Cylleneus

next-gen corpus search for ancient languages

Next-gen search engine for ancient languages

a full-featured and easily extensible open-source search engine library written in Python

enables texts to be searched on the basis of their semantic as well as morphosyntactic properties

takes advantage of rich data from the Sanskrit, Greek, and Latin WordNets

easily integrates with all corpus types, from plain-text corpora to treebanks

texts can be searched by
the **meanings** of words as well as by the
kinds of grammatical constructions they
occur in

Intelligent
Draws on the rich lexical and semantic information of the Sanskrit, Greek, and Latin WordNets

Polyglot
Using the MultiWordNet, meanings can be specified in English, Italian, Spanish, or French

Flexible
Find words, or filter the results of other queries, based on morphological properties – including for plain-text corpora

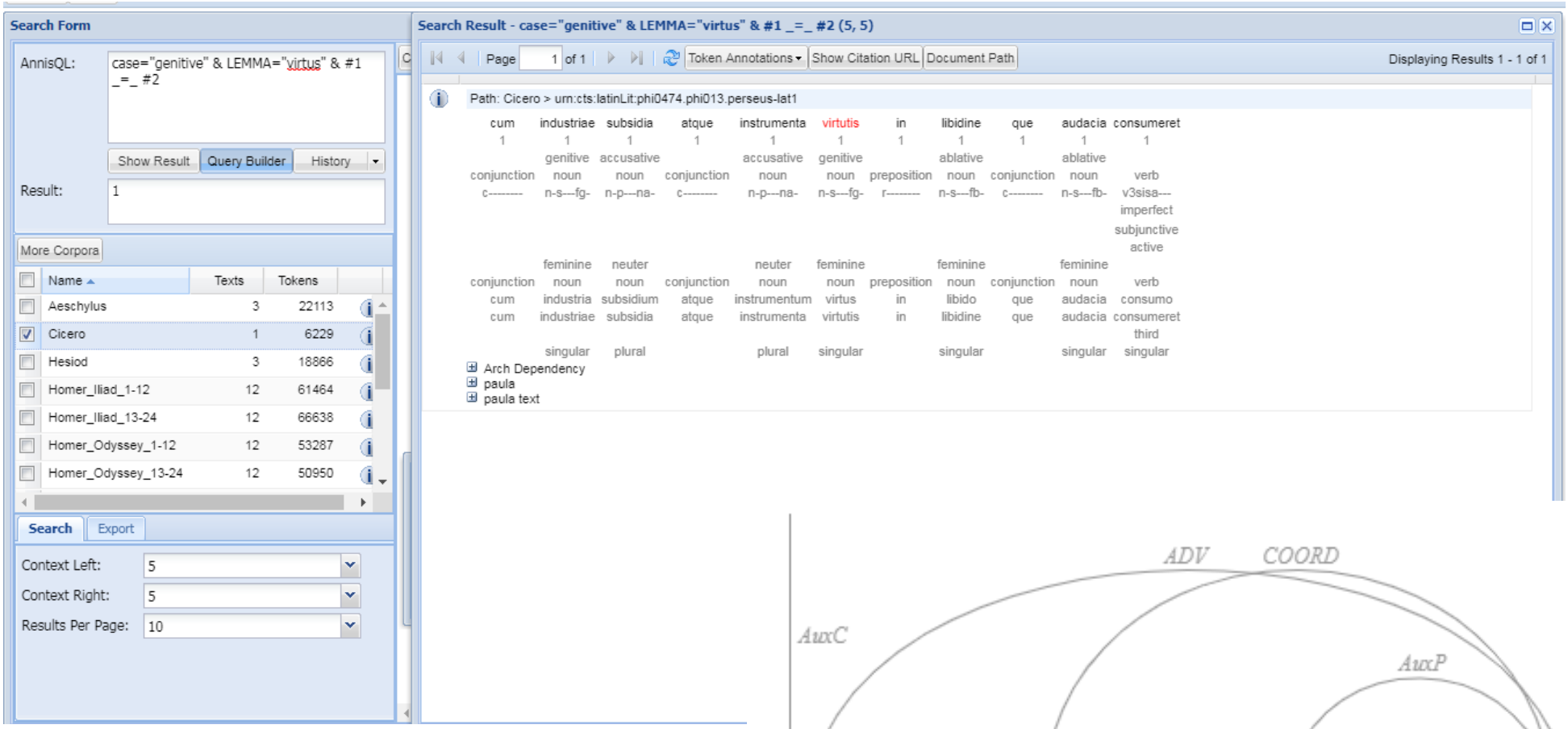
Fast
Once a corpus is indexed, searching is nearly instantaneous for most query types

Advanced
Query types can be combined into complex contextual or phrasal search patterns

Extensible
Indexing pipelines can be created for any corpus type with different annotation schemas (or none!)

‘First-gen’ search tools

- large collections of Latin and Greek texts
- word-form searches
- wildcard queries
- lemma queries
- special cases: e.g.,
  - *Tesserae Project,* intertextual searches between texts and even across languages
  - *Pede Certo,* metrical pattern searches
- syntactic properties and relations



Ancient Greek and Latin Dependency Treebank

A problem of metaphor research

flagrabat ingens bellum ‘a huge war was burning’ (Tac. *Hist*. 2.86)

adolere, (ad)uro, aestuare, ardere, fervere, incendere, torrere . . .

bellum, certatus, certamen, colluctatio, concertatio, conflictus, congressio, congressus, dimicatio, proelium, pugna, Mars . . .

Lemmas

Headword

Part of speech

Morphological description

Sense attributions

Synsets

Part of speech

Unique offset identification number

Gloss

Relations

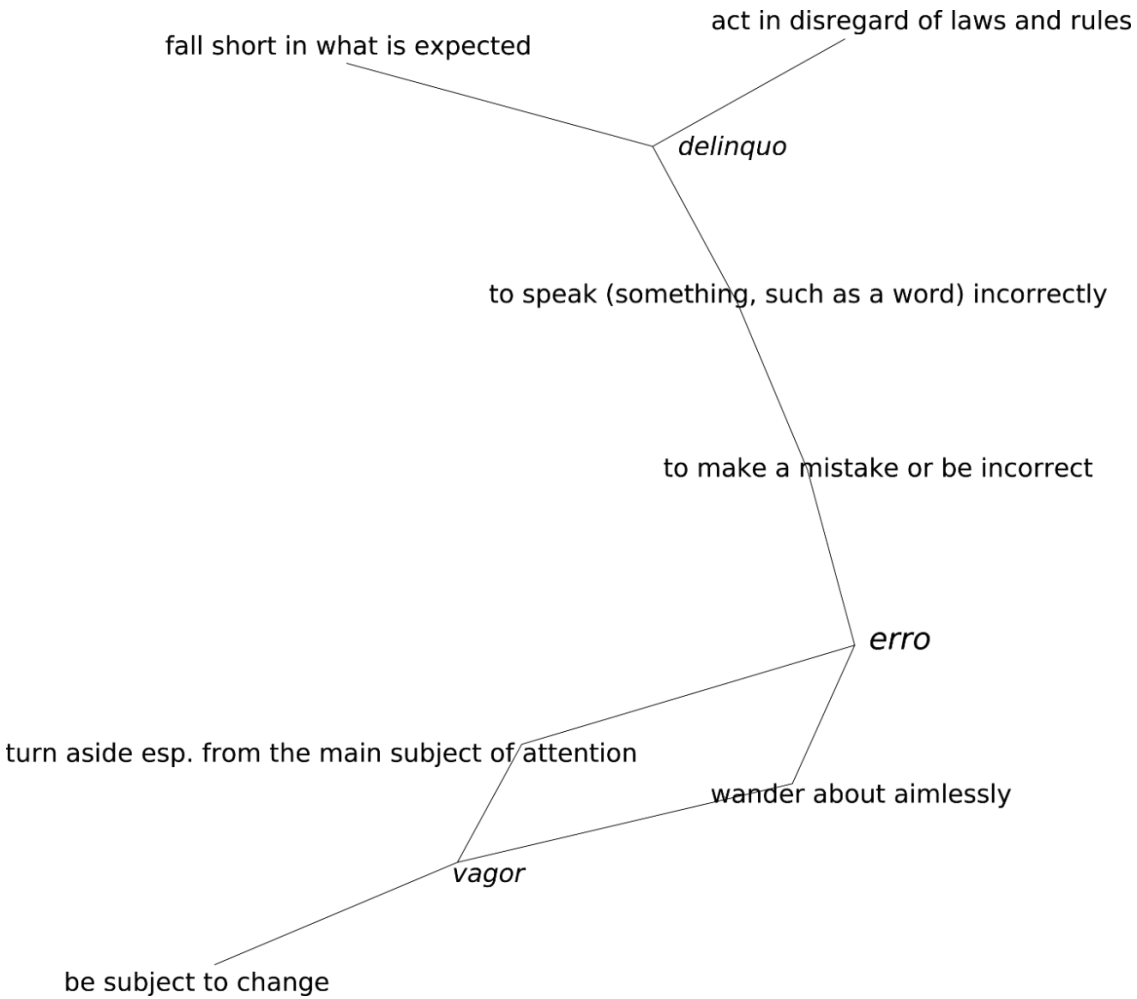
Lexical (derivation, parasynthesis, composition . . .)

Semantic (antonymy, hypernymy, hyponymy . . .)

Semfields

Broad semantic domains including multiple synsets

WordNet data structures



Semantic network of **Latin *erro***

Ancient Language WordNets

Sanskrit about 235,000 lemmas
https://sanskritwordnet.unipv.it

Greek about 112,000 lemmas
https://greekwordnet.chs.harvard.edu

Latin about 70,000 lemmas
https://latinwordnet.exeter.ac.uk

Ancient Language WordNets

- Discrimination between literal, metonymic, and metaphor sense of words
- Etymological information
- Metaphorical and metonymic mappings that capture supra-lexical relations between concepts
- Diachronic and generic tagging, at the level of sense attribution
- RESTful API for programmatic access

Ancient Greek and Latin Sembank

Cat. *Agr.* pr.1

est *interdum* *praestare* *mercaturis* *rem* *quaerere*

(v#01246259)
'value more highly'

(v#01513874)
'come into possession of'

(r#00020741) 'on certain occasions'

(n#09639711) 'the most common medium of exchange'

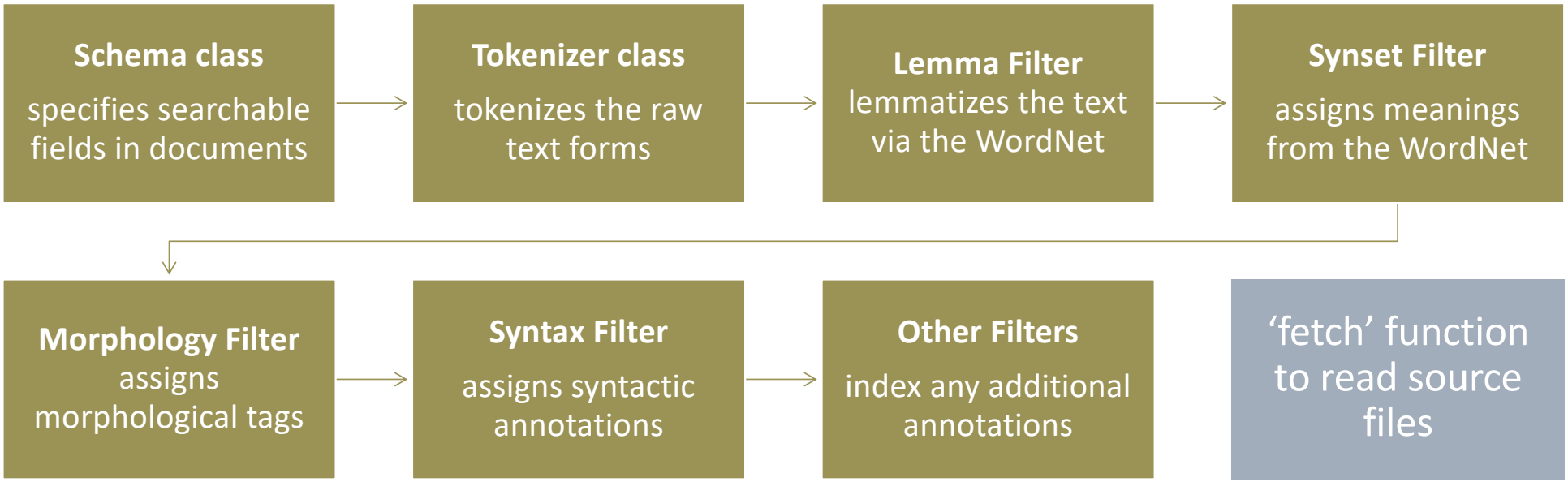
(v#01775163) 'have an existence, be extant'

(n#00707408) 'the commercial of goods and services'

Forms, wildcards	‘animos’	contum*	
Lemmas	<animus>		
Meanings	[en?courage]	[it?guerra]	
Semfields	{611}		
Lexical & semantic relations	</::bellum>	[!::en?courage]	
Morphology	:ACC.PL.		
. . . as a filter	[en?courage]|ABL.SG.		
Morphosyntax	/ablative absolute/		
Contextual & sequential queries	“ . . . ”		

Permissible query types and specification

The indexing pipeline



Works with any structured or plaintext corpus

Ready-made indexing infrastructure is provided for many corpus formats

Perseus Digital Library (in JSON or TEI XML format)
LASLA
PROIEL
AGLDT
CAMENA
DigilibLT
Digital Corpus of Sanskrit
ATLAS
Diorisis
The Latin Library
Perseids Project **translation alignments**

Searches can be performed over ‘collections’ which may include documents from different corpora

Metaphor research, by enabling efficient searching of relations between whole semantic fields

Intertextual research, by finding lexically and even semantically similar expressions

Translation research, by abstracting away from the specific lexicalization of concepts

Permits exploration of ancient literature even without expert linguistic knowledge

Some research use cases

https://mybinder.org/v2/gh/cylleneus/cylleneus/master?filepath=notebooks/quick_search.ipynb