# Introduction to the CLTK

June 8, 2016

Kyle P. Johnson, PhD kyle@kyle-p-johnson.com http://cltk.org

Linking the Big Ancient Mediterranean University of Iowa, June 6-8, 2016 #BAM2016

## The CLTK's goals ...



 Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)

## The CLTK's goals ...



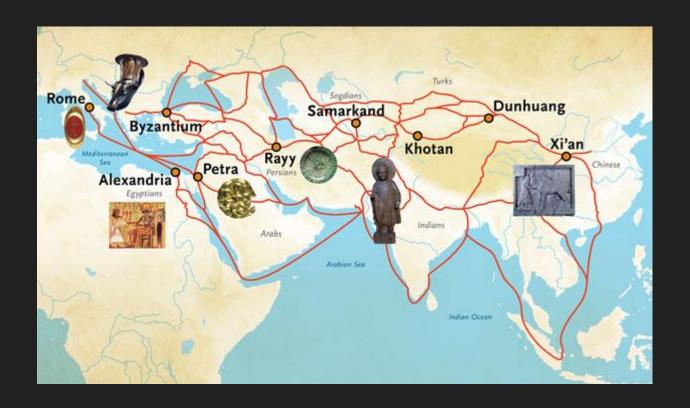
- Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)
- Medium: Quantified Classics

## The CLTK's goals ...



- Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)
- Medium: Quantified Classics
- High: Framework for an integrated study of the ancient world

## ... for a connected ancient world



## By the numbers

- Began 2014
- 1,523 commits
- 24 contributors
- 27 watchers, 103 stars, 80 forks
- 39 people, 18 teams
- 24 releases (with DOI for every release)
- 81% code coverage
- Supports POSIX OS (and partially Windows)
- 2 students, Google Summer of Code
  - Patrick Burns, PhD (ISAW)
  - Suhaib Khan (Netaji Subhas Institute of Technology, Delhi, India); mentored by Luke Hollis of Archimedes Digital)

 Python: A programming language known for its easy-to-read syntax and general friendliness

- Python: A programming language known for its easy-to-read syntax and general friendliness
- NLP: Natural language processing

- Python: A programming language known for its easy-to-read syntax and general friendliness
- NLP: Natural language processing
- Package or "library": Collection of software for a particular set of tasks

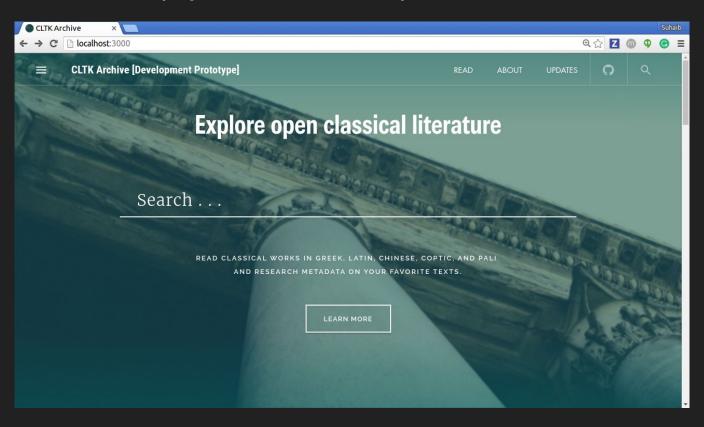
- Python: A programming language known for its easy-to-read syntax and general friendliness
- NLP: Natural language processing
- Package or "library": Collection of software for a particular set of tasks
- NLTK: A prominent NLP package for the Python language

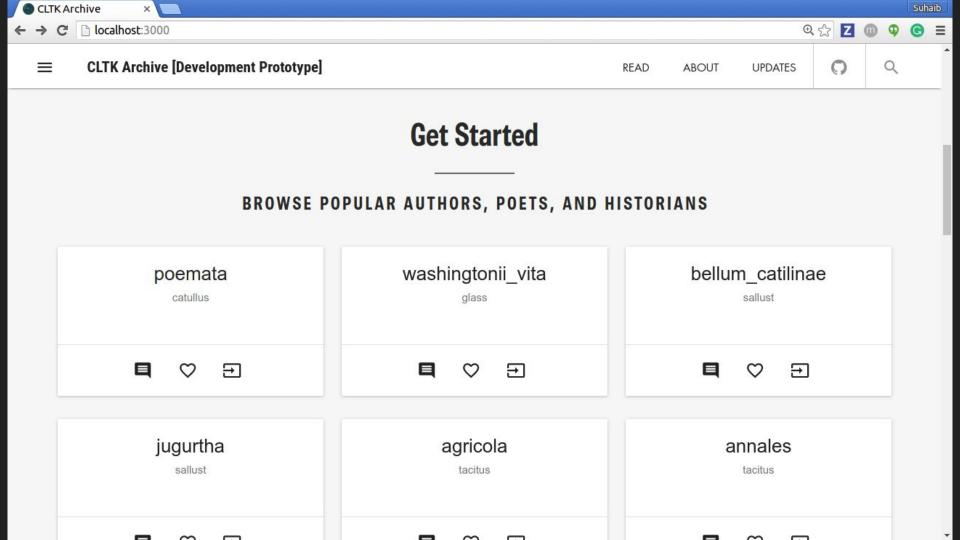
- Python: A programming language known for its easy-to-read syntax and general friendliness
- NLP: Natural language processing
- Package or "library": Collection of software for a particular set of tasks
- NLTK: A prominent NLP package for the Python language
- Git: Software for distributed software development

- Python: A programming language known for its easy-to-read syntax and general friendliness
- NLP: Natural language processing
- Package or "library": Collection of software for a particular set of tasks
- NLTK: A prominent NLP package for the Python language
- Git: Software for distributed software development
- GitHub: A website which makes Git easy

- Python: A programming language known for its easy-to-read syntax and general friendliness
- NLP: Natural language processing
- Package or "library": Collection of software for a particular set of tasks
- NLTK: A prominent NLP package for the Python language
- Git: Software for distributed software development
- GitHub: A website which makes Git easy
- Jupyter (formerly IPython): "Scientific notebooks", an easy way to share code

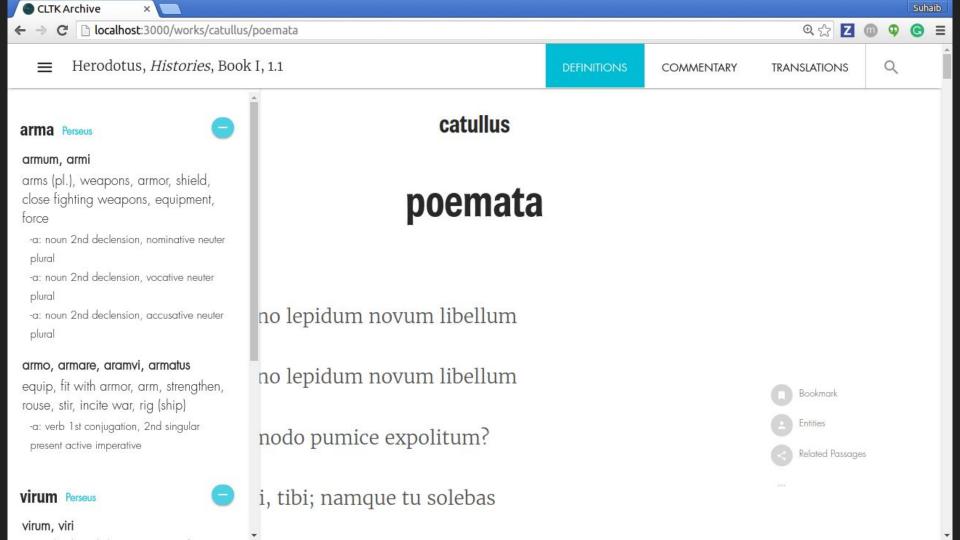
# CLTK Archive (by Luke Hollis)



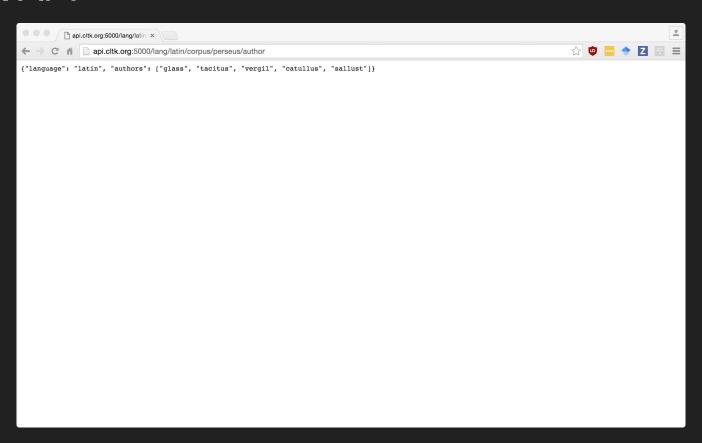








## **CLTK API**













{"language": "latin", "author": "vergil", "text": {"643": "Aeneas (negue enim patrius consistere mentem", "51": "nimborum in patriam, loca feta furentibus austris,", "702": "expediunt, tonsisque ferunt mantelia villis.", "586": "Vix ea fatus erat, cum circumfusa repente", "705": "centum aliae totidemque pares aetate ministri,", "351": "germanae; factumque diu celavit, et aegram,", "288": "Iulius, a magno demissum nomen Iulo.", "576": "adforet Aeneas! Equidem per litora certos", "414": "molirive moram, aut veniendi poscere causas.", "1": "Arma virumque cano, Troiae qui primus ab oris", "719": "insidat quantus miserae deus; at memor ille", "468": "hac Phryges, instaret curru cristatus Achilles.", "307": "explorare novos, quas vento accesserit oras,", "491": "Penthesilea furens, mediisque in milibus ardet,", "365": "Devenere locos, ubi nunc ingentia cernis", "436": "fervet opus, redolentque thymo fragrantia mella.", "17": "hic currus fuit; hoc regnum dea gentibus esse, ", "173": "et sale tabentis artus in litore ponunt.", "562": "\u201cSolvite corde metum, Teucri, secludite curas.", "25": "necdum etiam causae irarum saevique dolores", "324": "aut spumantis apri cursum clamore prementem.\u201d", "32": "errabant, acti fatis, maria omnia circum.", "34": "Vix e conspectu Siculae telluris in altum", "540": "permittit patria? Hospitio prohibemur harenae;", "375": "Nos Troia antiqua, si vestras forte per auris", "415": "Ipsa Paphum sublimis abit, sedesque revisit", "103": "velum adversa ferit, fluctusque ad sidera tollit.", "246": "it mare proruptum et pelago premit arva sonanti.", "151": "tum, pietate gravem ac meritis si forte virum quem", "203": "mittite: forsan et haec olim meminisse iuvabit.", "134": "miscere, et tantas audetis tollere moles?", "137": "Maturate fugam, regique haec dicite vestro:", "670": "Hunc Phoenissa tenet Dido blandisque moratur", "147": "atque rotis summas levibus perlabitur undas.", "159": "Est in secessu longo locus: insula portum", "152": "conspexere, silent, arrectisque auribus adstant:", "142": "Sic ait, et dicto citius tumida aeguora placat,", "53": "luctantes ventos tempestatesque sonoras", "108": "Tris Notus abreptas in saxa latentia torquet\u2014", "187": "Constitit hic, arcumque manu celerisque sagittas", "390": "Namque tibi reduces socios classemque relatam", "441": "Lucus in urbe fuit media, laetissimus umbra,", "47": "et soror et coniunx, una cum gente tot annos", "527": "Non nos aut ferro Libycos populare Penatis", "394": "aetheria guos lapsa plaga Iovis ales aperto", "185": "prospicit errantis; hos tota armenta sequuntur", "311": "arboribus clausam circum atque horrentibus umbris", "673": "Quocirca capere ante dolis et cingere flamma", "131": "Eurum ad se Zephyrumque vocat, dehinc talia fatur:", "624": "Troianae nomenque tuum regesque Pelasqi.", "72": "quarum quae forma pulcherrima Deiopea, ", "218": "spemque metumque inter dubii, seu vivere credant, ", "526": "parce pio generi, et propius res aspice nostras.", "639": "arte laboratae vestes ostroque superbo,", "132": "\u201cTantane vos generis tenuit fiducia vestri?", "81": "Haec ubi dicta, cavum conversa cuspide montem", "445": "egregiam et facilem victu per saecula gentem.", "385": "Europa atque Asia pulsus.\u201d Nec plura querentem", "207": "Durate, et vosmet rebus servate secundis.\u201d", "29": "His accensa super, iactatos aequore toto", "305": "At pius Aeneas, per noctem plurima volvens,", "226": "constitit, et Libyae defixit lumina regnis.", "232": "quid Troes potuere, quibus, tot funera passis,", "507": "Iura dabat legesque viris, operumque laborem", "235": "hinc fore ductores, revocato a sanguine Teucri,", "310": "Classem in convexo nemorum sub rupe cavata", "423": "Instant ardentes Tyrii pars ducere muros,", "238": "Hoc equidem occasum Troiae tristisque ruinas", "426": "Iura magistratusque legunt sanctumque senatum;", "550": "arvaque, Troianoque a sanguine clarus Acestes.", "276": "Romulus excipiet gentem, et Mayortia condet", "572": "Voltis et his mecum pariter considere regnis;", "756": "omnibus errantem terris et fluctibus aestas.\u201d", "294": "claudentur Belli portae; Furor impius intus,", "333": "erramus, vento huc vastis et fluctibus acti:", "583": "omnia tuta vides, classem sociosque receptos.", "535": "cum subito adsurgens fluctu nimbosus Orion", "260": "magnanimum Aenean; neque me sententia vertit.", "36": "cum Iuno, aeternum servans sub pectore volnus,", "93": "ingemit, et duplicis tendens ad sidera palmas", "304": "accipit in Teucros animum mentemque benignam.", "696": "regia portabat Tyriis, duce laetus Achate.", "265": "tertia dum Latio reqnantem viderit aestas,", "720": "matris Acidaliae paulatim abolere Sychaeum", "378": "Sum pius Aeneas, raptos qui ex hoste Penates", "123": "accipiunt inimicum imbrem, rimisque fatiscunt.", "340": "Imperium Dido Tyria regit urbe profecta,", "392": "ni frustra augurium vani docuere parentes.", "607": "In freta dum fluvii current, dum montibus umbrae", "421": "Miratur molem Aeneas, magalia quondam,", "456": "miratur, videt Iliacas ex ordine pugnas,", "499": "exercet Diana choros, quam mille secutae", "14": "ostia, dives opum studiisque asperrima belli;", "618": "alma Venus Phrygii genuit Simoentis ad undam?", "474": "Parte alia fugiens amissis Troilus armis,", "280": "quae mare nunc terrasque metu caelumque fatigat,", "270": "imperio explebit, regnumque ab sede Lavini", "633": "Nec minus interea sociis ad litora mittit", "117": "torquet agens circum, et rapidus vorat aequore vortex.", "63": "et premere et laxas sciret dare iussus habenas.", "57": "sceptra tenens, mollitque animos et temperat iras.", "435": "ignavom fucos pecus a praesepibus arcent:", "252": "prodimur atque Italis longe disiunqimur oris.", "233": "cunctus ob Italiam terrarum clauditur orbis?", "530": "Est locus, Hesperiam Grai cognomine dicunt,", "716": "et magnum falsi implevit qenitoris amorem,", "419": "Iamque ascendebant collem, qui plurimus urbi", "24": "prima quod ad Troiam pro caris gesserat Argis\u2014", "537": "perque undas, superante salo, perque invia saxa", "438": "Aeneas ait, et fastigia suspicit urbis.", "213": "litore aena locant alii, flammasque ministrant.", "524": "Troes te miseri, ventis maria omnia vecti,", "654": "maxima natarum Priami, colloque monile", "242": "Antenor potuit, mediis elapsus Achivis,", "718": "haeret et interdum gremio fovet, inscia Dido, ", "106": "Hi summo in fluctu pendent; his unda dehiscens", "660": "incendat reginam, atque ossibus implicet ignem;", "368": "taurino quantum possent circumdare tergo.", "315": "virginis os habitumque gerens, et virginis arma", "577": "dimittam et Libyae lustrare extrema iubebo,", "608": "lustrabunt convexa, polus dum sidera pascet,", "722": "iam pridem resides animos desuetaque corda.", "475": "infelix puer atque impar congressus Achilli,", "298": "ut terrae, utque novae pateant Karthaginis arces", "240": "nunc eadem fortuna viros tot casibus actos", "565": "Quis genus Aeneadum, quis Troiae nesciat urbem,", "381": "Bis denis Phrygium conscendi navibus aequor,", "400": "aut portum tenet aut pleno subit ostia velo.", "13": "Karthago, Italiam contra Tiberinaque

longe", "73": "conubio iungam stabili propriamque dicabo,", "165": "desuper horrentique atrum nemus imminet umbra.", "251": "navibus (infandum!) amissis, unius ob

Disintermediation

- Independent of academic bureaucracy
- Software direct into researchers' hands

Disintermediation

- Independent of academic bureaucracy
- Software direct into researchers' hands

**Decentralization** 

- Distributed by Git
- Not "pet project" of one person of many!

Disintermediation	<ul> <li>Independent of academic bureaucracy</li> <li>Software direct into researchers' hands</li> </ul>
Decentralization	<ul> <li>Distributed by Git</li> <li>Not "pet project" of one person – of many!</li> </ul>
Transparency	<ul><li>Public development on GitHub</li><li>Public, readable code</li></ul>

Disintermediation	<ul> <li>Independent of academic bureaucracy</li> <li>Software direct into researchers' hands</li> </ul>
Decentralization	<ul> <li>Distributed by Git</li> <li>Not "pet project" of one person – of many!</li> </ul>
Transparency	<ul><li>Public development on GitHub</li><li>Public, readable code</li></ul>
Standardization	<ul><li>Scientific reproducibility</li><li>Good basic texts, but editable</li></ul>

Extensibility

- Accepting of any proven NLP algorithms
- 100% NLP coverage of all ancient langs

Extensibility

- Accepting of any proven NLP algorithms
- 100% NLP coverage of all ancient langs

Multi-disciplinary

- Academic depts, CS, faith traditions
- Intersection of industry & academe



Extensibility	<ul> <li>Accepting of any proven NLP algorithms</li> <li>100% NLP coverage of all ancient langs</li> </ul>
Multi-disciplinary	<ul> <li>Academic depts, CS, faith traditions</li> <li>Intersection of industry &amp; academe</li> </ul>
Mutual benefit	<ul><li>Full public record of all commits</li><li>Researchers develop own work</li></ul>
Inclusion	<ul><li>Collaborative, encouraging</li><li>Free, easy communication</li></ul>

Free & open source

- Fork, modify, merge ... whatever!
- MIT licence (OK for commercial use)

- Make observations
- 2. Think of interesting questions
- 3. Formulate hypotheses
- 4. Develop testable predictions
- 5. Gather data to test predictions
- 6. Develop general theories (repeat)

- Make observations
- 2. Think of interesting questions
- 3. Formulate hypotheses
- 4. Develop testable predictions
- 5. Gather data to test predictions
- 6. Develop general theories (repeat)



- Peer review
- 2. Documentation
- 3. Reproducibility
  - Archiving
  - Data sharing

- 1. Make observations
- 2. Think of interesting questions
- 3. Formulate hypotheses
- 4. Develop testable predictions
- 5. Gather data to test predictions
- 6. Develop general theories (repeat)



2. Documentation

Reproducibility

- Archiving
- Data sharing



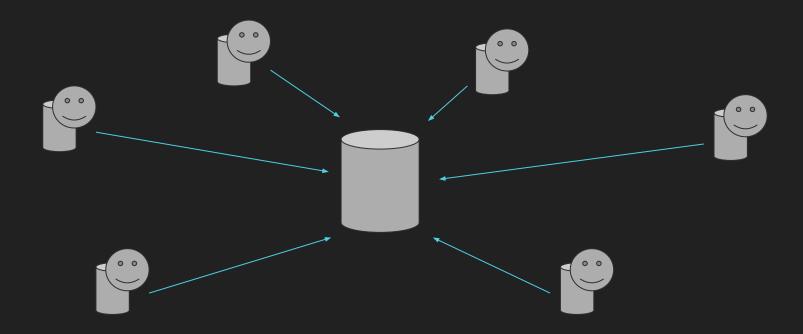
- Make observations
- 2. Think of interesting questions
- 3. Formulate hypotheses
- 4. Develop testable predictions
- 5. Gather data to test predictions
- 6. Develop general theories (repeat)

- Peer review
- 2. Documentation
- 3 Reproducibility
  - Archiving
  - Data sharing

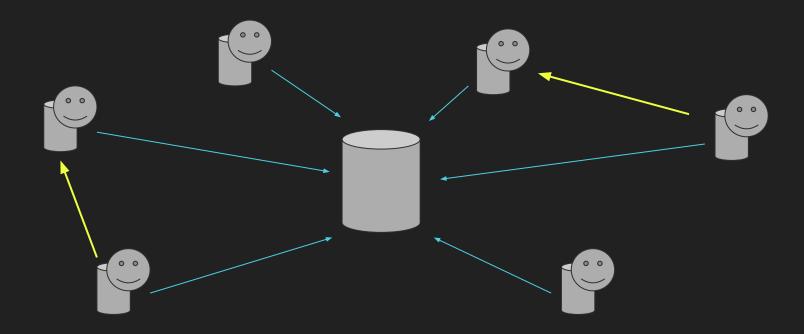
#### Data sets should be:

- Versioned
- Author-attributed
- Auditable
- Editable
- Easily obtained

# Technical organization: Repositories



# Technical organization: Repositories



## Technical organization: Core vs. Data

- CLTK Core software
  - Led by programmers
  - Coordinates data processing
  - Downloads and installs data repositories

## Technical organization: Core vs. Data

- CLTK Core software
  - Led by programmers
  - Coordinates data processing
  - Downloads and installs data repositories
- Linguistic data repositories
  - Led by language experts
  - Plaintext corpora
  - Trained models (for machine learning)
  - Dictionaries, word lists
  - Tagged texts (for part-of-speech, dependency grammar)

# Technical organization: Core vs. Data

#### CLTK Core software

- Led by programmers
- Coordinates data processing
- Downloads and installs data repositories

#### Linguistic data repositories

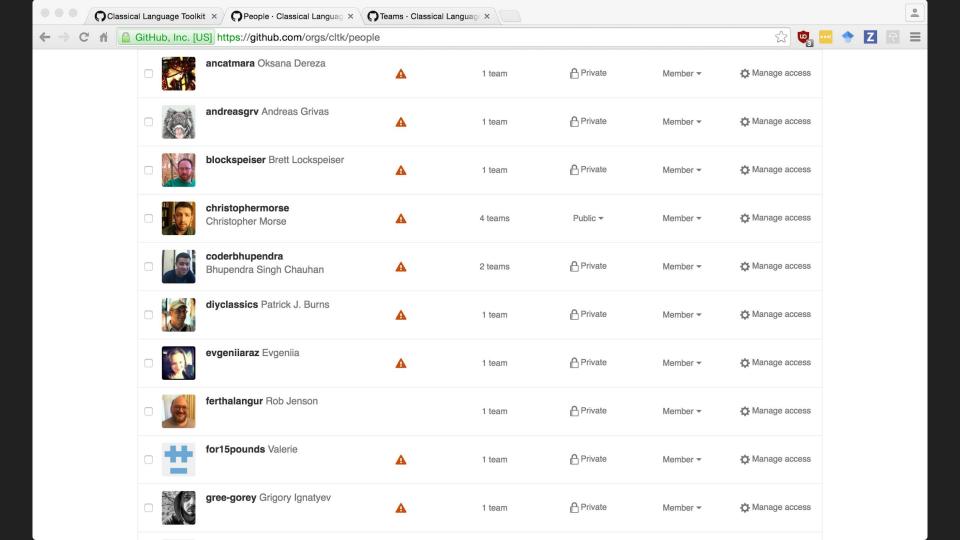
- Led by language experts
- Plaintext corpora
- Trained models (for machine learning)
- Dictionaries, word lists
- Tagged texts (for part-of-speech, dependency grammar)

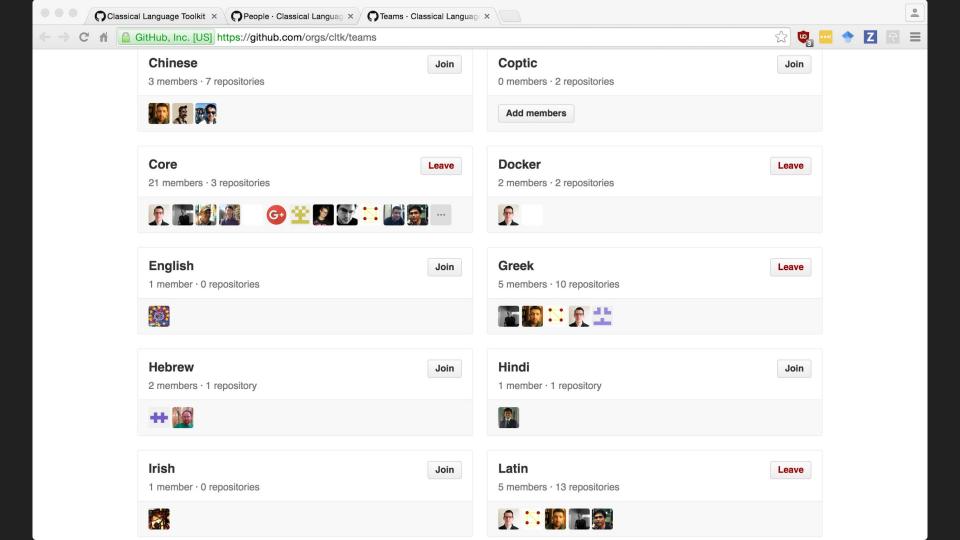
#### CLTK Archive and API

- Reading environment, with NLP and research extras
- Totally led by Luke Hollis

# Personnel organization: People and Teams

- CLTK organization on GitHub
  - o 39 People
  - o 18 Groups
    - Languages, plus Core, API, website,
  - Admins (a few, mostly housekeeping)
  - Unaffiliated Contributors





# (Really quick) quickstart

Make virtualeny and download core

```
o $ pyvenv venv
o $ source venv/bin/activate
o $ pip install cltk
```

#### Download and import corpora

## Setup for PHI and TLG corpora

- PHI5, PHI7, and TLG\_E
  - Not downloaded, but imported from local files
  - o >>> ci.import corpus('tlg', '~/Documents/corpora/TLG E/')
  - Makes copy of corpus at ~/cltk data/originals
- Convert TLG from Beta Code into Unicode
  - o >>> from cltk.corpus.greek.tlgu import TLGU
  - $\circ$  >>> t = TLGU()
  - o >>> t.convert corpus(corpus='tlg')
  - o >>> t.convert corpus(corpus='phi5')
  - Makes copy of corpus in ~cltk\_data/greek/text/tlg or

```
~/cltk_data/latin/text/phi5
```

## NLP for all languages

- Concordance
- Information retrieval
  - Plain and regex searching
  - Robust boolean search on the way
- n-gram: 'Ut primum nocte discussa sol'
  - bigrams: ('ut', 'primum'), ('primum', 'nocte'), ('nocte', 'discussa'), ('discussa', 'sol')
  - o trigrams: ('ut', 'primum', 'nocte'), ('primum', 'nocte', 'discussa'), ('nocte', 'discussa', 'sol')
  - o 5-grams: ('ut', 'primum', 'nocte', 'discussa', 'sol')
- Word frequencies
  - simple count for a word
  - complete reports for a text
- Word tokenization (via NLTK)

#### NLP for Greek and Latin

- Text normalization
  - $\circ$  j  $\rightarrow$  i, v  $\rightarrow$  u (Latin)
  - Beta Code conversion (for legacy Greek texts)
  - TLG and PHI corpus specific (remove formatting)
  - Unicode normalization
- Sentence tokenizer
- Lemmatizer
- Stemmer (Latin)
- Word tokenizer, for enclitics (Latin)
- Stopword filtering

## NLP for Greek and Latin (cont'd.)

- Named Entity Recognition (NER)
- Part-of-speech (POS) tagger
  - From Perseus/Alphaeus treebank
  - Great work remaining to be done, convert their codes to others (Brill, PROIEL, etc)
- Dependency grammar tagger # In progress!
- Prosody scanner
- Syllabifier (Greek)
- TLG and PHI5 indices
  - File to author, genre to authors, date to authors, gender to authors, etc.
- Word2Vec

## Beyond Greek and Latin

- ~60 repos at <a href="https://github.com/cltk">https://github.com/cltk</a>
- Chinese, Coptic, Pali, Tibetan, Middle English, Telugu, Classical Hindi, Sanskrit, Hebrew, Aramaic
  - o 2.5 GB (!) of Chinese Buddhist texts
  - Coptic texts (via Coptic Scriptorium)
  - Pali Tipitaka
  - Tibetan POS tagged texts and a lexicon
  - Parallel corpora ready for statistical machine translation (hint, hint)
  - Corpus of ~50 million Hebrew words, ~20 million Aramaic (via Sefaria)
  - Entirety of Perseus/Open Philology

### Citation

- Developed by <u>many talented contributors!</u>
- BibTex

```
O @Misc{johnson2014,
    author = {Kyle P. Johnson et al.},
    title = {CLTK: The Classical Language Toolkit},
    howpublished = {\url{https://github.com/cltk/cltk}},
    note = {{DOI} 10.5281/zenodo.<current_release_id>},
    year = {2014--2016},
}
```

Chicago author-date

Note: Current DOI release id available at: https://github.com/cltk/cltk

#### Resources

- This and other lectures
- Core software: <a href="https://github.com/cltk/cltk">https://github.com/cltk/cltk</a>
  - Bug tracking: <a href="https://github.com/cltk/cltk/issues">https://github.com/cltk/cltk/issues</a>
    - Beginners' issues labeled easy
- Project repositories: <a href="https://github.com/cltk">https://github.com/cltk</a>
- Docs: <a href="http://docs.cltk.org">http://docs.cltk.org</a>
  - Installation: <a href="http://docs.cltk.org/en/latest/installation.html">http://docs.cltk.org/en/latest/installation.html</a>
- Python + Command line basics
  - Intro to the command line: <a href="http://blog.teamtreehouse.com/introduction-to-the-mac-os-x-command-line">http://blog.teamtreehouse.com/introduction-to-the-mac-os-x-command-line</a>
  - Python installation: <a href="https://www.python.org/downloads">https://www.python.org/downloads</a> (choose 3.5)
  - Good self-paced Python lessons: <a href="http://learnpythonthehardway.org">http://learnpythonthehardway.org</a>

#### Contribute & contact

- Classical Language Toolkit
  - Home: <a href="http://cltk.org">http://cltk.org</a>
  - Docs: <a href="http://docs.cltk.org/en/latest">http://docs.cltk.org/en/latest</a>
  - Source: <a href="https://github.com/cltk/cltk">https://github.com/cltk/cltk</a>
  - Corpora: <a href="https://github.com/cltk">https://github.com/cltk</a>
  - Import module: <a href="https://github.">https://github.</a>
     com/cltk/cltk/blob/master/cltk/corpus/utils/importer.py
- Contribute
  - Issue tracking: <a href="https://github.com/cltk/cltk/issues">https://github.com/cltk/cltk/issues</a>
  - Other questions: <u>kyle@kyle-p-johnson.com</u>

#### Sources

- Images
  - http://www.penn.museum/silkroad/exhibit\_silkroad.php
- Git
  - o GitPython: https://github.com/gitpython-developers/GitPython
  - https://en.wikipedia.org/wiki/Git\_(software)
- Science
  - https://en.wikipedia.org/wiki/Scientific\_method
  - https://en.wikipedia.org/wiki/Reproducibility