | Monza
YouTube link: https://youtu.be/zjkyZUpvhAQ

Computational Linguistics & Deep Learning for Ancient Texts

Thea Sommerschield
DPhil candidate in Ancient History

It’s a great time to be working on Computational Linguistics and Machine Learning.

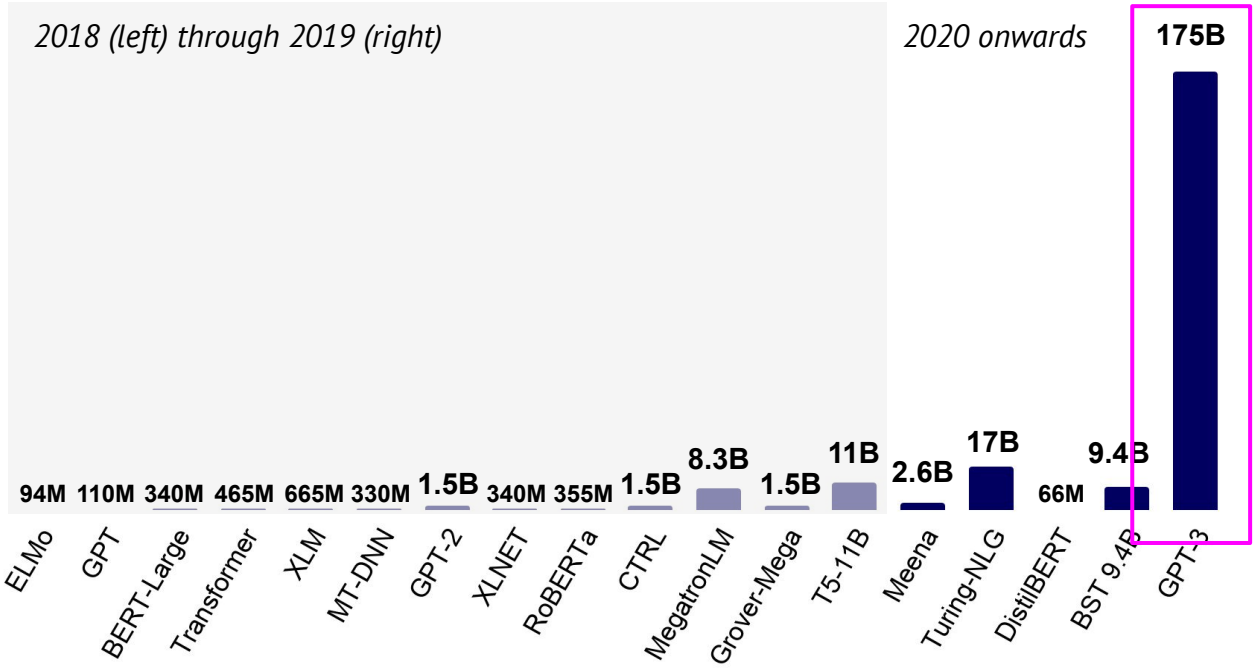
- The *2020 State of AI and Machine Learning* report illustrates the current state of artificial intelligence and machine learning, showcasing where the industry is as a whole in 2020 compared to 2019.
- https://www.stateof.ai/

UNIVERSITY OF OXFORD

Our 2019 Prediction	Grade	Evidence
New natural language processing companies raise $100M in 12 months.	Yes	Gong.io ($200M), Chorus.ai ($45M), Ironscales ($23M), ComplyAdvantage ($50M), Rasa ($26M), HyperScience ($60M), ASAPP ($185M), Cresta ($21M), Eigen ($37M), K Health ($48M), Signal ($25M), and many more!
No autonomous driving company drives >15M miles in 2019.	Yes	Waymo (1.45M miles), Cruise (831k miles), Baidu (108k miles).
Privacy-preserving ML adopted by a F2000 company other than GAFAM (Google, Apple, Facebook, Amazon, Microsoft).	Yes	Machine learning ledger orchestration for drug discovery (MELLODY) research consortium with large pharmaceutical companies and startups including Glaxosmithkline, Merck and Novartis.
Unis build *de novo* undergrad AI degrees.	Yes	CMU graduates first cohort of AI undergrads, Singapore's SUTD launches undergrad degree in design and AI, NYU launches data science major, Abu Dhabi builds an AI university.
Google has major quantum breakthrough and 5 new startups focused on quantum ML are formed.	Sort of	Google demonstrated quantum supremacy in October 2019! Many new quantum startups were launched in 2019 but only Cambridge Quantum, Rahko, Xanadu.ai, and QCWare are explicitly working on quantum ML.
Governance of AI becomes key issue and one major AI company makes substantial governance model change.	No	Nope, business as usual.

Language models: Welcome to the Billion Parameter club

Huge models, large companies and massive training costs dominate the hottest area of AI today, NLP.



Source: https://arxiv.org/pdf/2004.08900.pdf

Note: The number of parameters indicates how many different coefficients the algorithm optimizes during the training process.

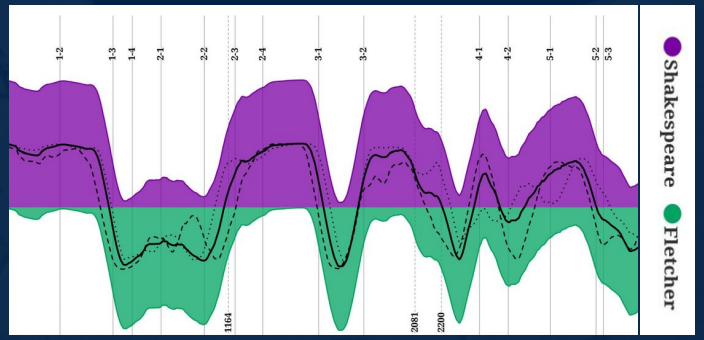
It’s a great time to be working on texts from the Ancient World.

- The 2020 *BES Autumn Colloquium* of the *British Epigraphy Society* features 5/10 digital epigraphy initiatives in its yearly Epigraphic Gazetteer.
- http://www.britishepigraphysociety.org/bes-autumn-colloquium.html

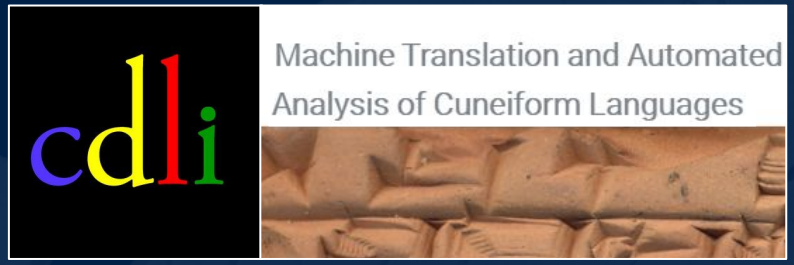
UNIVERSITY OF OXFORD

http://sicily.classics.ox.ac.uk/

https://artsexperiments.withgoogle.com



https://arxiv.org/pdf/1911.05652.pdf



https://cdli.ucla.edu/ & https://cdli-gh.github.io/mtaac/

REKREI

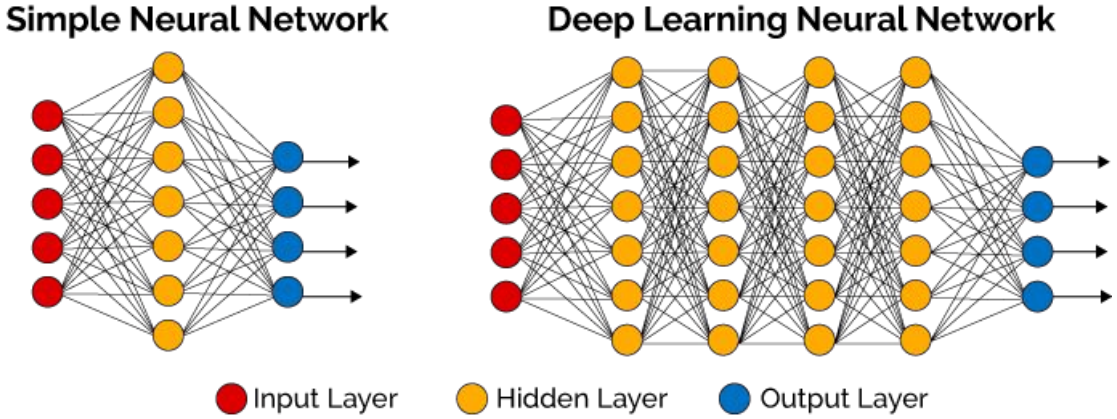
https://rekrei.org/

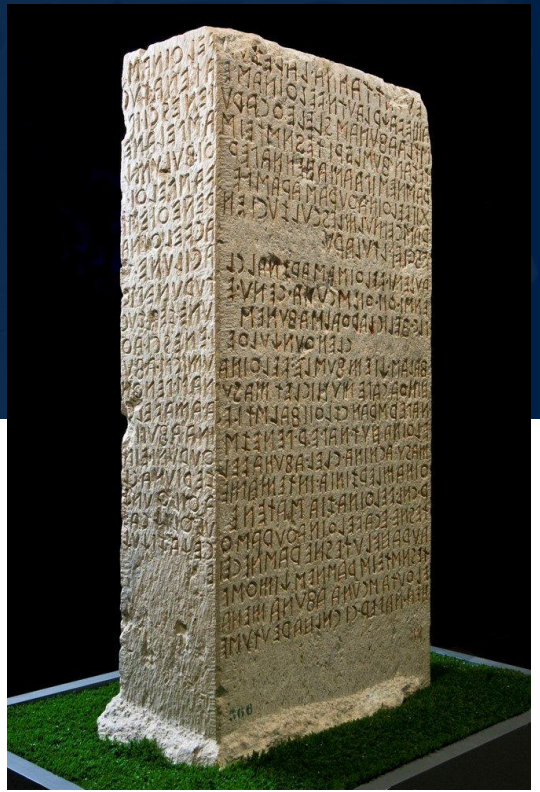
https://arxiv.org/abs/1906.06718

https://hieroglyphicsinitiative.ubisoft.com/

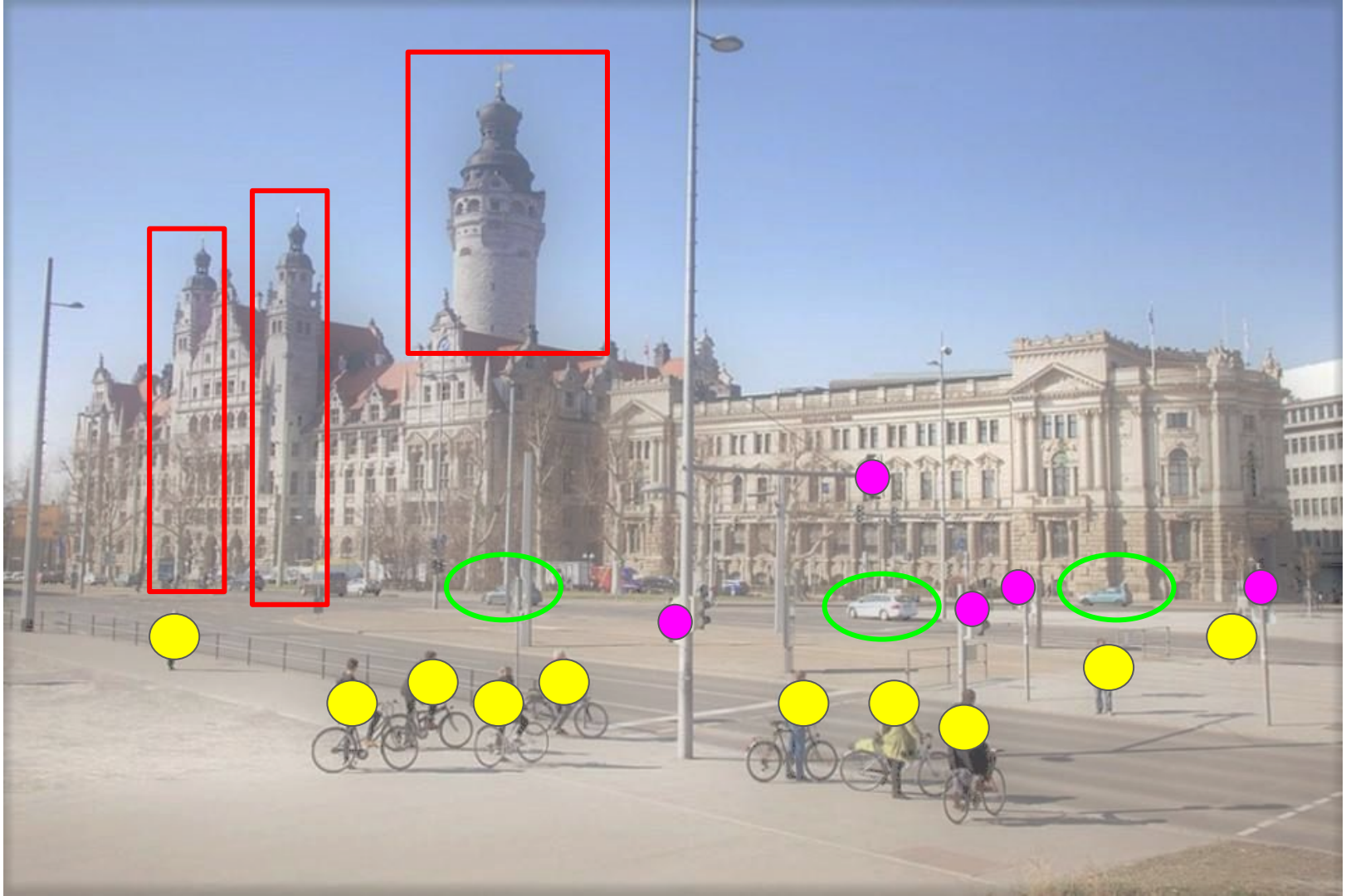
But first, some definitions.

- Machine Learning
- Deep Learning
- Epigraphy (in a bit)



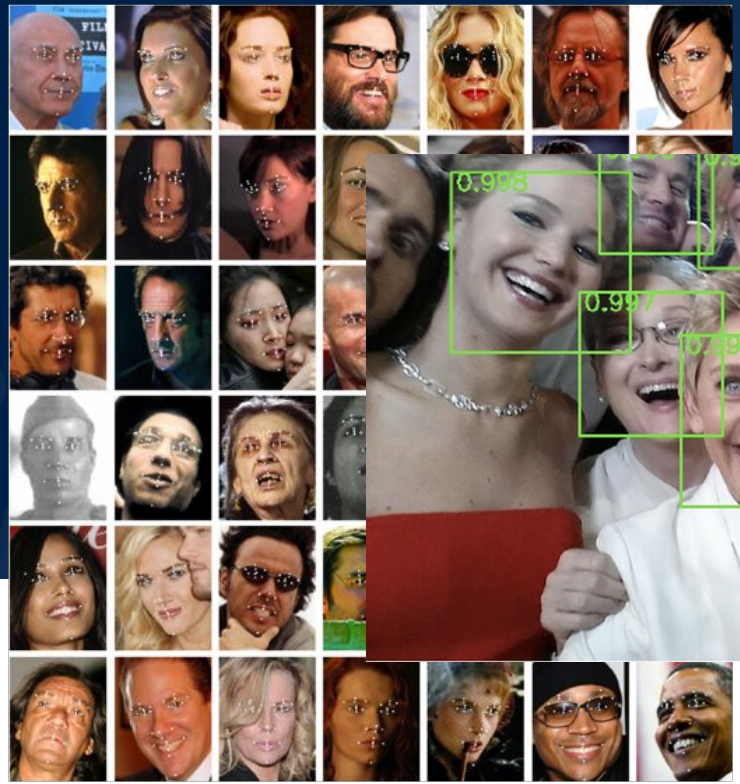


The Cippus Perusinus

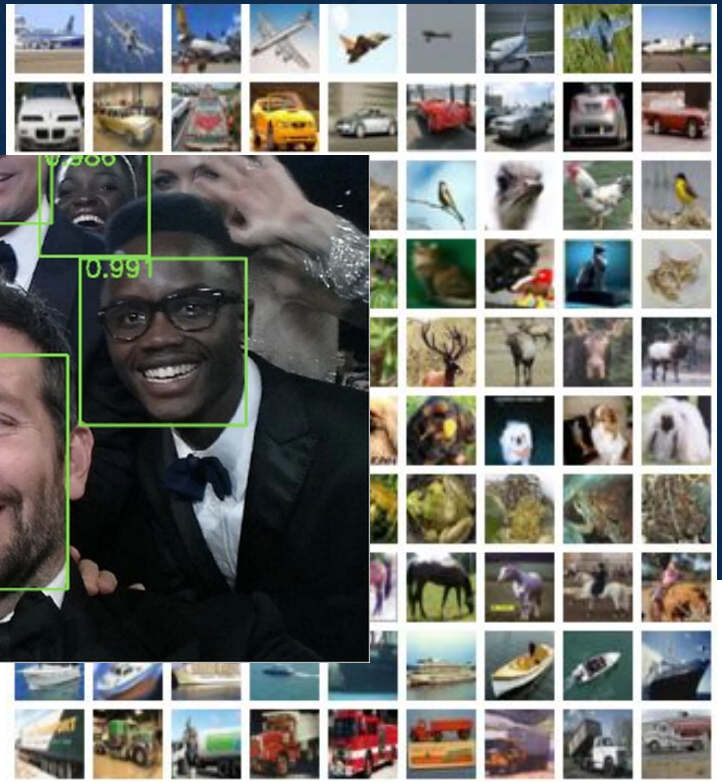


Learning features from data

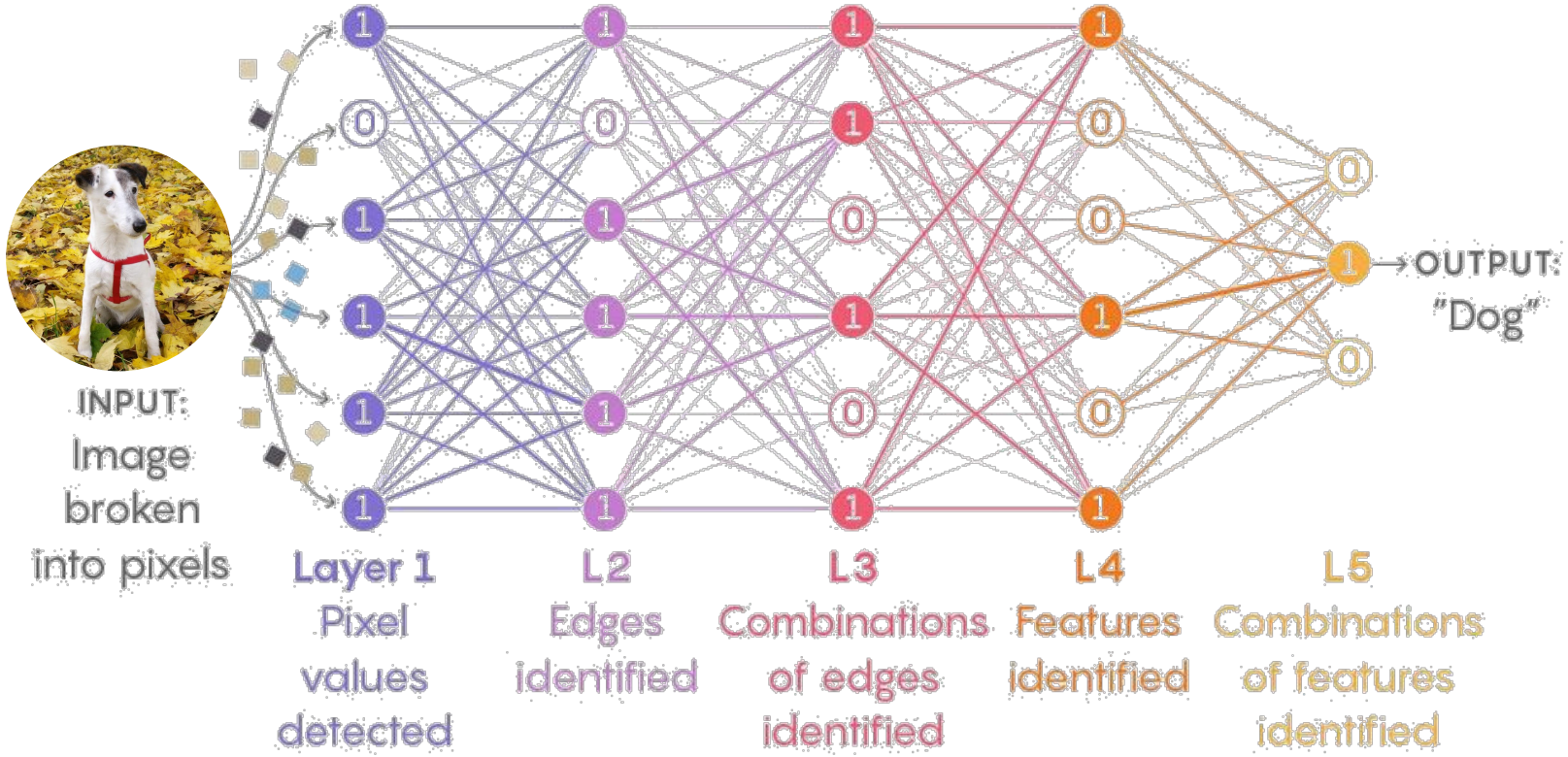
faces



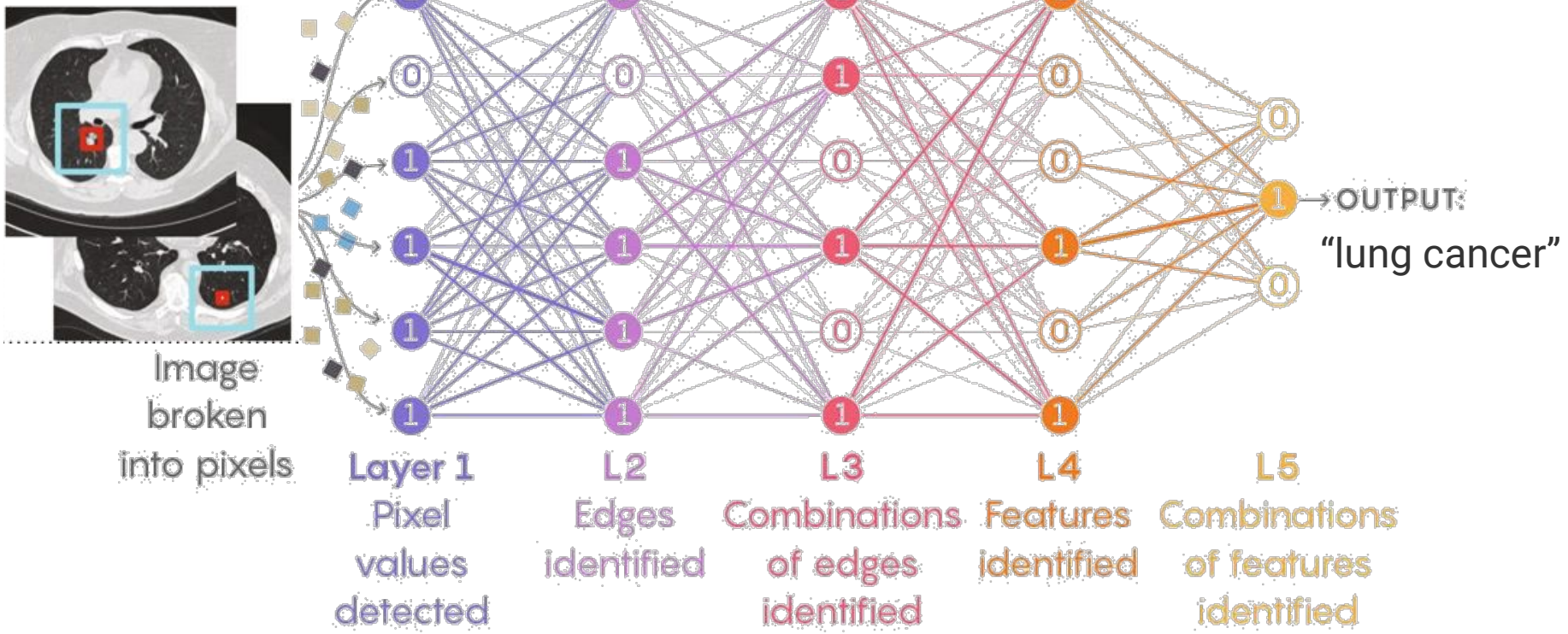
not faces



Neural Networks



Neural Networks



[Ardila et al. 2019]

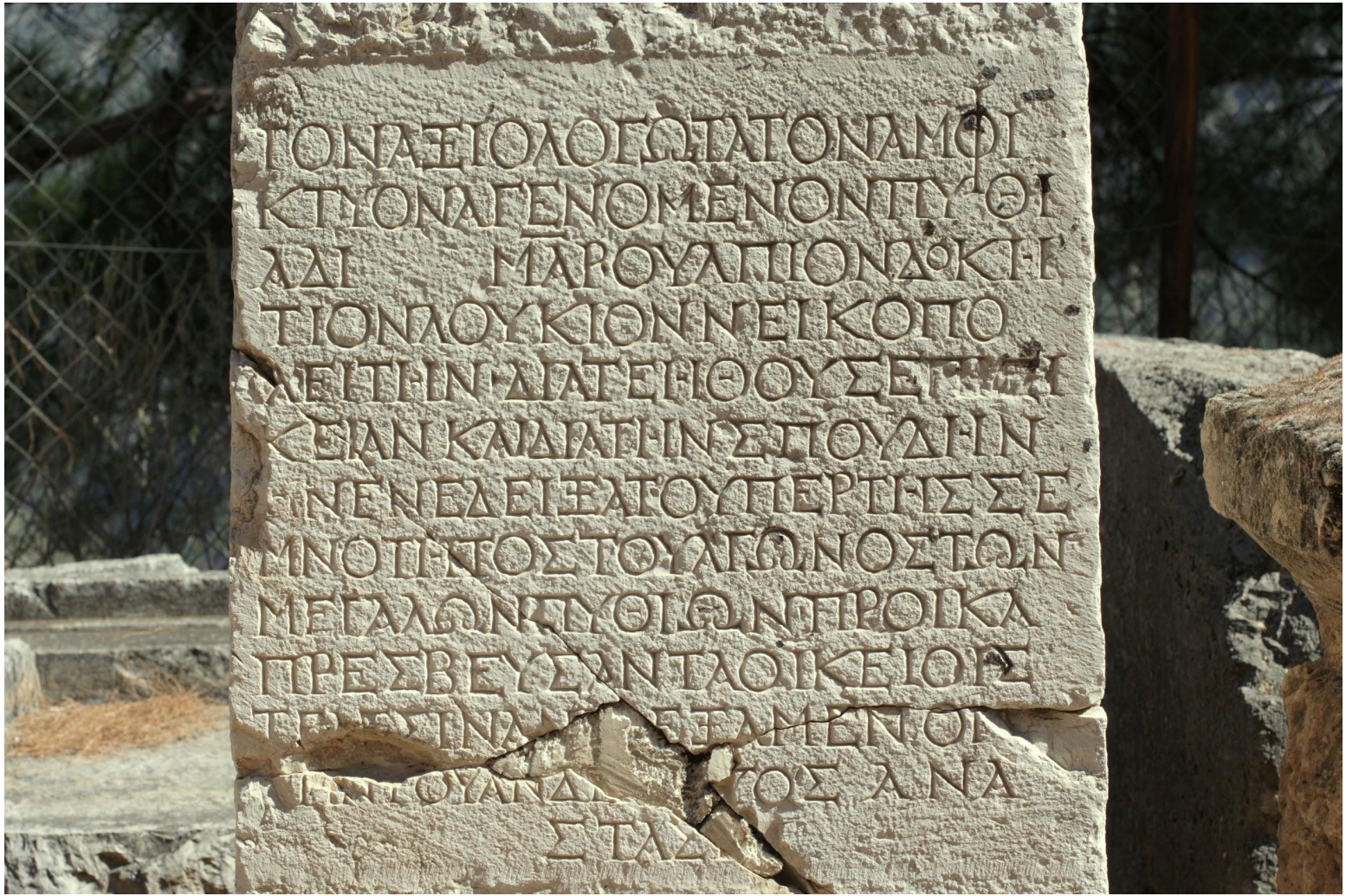
“Restoring ancient text using deep learning: a case study on Greek epigraphy”

Yannis Assael*, Thea Sommerschield*, Jonathan Prag

Association for
Computational Linguistics

EMNLP 2019
Conference on Empirical Methods
in Natural Language Processing

Hong Kong, China
November 3–7, 2019



https://commons.wikimedia.org/wiki/File:Ancient_Greek_inscription,_Delphi_060055.jpg : CC-BY-SA-4.0

Aims:

- A fully automated aid to the epigraphist’s restoration task: a ML model.
- An epigraphic dataset of machine actionable text to train the model.
- Design a model which can be used by all disciplines dealing with ancient texts and which applies to any language.

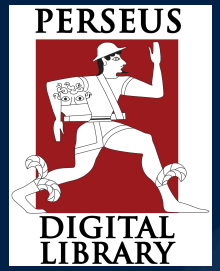
UNIVERSITY OF OXFORD

Why ancient Greek epigraphy?

1. The variability of its contents and context within this dataset makes it an excellent challenge for Machine Learning.
2. Increased availability of digitised data.
3. Personal interest/expertise.

UNIVERSITY OF
OXFORD

Where can I find thousands of digitised Greek texts?



The Packard Humanities Institute
Los Altos, Santa Clarita, Cambridge

A system is as good only as the data it collects!

1. We need a lot of data.
2. We need a lot of clean data.

UNIVERSITY OF
OXFORD

[ ]

Att. — Athens: Agora — stoich. — c. 450 a. — IG I³ 16 — Hesp. 14.1945.82,2; 15.1946.246,77 — SEG 10.11

frg. a

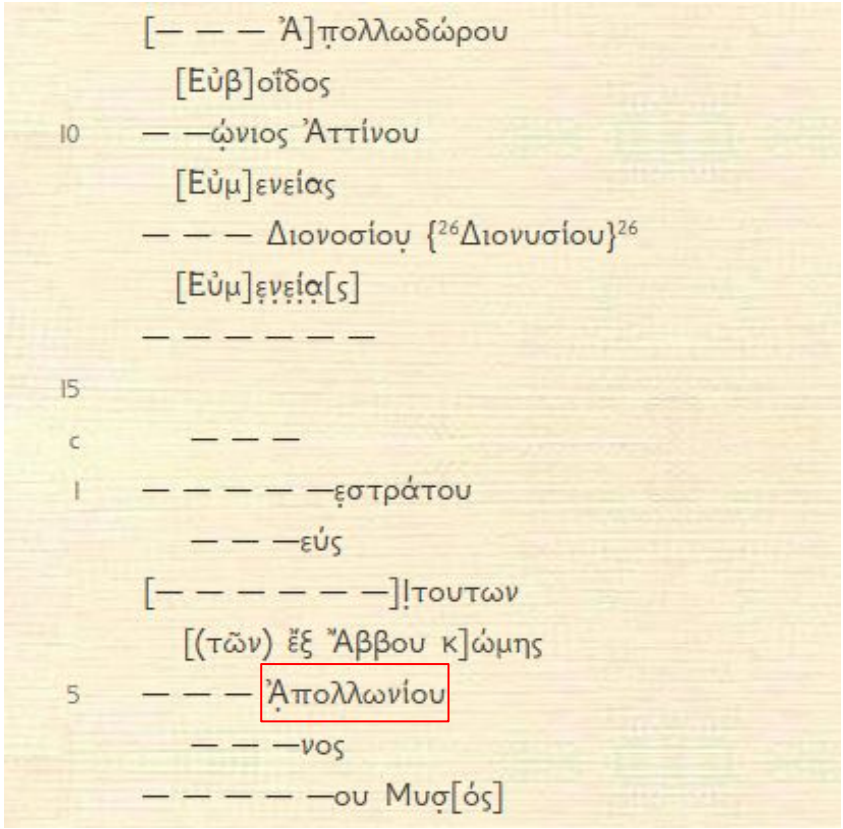
{²reliqiae nonnullae}²

[.]ται ἐὰμ[— — — — — — — — — — —]
[.]ϙθο[.]ατ[— — — — — — — — — — —]
[...]Λ!Ọ#⁷[— — — — — — — — — — —]
[....]τα καὶ [— — — — — — — — — —]
[...7...]εμ[— — — — — — — — — — —]
[..6...]ỌΛẸ[— — — — — — — — — — —]
[..] τὲν βολὲ[ν — — — — — — — — —]
[φ]ρόραρχον [— — — — — — — — — φ]-
[ρ]ορῆν πεντέ[κοντα — — — — — —]
ϛ ἐν ἀκροπόλ[ει — — — — — — — —]
φρόραρχον ϙ[— — — — — — — — — —]
[.]· τὸν δὲ φρ[όραρχον — — — — τ]-
ριάκοντα ἔ[τε γεγονότας(?) — — βο]-
ε̣θ̣ε̣ίαι [.]ϙκ[— — — — — — — — —]

Machine readable text:

1. **Text cleanup:** compute character frequencies, standardise alphabet, strip human annotations and inconsistencies.
2. **Text processing:** '-' for missing characters and '?' for characters to predict; match the number of '-' with those conjectured by epigraphists.

PHI version of MDAI(A) 32 (1907) 428,275



[— — — Ἀ]πολλωδώρου
[Εὐβ]οΐδος
— —ώνιος Ἀττίνου
[Εὐμ]ενείας
— — — Διονοσίου {[26]Διονυσίου}[26]
[Εὐμ]ενείά[ς]
— — — — — —
— — —
— — — — —εστράτου
— — —εύς
[— — — — — —]ιτουτων
[(τῶν) ἔξ Ἄββου κ]ώμης
— — — Ἀπολλωνίου
— — —νος
— — — — —ου Μυσ[ός]

PHI-ML version of MDAI(A) 32 (1907) 428,275

ἀπολλοδώρου εὐβοΐδος --ώνιος ἀττίνου

εὐμενείας --- διονοσίου εὐμενείας

----------εστράτου ---εύς ------ιτουτων

τῶν ἔξ ἄββου κώμης --- **ἀπολλωνίου**

---νος -----ου μυσός

????????ου

The PHI-ML training set:

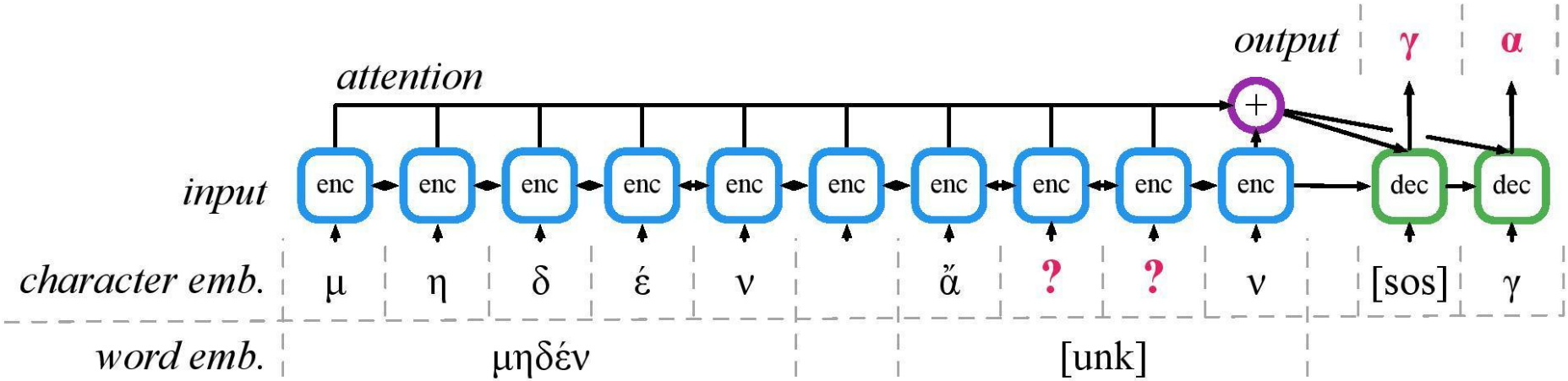
Split	Inscriptions	Words	Chars
Train	34,952	2,792k	16,300k
Valid	2,826	211k	1,230k
Test	2,949	223k	1,298k

Statistics for the PHI-ML corpus.

Architecture design:

- **Input:** sequence of characters
- **Output:** restoration predictions
- **seq2seq** architecture turning input sequence into output sequence
- **LSTM** recurrent neural network
- Works at a **word and a character level**

Pythia:



Pythia BI-WORD processing the phrase **μηδέν ἄγαν.** The letters “γα” are the characters to be predicted, and are annotated with ‘?’. Since ἄ??ν is not a complete word, its embedding is treated as unknown (‘unk’). The decoder outputs correctly “γα”.

Ablation study:

Method	CER	Top-20
Ancient Historian	57.3%	–
LM Philology	68.1%	26.0%
LM Philology & Epigraphy	65.0%	28.8%
LM Epigraphy	52.7%	47.0%
PYTHIA-UNI	42.2%	60.6%
PYTHIA-BI	32.5%	71.1%
PYTHIA-BI-WORD	**30.1%**	**73.5%**

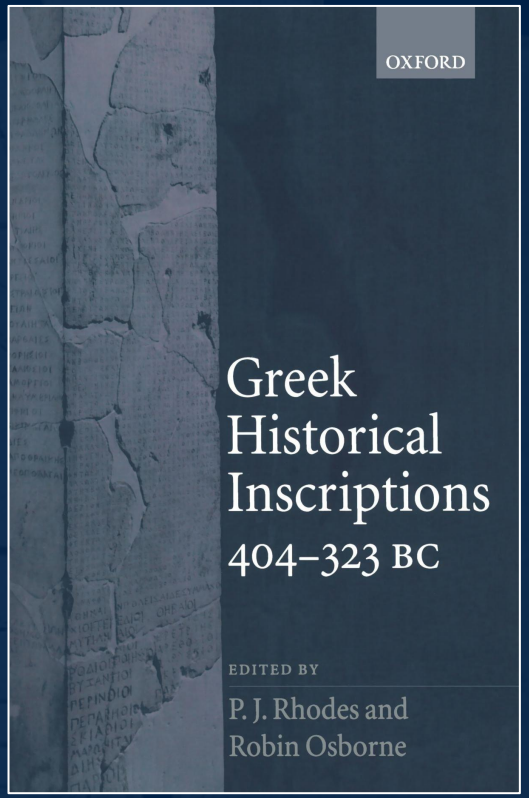
Predictive performance on PHI-ML.

Text restoration demo:

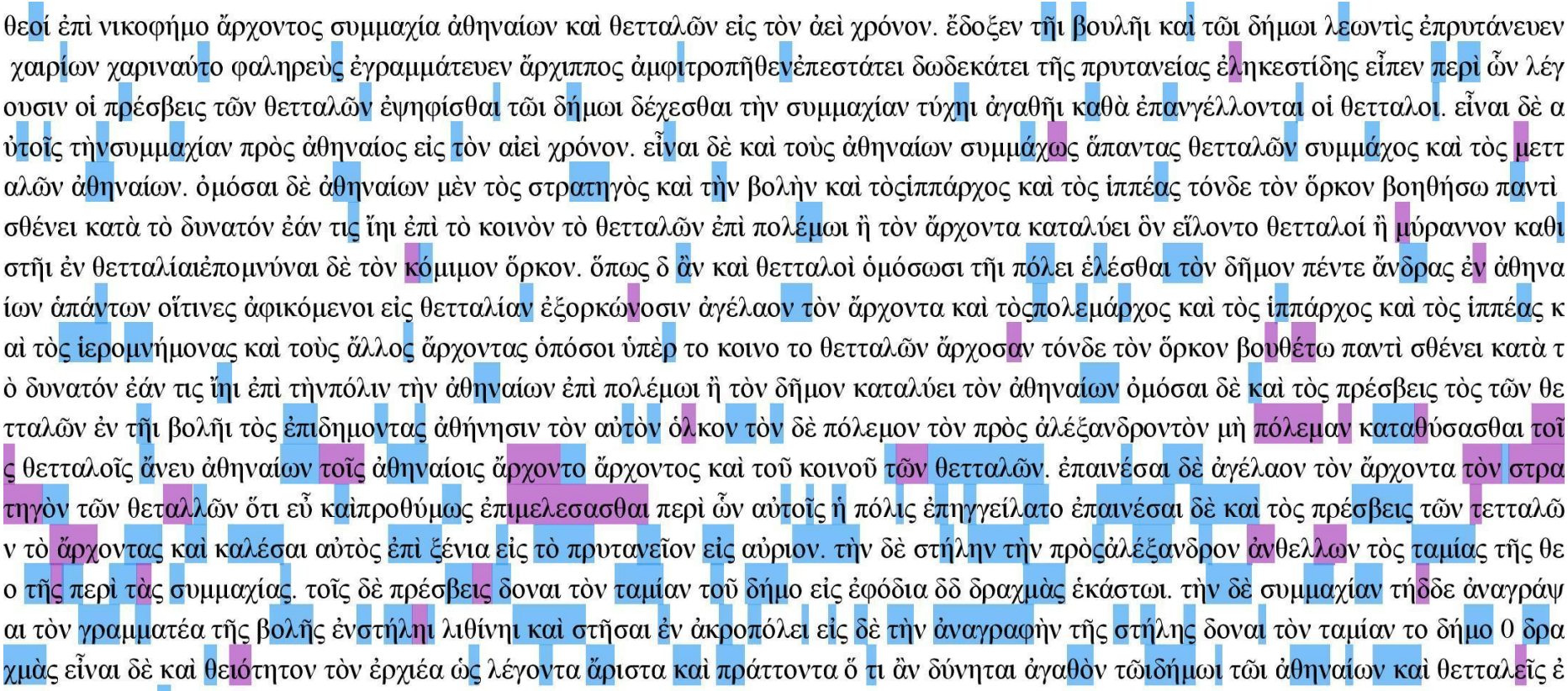
https://colab.research.google.com/drive/16RfCpZLm0M6bf3eGIA7VUPclFdW8P8pZ

UNIVERSITY OF OXFORD

IG II ^{2} 116



UNIVERSITY OF
OXFORD



Sample restoration of the inscription *IG* II ^{2} 116.
Restorations are in blue when correct, purple when incorrect.

Results:

1. Our experimental results illuminate the ways Pythia can assist, guide and advance the epigraphist's task.
2. The combination of Machine Learning and Epigraphy can impact meaningfully the study of inscribed textual cultures, ancient and modern.

Impact and significance:

1. Track textual connections and correlations with computational methods.
2. Design educational tools to explore the data.
3. Make standardised and organised data more accessible.

For more information:

- Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): https://www.aclweb.org/anthology/D19-1668/
- DeepMind blog post: https://deepmind.com/research/publications/Restoring-ancient-text-using-deep-learning-a-case-study-on-Greek-epigraphy
- University of Oxford blog post: http://www.ox.ac.uk/news/arts-blog/restoring-ancient-greek-inscriptions-using-ai-deep-learning
- Financial Times article: https://www.ft.com/content/2b72ed2c-907b-11ea-bc44-dbf6756c871a?FTCamp=engage%2FCAPI%2F%2FChannel_signal%2F%2FB2B

UNIVERSITY OF OXFORD

Digital Humanities
UNIVERSITÄT LEIPZIG

Thank you!

Thea Sommerschield

DPhil candidate in Ancient History

thea.sommerschield@classics.ox.ac.uk

UNIVERSITY OF OXFORD

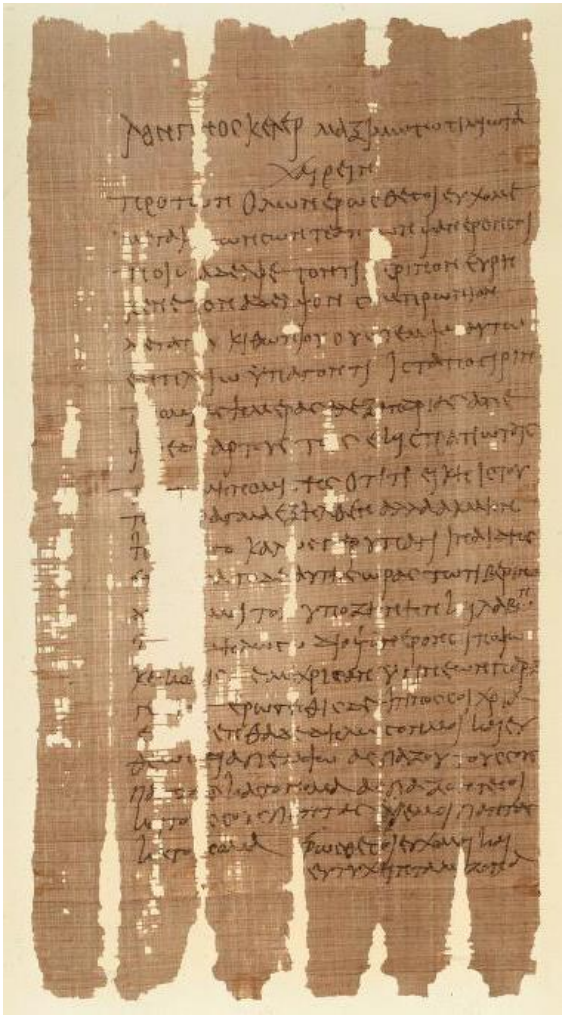
A natural language processing ‘pipeline’ of the papyri

Alek Keersmaekers

University of Leuven
RU Quantitative Lexicology and Variational Linguistics

Introduction: the papyri

- "Documentary" texts from Greek/Roman Egypt (3th century BC – 8th century AD)
- Letters, petitions etc.: closer to everyday language
- More sociolinguistic variation: gender, social class, age, mother tongue etc.
- 4.5 million tokens
- Current linguistic work on papyri: mostly small-scale (e.g. Porter & O'Donnell 2010: 3341 words)



Towards an automatic analysis ‘pipeline’

– All (published) texts are available in a computer-readable (XML) format, but lacking any sort of tokenization, linguistic annotation etc.

```
▼<div xml:lang="grc" type="edition" xml:space="preserve">
  ▼<div n="r,ctr" type="textpart">
    ▼<ab>
        <lb n="1"/>
        ὁ ἐπιφ
        <supplied reason="lost">α</supplied>
        <unclear>ν</unclear>
        έστατος Φωνην βασιλεὺς Βλεμμύων
```

– Supplying missing information via an NLP ‘pipeline’

Main problems

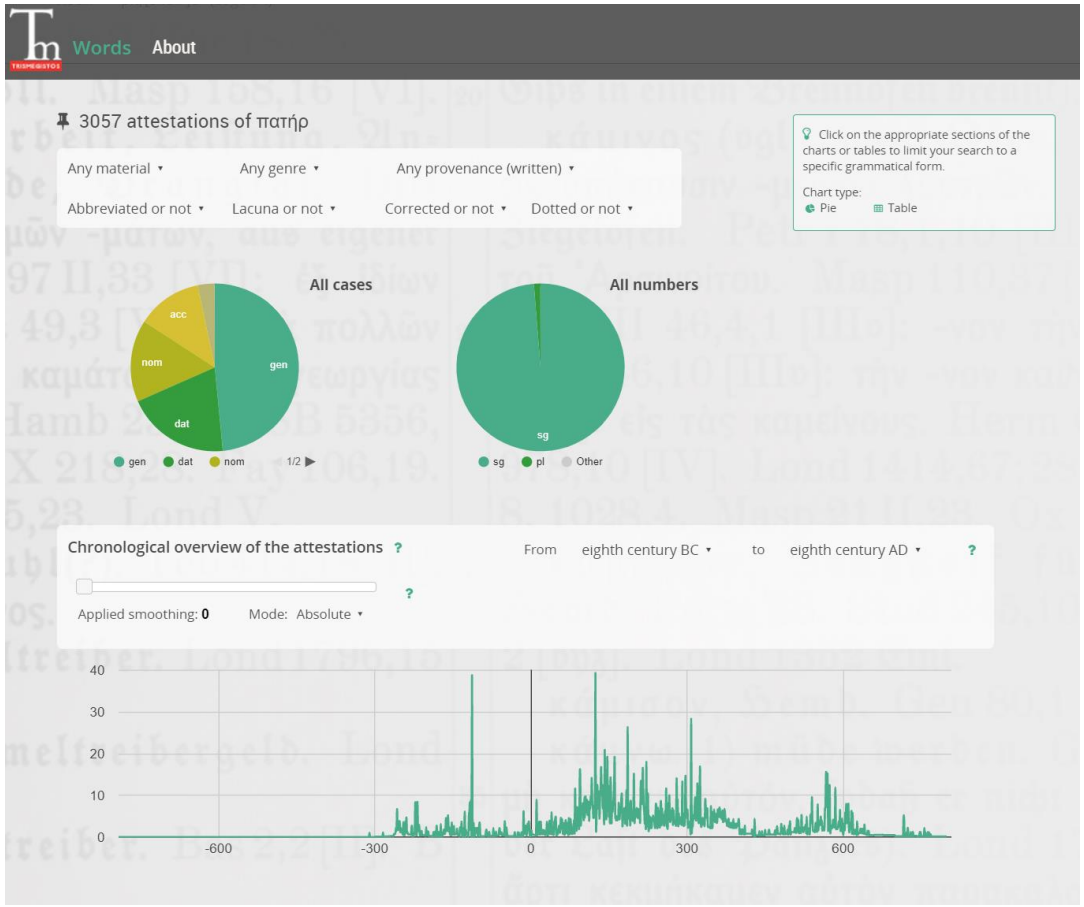
- Greek: highly inflectional, free word order, large rate of elliptic constructions
- Incompletely preserved texts (i.e. ‘gaps’ in some sentences)
- Not a homogeneous spelling
- Mismatch between training corpus and test data

The Trismegistos project

- Collaboration with Trismegistos (Ancient History, KU Leuven) through Mark Depauw (co-supervisor of PhD project)
  - 800BC-800AD
  - ‘Metadata’ (e.g. date, provenance, onomastic data)…
  - … of all non-literary texts (including papyri)
- Linked open data
- *Trismegistos Words*: linguistic database
- Mutually beneficial for linguistics and ancient history

Trismegistos Words

– *https://www.trismegistos.org/words*
– See also *https://github.com/alekkeersmaekers/duke-nlp*



Tokenization

- Tokenization: rule-based approach (spaces and punctuation), some persistent problems (capitalization, multiple separate words written together) resolved semi-automatically
- Representing different text versions (original/editorial) under same token
- Possible to link to existing info in Trismegistos databases (People, Places, Text Irregularities)

ὅπως καὶ γὰρ σοὶ(*) δέ, ἔχω παῖδαν(*) καὶ κἀγώ(*).

^ r,ctr.5. l. σὺ
^ r,ctr.5. l. παῖδα
^ r,ctr.5. l. καὶ ἐγώ

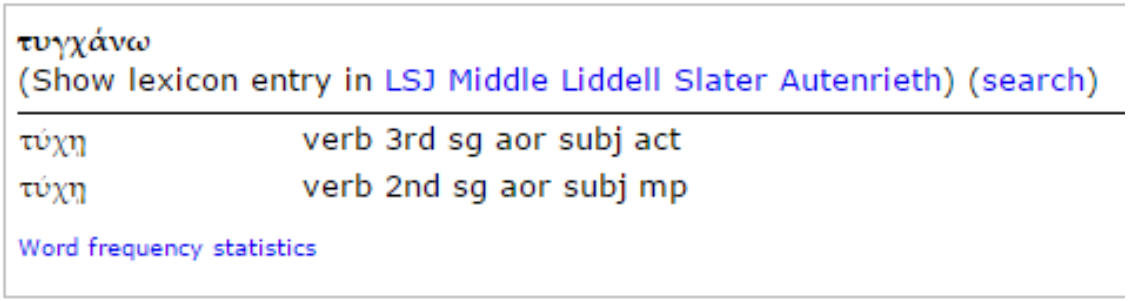
```
<sentence id="5">
  <word id="1" row="5" wordNum="5" hand="m1" regularized="ὅπως" original="ὅπως" compl
  <word id="2" row="5" wordNum="6" hand="m1" regularized="καὶ" original="καὶ" complet
  <word id="3" row="5" wordNum="7" hand="m1" regularized="γὰρ" original="γὰρ" complet
  <word id="4" row="5" wordNum="8" hand="m1" regularized="σὺ" original="σοὶ*" complet
  <word id="5" row="5" wordNum="9" hand="m1" regularized="δέ" original="δέ" completen
  <word id="6" row="5" wordNum="10" hand="m1" regularized="," original="," completene
  <word id="7" row="5" wordNum="11" hand="m1" regularized="ἔχω" original="ἔχω" comple
  <word id="8" row="5" wordNum="12" hand="m1" regularized="παῖδα" original="παῖδαν*"
  <word id="9" row="5" wordNum="13" hand="m1" regularized="καὶ" original="καὶ" comple
  <word id="10" row="5" wordNum="14" hand="m1" regularized="καὶ" original="κἀγώ*" com
  <word id="11" row="5" wordNum="15" hand="m1" regularized="ἐγώ" original="|extra|*"
  <word id="12" row="5" wordNum="16" hand="m1" regularized="." original="." completen
</sentence>
```

Morphological tagging: introduction

- For Ancient Greek: part-of-speech (e.g. noun, verb) + morphology (e.g. genitive, active, … > 33 possible features)
- Typical approach: using lexical probabilities from corpus + N preceding words (e.g. *I* ***can*** *do this*)

Morphological tagging: problems

- Number of possible forms for a given lemma is much higher than for English (e.g. λελυκώς: participle, singular, perfect, active, masculine, nominative)
  - Data sparsity problem: typically inflectional languages are analyzed first formally (in this case via *Morpheus*)
  - Special machine learning algorithms are needed to deal with morphologically complex languages
- Syntax/morphology highly interrelated (e.g. nominative/accusative)
- Irregular spelling > editorial corrections
- Unknown words (e.g. names): extension Morpheus, information Trismegistos databases



Morphological tagging: results

- Overall accuracy: about 94.7% (*RFTagger*, *Mate*)
- Some morphological attributes are easier to tag than others > less ambiguous forms

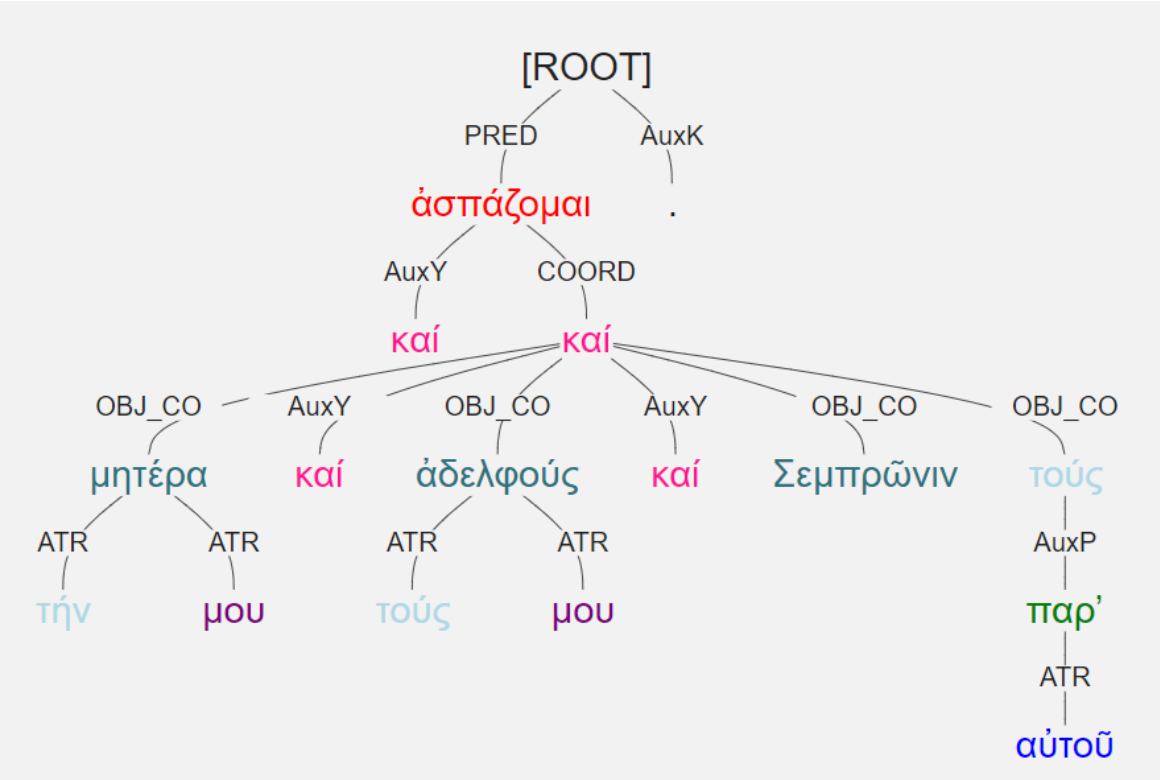
	Median accuracy
Derivative category	0.996
Part-of-speech	0.994
Number	0.990
Voice	0.989
Tense	0.985
Degree	0.974
Case	0.974
Person	0.963
Mood	0.959
Gender	0.951

Morphological tagging: remaining problems

- How to resolve the ~5% remaining problems?
- Some obvious steps in NLP are moving to ‘better’ algorithms (e.g. deep learning) and expanding the training data
- However, taking a deeper look at remaining problems: almost all of them syntactical (e.g. long distance between verb and argument: **ἔστι** γὰρ τὸ πλῆθος τοῦ ἀργυρίου οὐκ **ὀλίον** “Since the sum of the money **is** not **small**”)
- Smaller category of problems: related to world knowledge (e.g. past verbs on –ov: ambiguous between 1 singular and 3 plural)

Syntactic parsing

- For Ancient Greek: dependency parsing
- Two predictions: syntactic head + label



Graph-based parsing

- Initially, all words have a possible dependency link with each other words
- Machine learning is used to determine which dependency link is the most plausible

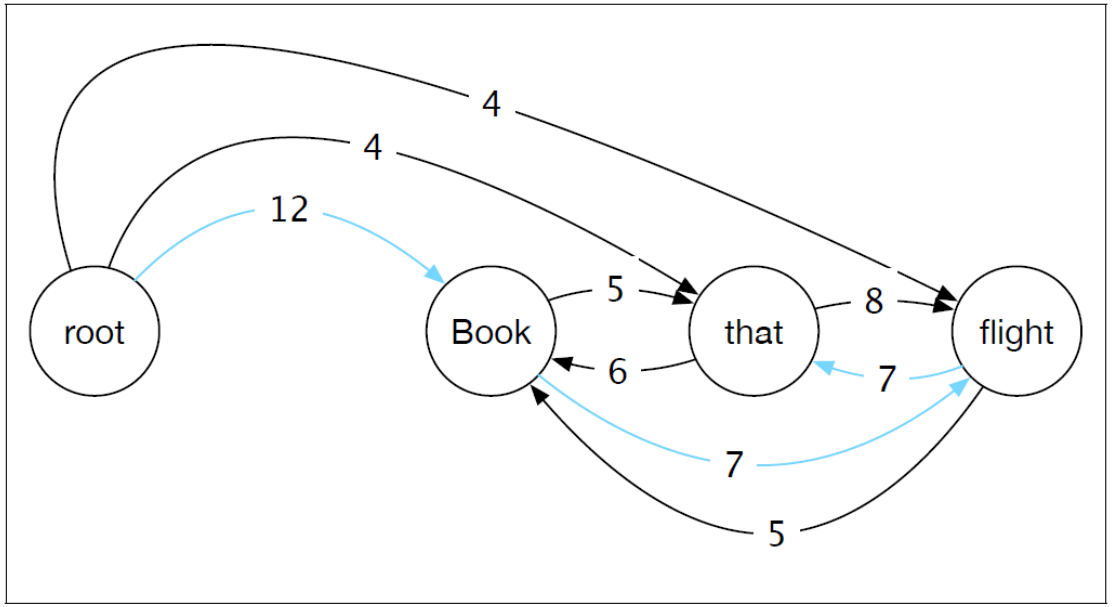


Figure 13.12 Initial rooted, directed graph for *Book that flight*.

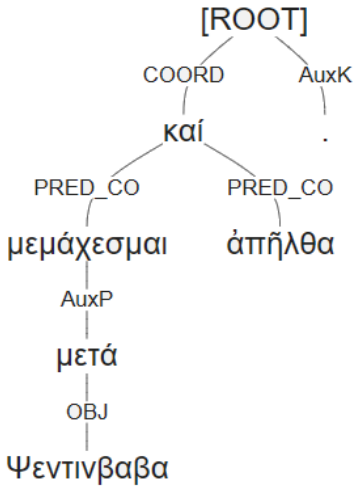
Syntactic parsing: main problems (1)

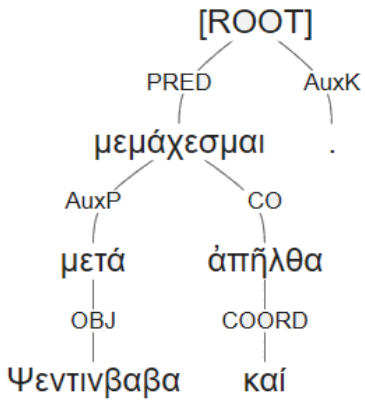
– Major Greek treebanks (e.g. AGDT, Gorman, PROIEL): high rate of inconsistencies (PROIEL even completely different style)

➢ **Homogenization**
  - ➢ Manual
  - ➢ Rule-based
  - ➢ In the future: machine learning techniques

– More difficult structures: e.g. coordination

➢ Modifying annotation format for coordination structures





Syntactic parsing: main problems (2)

– Damaged text

σε(*) ἱμά[τιο]ν χρωμ[ά]τ[ιον ε]ὔσχημον. πληρόθητι(*) δὲ
5 [. . . . .]α τῆς οἰκία[ς . .]οσθης τῇ τιμῇ τοῦ ἱμα-
[τίο]υ [. .]ω . ε . . σ . η[. . . .]ν τὸν Ἀρσινοείτην. μ-

– “Partial parsing” (i.e. feeding such a sentence to the parser like a regular sentence) > some dependencies are still possible to make (e.g. articles with nouns etc.), with OK parsing quality

Syntactic parsing: results

- Stanford's Graph-Based Neural Dependency Parser: accuracy rate of 84.5% (head+label correctly predicted)
- However, not all 'mistakes' are actually wrong
- 'Real' mistakes: often related to semantics and world knowledge

Error type	Frequency
Grammatical mistake	277/500 (55%)
Consistency issue	130/500 (26%)
Annotation error in the test data	48/500 (10%)
Technical (multiple nodes without head)	20/500 (4%)
Damaged text	13/500 (3%)
Ambiguous sentence structure	12/500 (2%)

Syntactic parsing: results by genre

Genre	N tokens	Mean LAS	Median LAS	Std Dev
Papyri	17,609	0.845	-	-
Religion	8,166	0.881	0.873	0.010
Example sentences	1,637	0.870	-	-
Biography	1,445	0.832	0.832	0.001
Epistolography	255	0.828	0.803	0.037
History	14,393	0.825	0.825	0.029
Oratory	4,482	0.822	0.818	0.025
Narrative	2,019	0.804	0.820	0.066
Dialogue	2,329	0.798	0.782	0.025
Philosophical Dialogue	2,431	0.790	0.790	0.040
Philosophy & Science	1,608	0.751	0.758	0.024
Poetry	622	0.740	0.726	0.058

Semantic role labeling

- 29 semantic relations (Pedalion grammar: en.pedalion.org)
- E.g. ἦλθε εἰς **πόλιν** > “direction”
- Features: form (e.g. εἰς+accusative), lemma, animacy, part-of-speech, morphology, syntactic relation, **word vectors/embeddings** (representation of the ‘meaning’ of a word in a series of numbers)
- Training examples: about 12500, from various treebanks (mostly Pedalion treebanks)

Semantic role labeling: results

– Overall accuracy: about 75.7% for all Greek texts

	Accuracy
Religion	0.838 (932/1112)
Documentary	0.809 (1332/1646)
History	0.765 (1439/1881)
Drama	0.751 (1091/1453)
Narrative	0.751 (2019/2689)
Rhetorical	0.723 (1086/1503)
Philosophy	0.714 (1076/1506)
Epic and lyric poetry	0.687 (485/706)

	F1
recipient (1289)	0.909
material (22)	0.842
direction (1006)	0.840
intermediary (16)	0.815
source (803)	0.797
agent (364)	0.785
time (943)	0.777
manner (1596)	0.775
degree (295)	0.768
companion (424)	0.765
location (1436)	0.752
possessor (127)	0.739
duration (221)	0.735
respect (800)	0.720
cause (753)	0.704
beneficiary (715)	0.669
instrument (507)	0.650
time frame (45)	0.603
comparison (198)	0.600
frequency (78)	0.576
experiencer (259)	0.556
goal (282)	0.525
modality (17)	0.333
extent of space (67)	0.278
result (15)	0.222
condition (5)	0.000
property (6)	0.000

Automatic analysis: an example

– τὴν ἐνβολὴν καὶ σφόδρα ὑφ’ ὑμῶν ἀμελουμένην ὁρῶ.

	FORM	LEMMA	POSTAG	HEAD	RELATION	SEMANTIC_ROLE
1	τὴν	ὁ	l-s---fa-	2	ATR	_
2	ἐνβολὴν	ἐνβολή	n-s---fa-	8	OBJ	stimulus
3	καὶ	καί	b--------	7	AuxY	_
4	σφόδρα	σφόδρα	d--------	7	ADV	degree
5	ὑφ’	ὑπό	r--------	7	AuxP	_
6	ὑμῶν	σύ	p-p----g-	5	OBJ	agent
7	ἀμελουμένην	ἀμελέω	v-sppefa-	2	ATR	_
8	ὁρῶ	ὁράω	v1spia---	0	PRED	_
9	.	.	u--------	0	AuxK	_