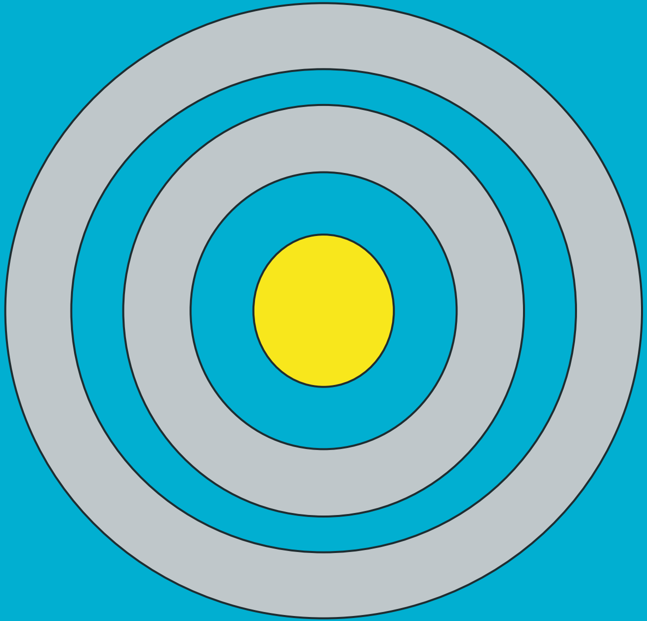
Scientific data set management: A lesson learned from building the Classical Language Toolkit

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

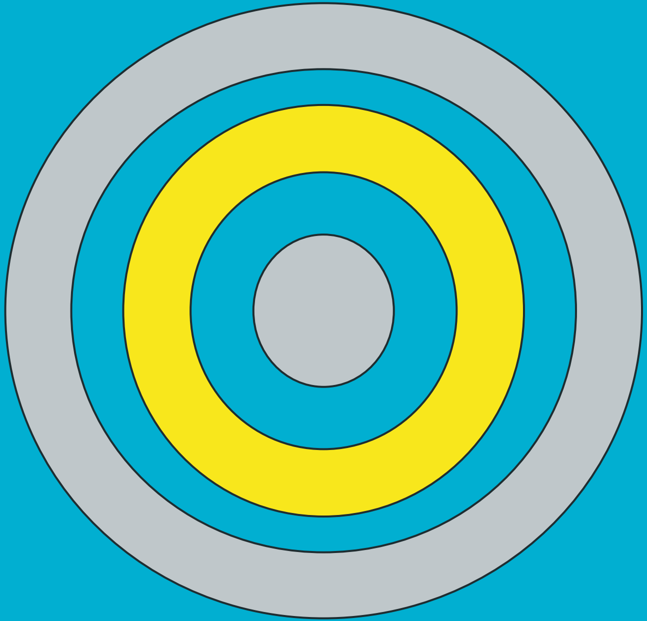
SF Python Meetup Group
@Yelp
February 10, 2016

The CLTK’s goals …



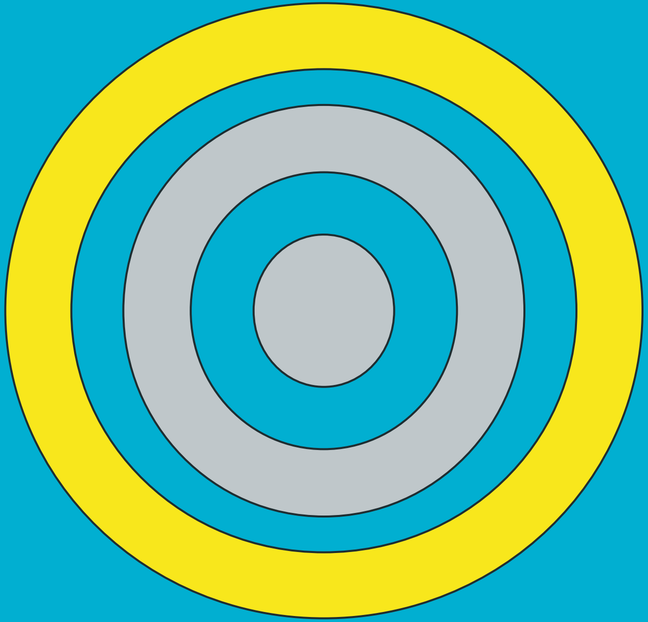
- Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)

The CLTK's goals …



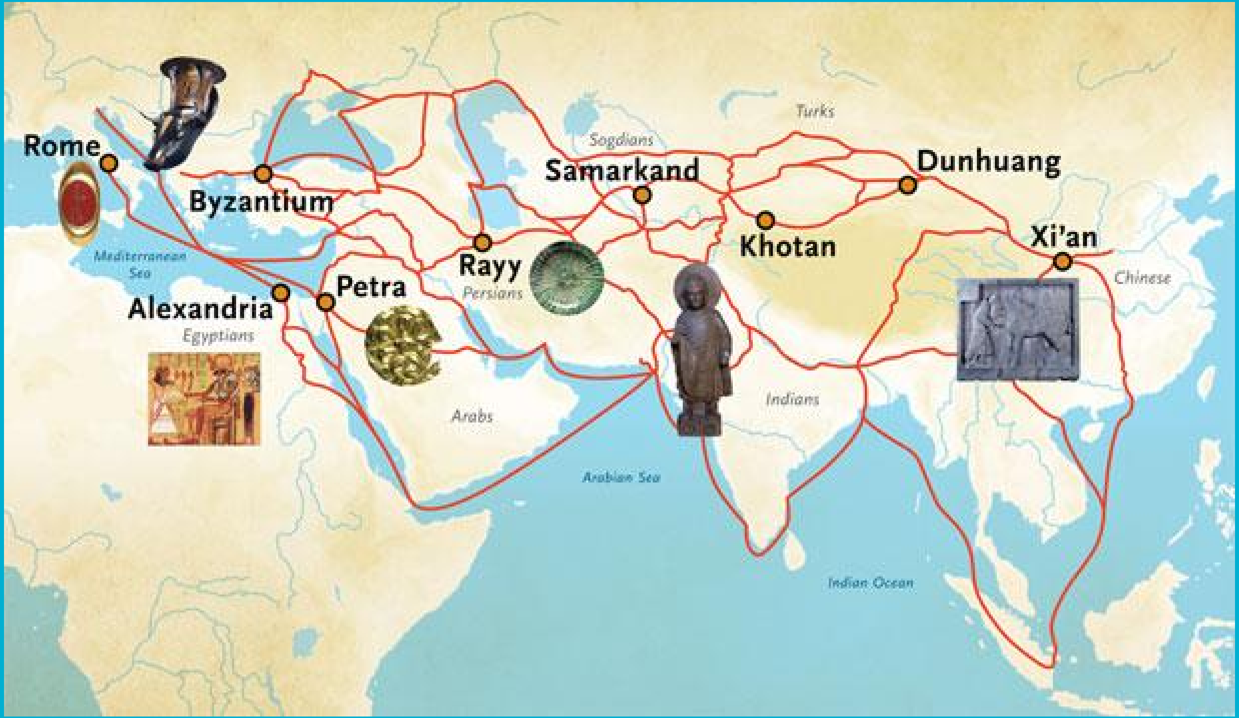
- Low: Good datasets for NLP of ancient languages (Egyptian hieroglyphs, Ancient Greek, Latin, Hebrew, Sanskrit, Tibetan, Classical Chinese, etc.)
- Medium: Quantified Classics

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

… a connected ancient world

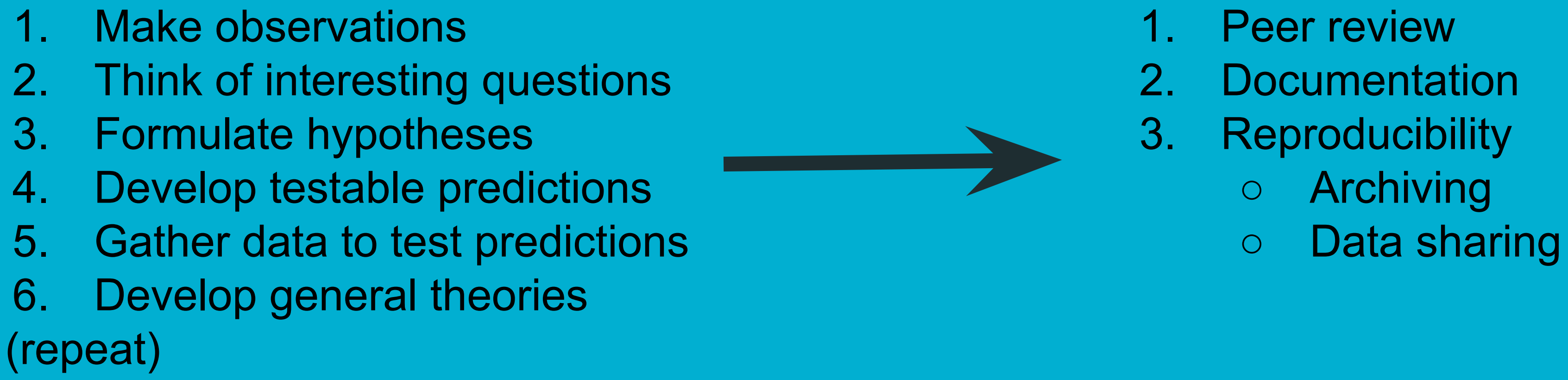


Scientific method and communication

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)

Scientific method and communication



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)

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

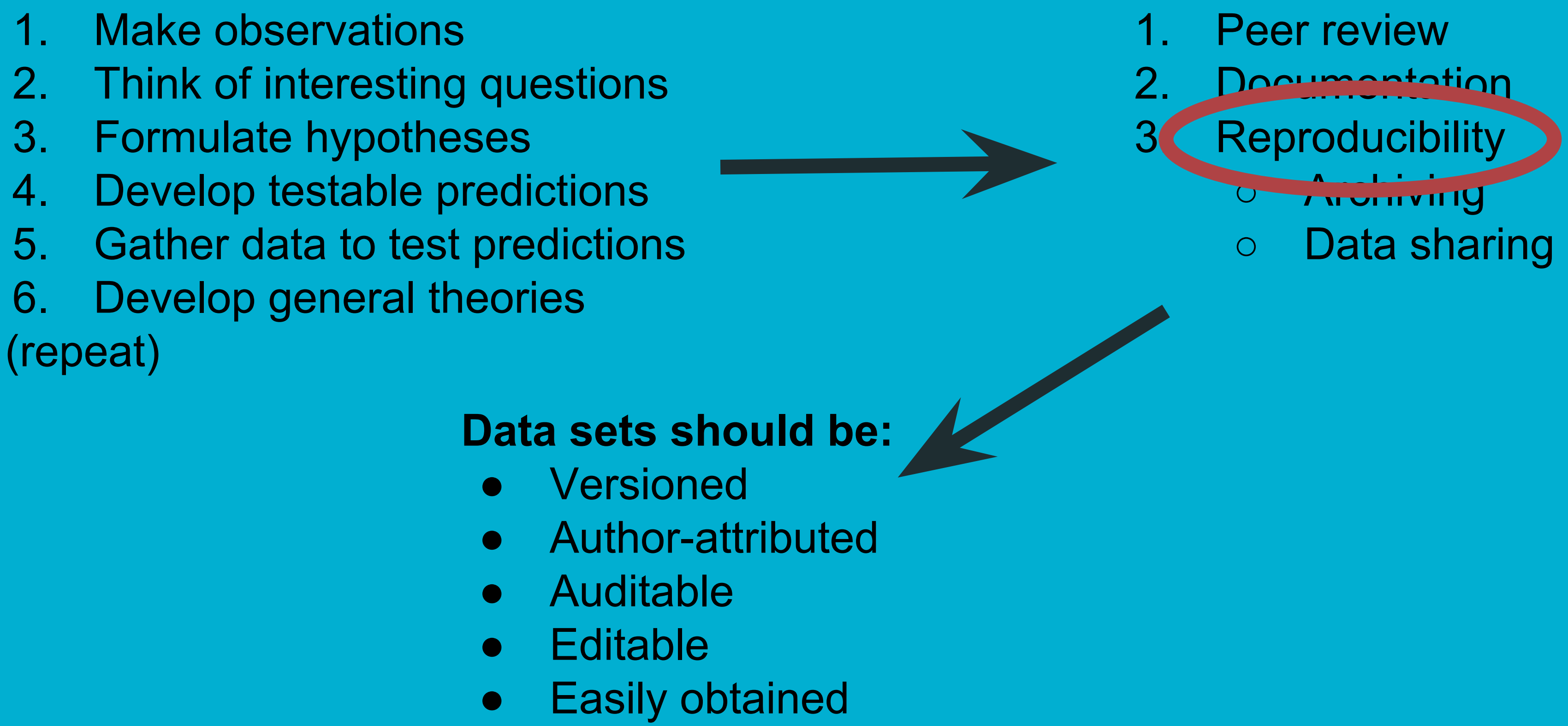
Scientific method and communication

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)

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

Scientific method and communication



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)

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

Data sets should be:

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

Problem: Data set realities

- Storage problems
    - Obscure university websites
    - Unreliable personal websites
    - Disappearing old versions
    - Buried in source code
    - USB drive, CD
- Format problems
    - Obfuscating compression
    - Ugly file organization
- Fetching problems
    - Bad protocols for large files
    - Unfriendly protocols

Problem: Data set realities

- Storage problems
  - Obscure university websites
  - Unreliable personal websites
  - Disappearing old versions
  - Buried in source code
  - USB drive, CD
- Format problems
  - Obfuscating compression
  - Ugly file organization
- Fetching problems
  - Bad protocols for large files
  - Unfriendly protocols

┐(´Д`)┌

Solution: Git-backed corpus management

Git is:

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

Solution: Git-backed corpus management

Git is:

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

and also:

- Distributed
- Non-linear
- ✓ Collaborative
- Scales
- ✓ fsck
- ✓ Compressed
- Diff updates
- Merge strategies
- Easy updates to end users

Solution: Git-backed corpus management

Git is:

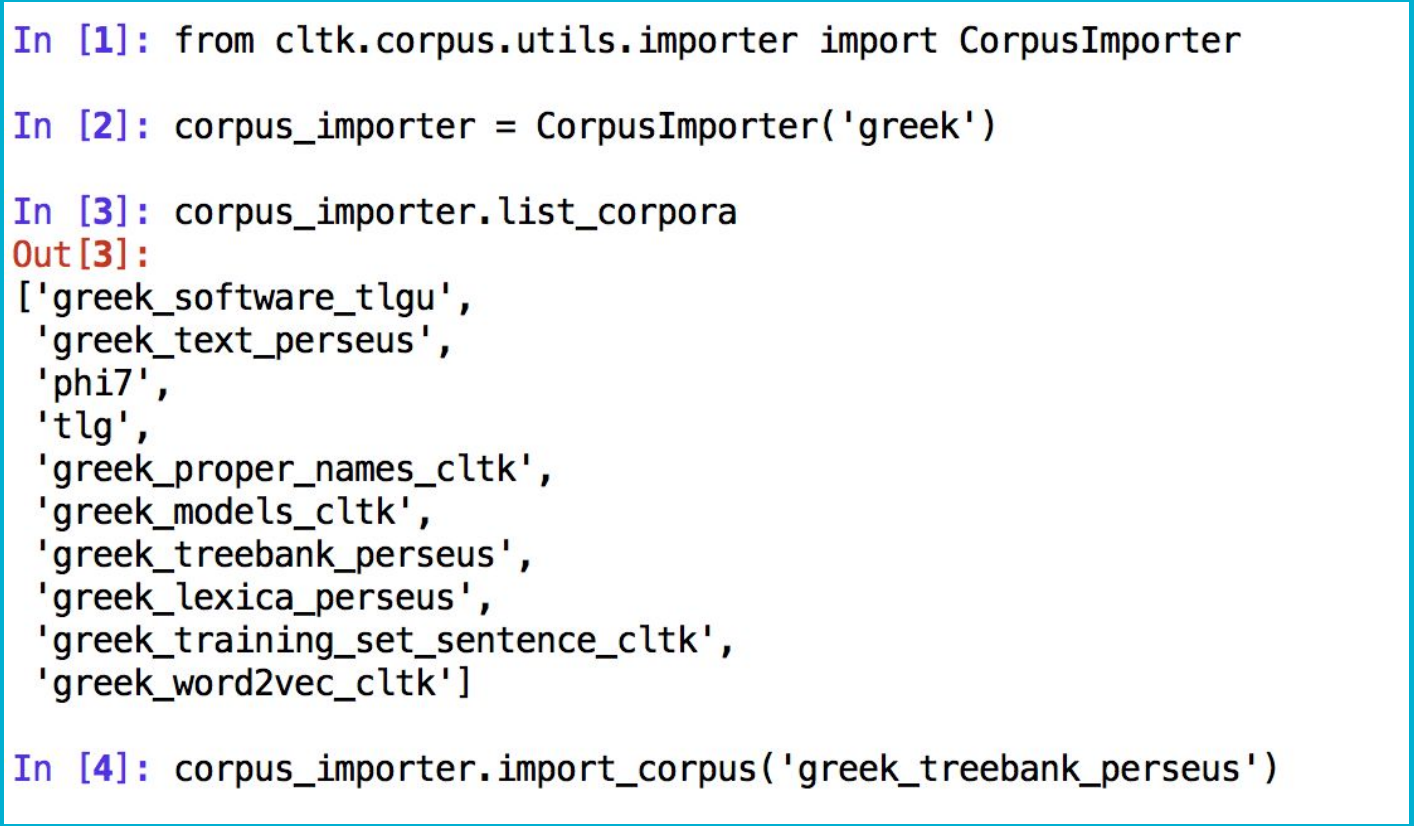
- ✓ Versioned
- ✓ Author-attributed
- ✓ Auditable
- ✓ Editable
- ? Easily obtained

and also:

- Distributed
- Non-linear
- ✓ Collaborative
- Scales
- ✓ fsck
- ✓ Compressed
- Diff updates
- Merge strategies
- Easy updates to end users

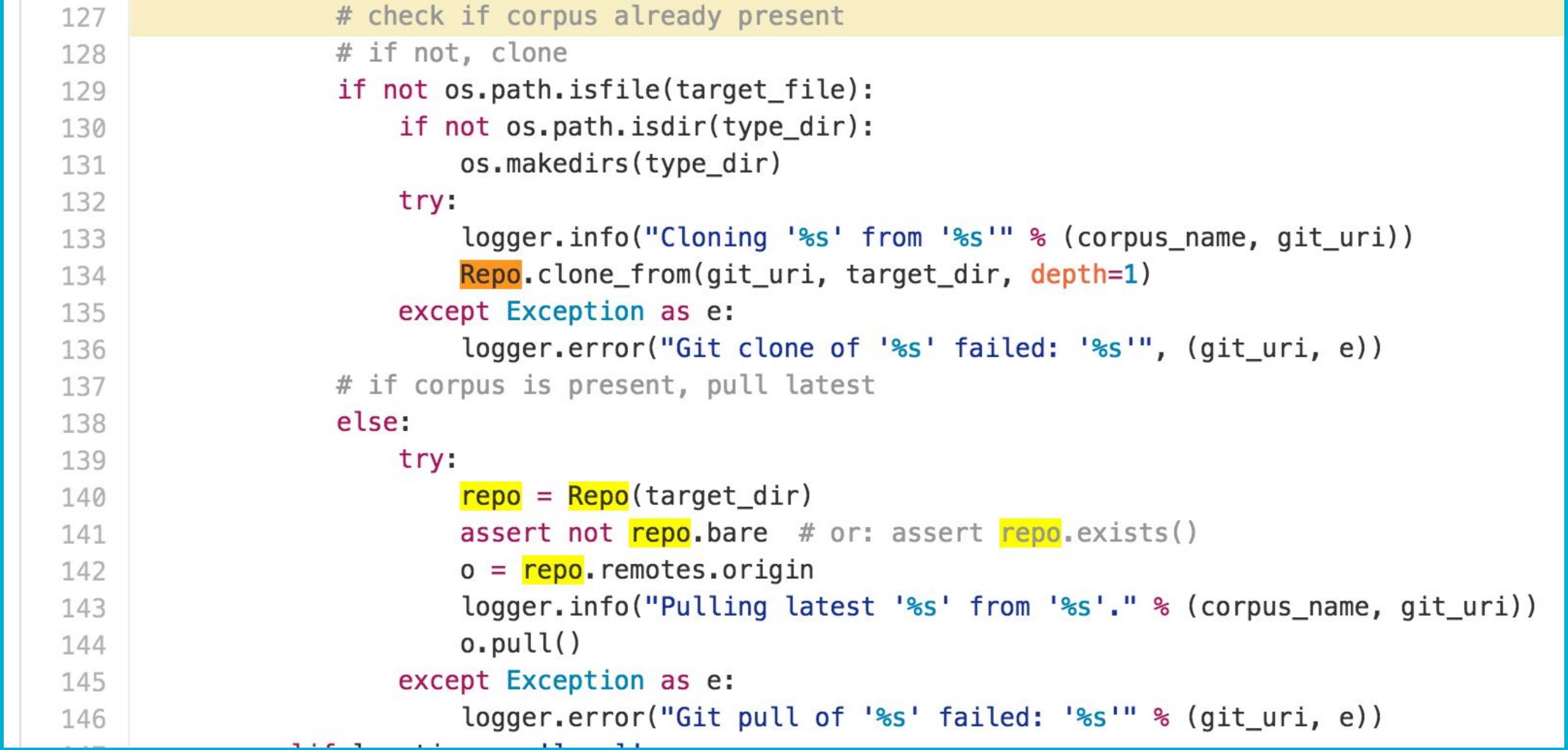
GitHub adds:

- ✓ Easily obtained
- ✓ High availability
- ✓ Community oriented



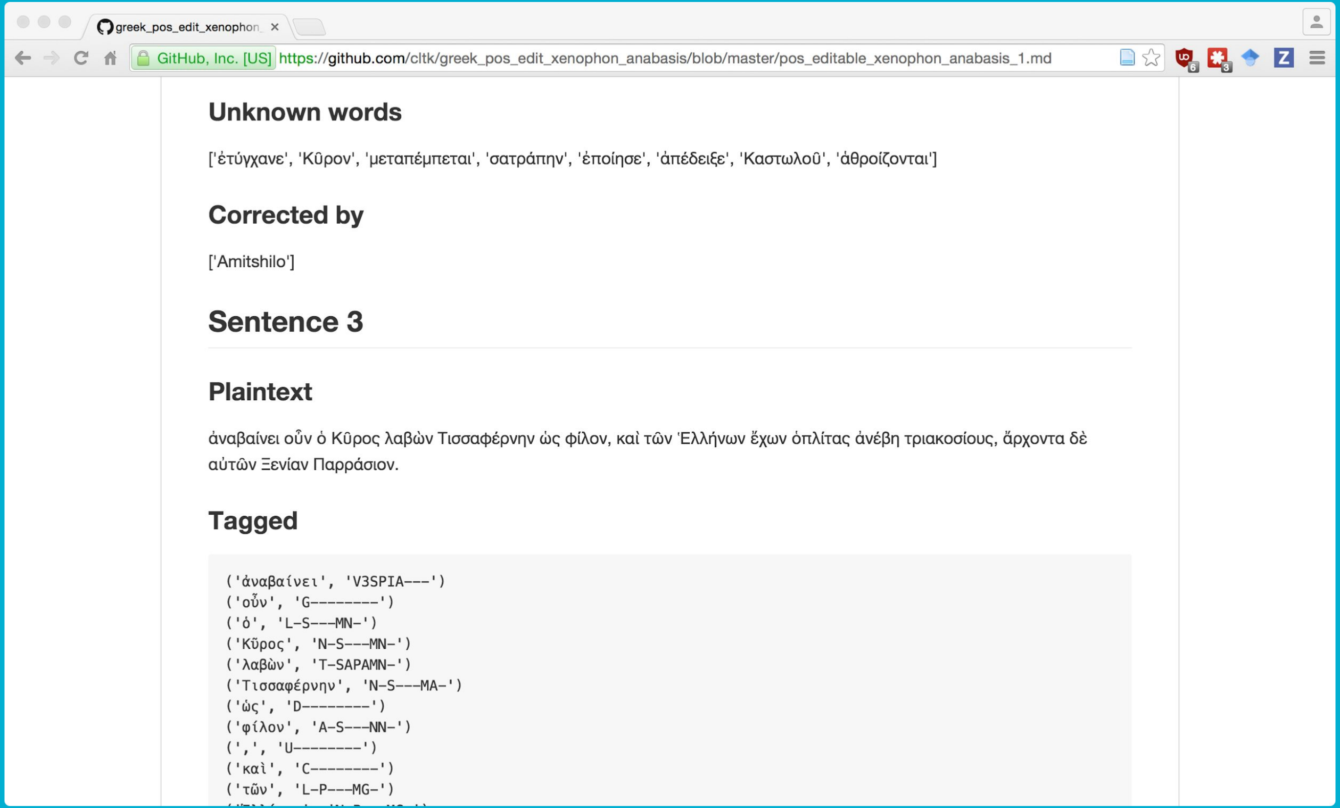
```
In [1]: from cltk.corpus.utils.importer import CorpusImporter

In [2]: corpus_importer = CorpusImporter('greek')

In [3]: corpus_importer.list_corpora
Out[3]:
['greek_software_tlgu',
 'greek_text_perseus',
 'phi7',
 'tlg',
 'greek_proper_names_cltk',
 'greek_models_cltk',
 'greek_treebank_perseus',
 'greek_lexica_perseus',
 'greek_training_set_sentence_cltk',
 'greek_word2vec_cltk']

In [4]: corpus_importer.import_corpus('greek_treebank_perseus')
```

Example of corpus import



```python
        # check if corpus already present
        # if not, clone
        if not os.path.isfile(target_file):
            if not os.path.isdir(type_dir):
                os.makedirs(type_dir)
            try:
                logger.info("Cloning '%s' from '%s'" % (corpus_name, git_uri))
                Repo.clone_from(git_uri, target_dir, depth=1)
            except Exception as e:
                logger.error("Git clone of '%s' failed: '%s'", (git_uri, e))
        # if corpus is present, pull latest
        else:
            try:
                repo = Repo(target_dir)
                assert not repo.bare  # or: assert repo.exists()
                o = repo.remotes.origin
                logger.info("Pulling latest '%s' from '%s'." % (corpus_name, git_uri))
                o.pull()
            except Exception as e:
                logger.error("Git pull of '%s' failed: '%s'" % (git_uri, e))
```

Source code snippet


greek_pos_edit_xenophon_
GitHub, Inc. [US]
https://github.com/cltk/greek_pos_edit_xenophon_anabasis/blob/master/pos_editable_xenophon_anabasis_1.md

Unknown words

['ἐτύγχανε', 'Κῦρον', 'μεταπέμπεται', 'σατράπην', 'ἐποίησε', 'ἀπέδειξε', 'Καστωλοῦ', 'ἀθροίζονται']

Corrected by

['Amitshilo']

Sentence 3

Plaintext

ἀναβαίνει οὖν ὁ Κῦρος λαβὼν Τισσαφέρνην ὡς φίλον, καὶ τῶν Ἑλλήνων ἔχων ὁπλίτας ἀνέβη τριακοσίους, ἄρχοντα δὲ αὐτῶν Ξενίαν Παρράσιον.

Tagged

```
('ἀναβαίνει', 'V3SPIA---')
('οὖν', 'G--------')
('ὁ', 'L-S---MN-')
('Κῦρος', 'N-S---MN-')
('λαβὼν', 'T-SAPAMN-')
('Τισσαφέρνην', 'N-S---MA-')
('ὡς', 'D--------')
('φίλον', 'A-S---NN-')
(',', 'U--------')
('καὶ', 'C--------')
('τῶν', 'L-P---MG-')
```

Example of online annotations

Contribute & contact

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

Sources

- Images
  - http://www.penn.museum/silkroad/exhibit_silkroad.php
- Git
  - 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