Next-Gen Search for Ancient Languages

MEANING- AND SYNTAX-BASED QUERYING OF ELECTRONIC CORPORA OF GREEK AND LATIN

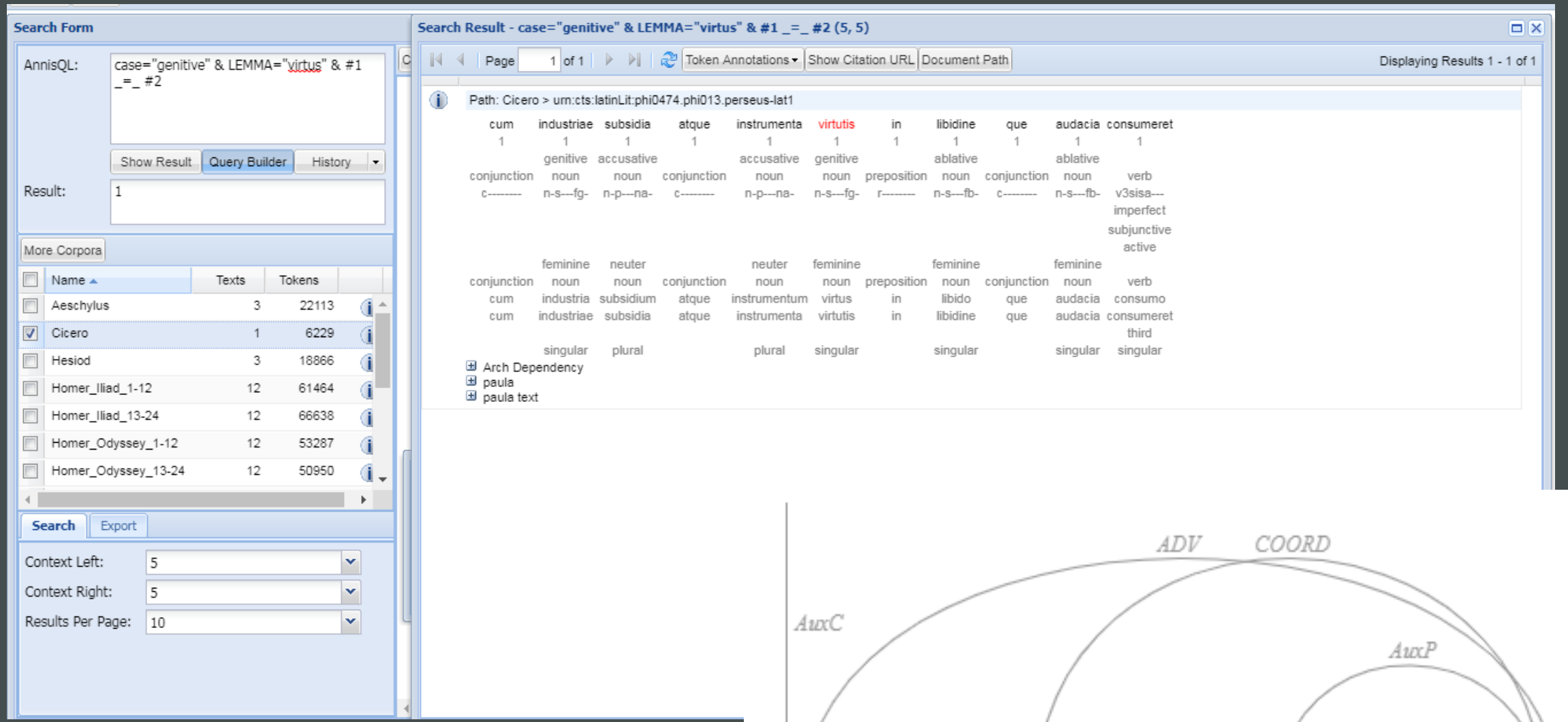
William Michael Short
University of Exeter *w.short@exeter.ac.uk*

First-generation search engine capabilities

- Large collections of Latin and Greek texts
- *n*-gram searches, adjacency or contextual
- ‘wild-card’ queries
- special cases
  - *Tesserae Project*

    intertextual searches between texts and even across languages
  - *Pede Certo*

    metrical pattern searches



Syntactic search

- Treebanks, morphological tagging

The WordNet specification

- Lemmas or Phrases

  headword, POS, sparse morphological tagging

- Synsets

  POS (open classes: n, v, a, r, ‘satellites’)

  offset (unique identification number)

- Relations

  lexical (derives-from, relates-to, composed-of)

  semantic (antonymy, hypernymy, hyponymy)

- Semfields

  Broad semantic domains including multiple synsets

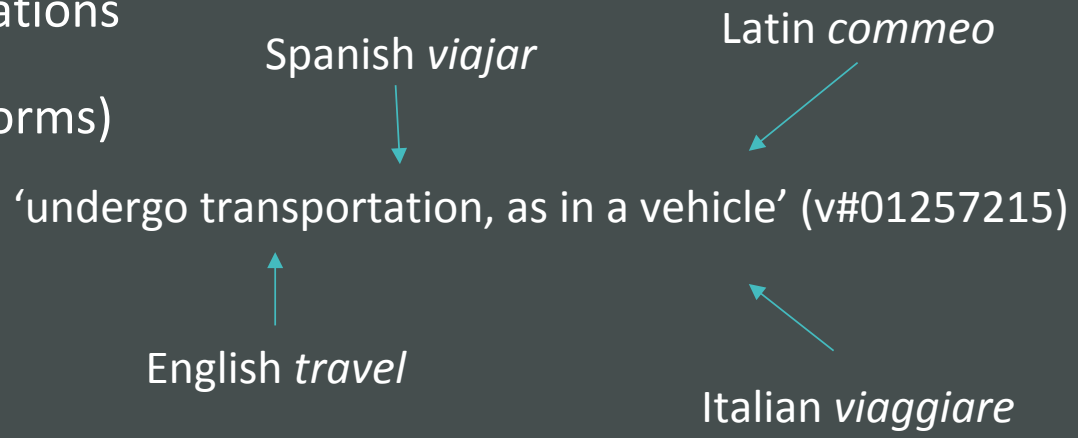
Latin WordNet 1.0

- Included around 9,000 lemmas, aligned with the other WordNets in the MultiWordNet
- Keyed through an automated process to English synsets, with often poor results

  E.g., *erectio* ‘an erect penis’

  Cf. *ērectĭo , ōnis*, f. I. *a setting up*, *erecting*: “tignorum,” Vitr. 10, 5 *fin.*; 8 praef.; Vulg. Num. 10, 21. II. Trop., *pride*, *insolence*, Vulg. Job, 22, 20.

- Relations derived from ‘common’ English relations
- Sparse morphological data (but ambiguous forms)



Latin WordNet 2.0

http://latinwordnet.exeter.ac.uk

- Over 70,000 lemmas covering archaic, classical, late, medieval Latin
- Keyed to English synsets through an automated process collating several electronic lexica, plus on-going manual curation
- Latin-specific synsets, as well as semantic and lexical relations, including derivational relations
- Rich morphological data, lemmatization and translation services

Latin WordNet 2.0

- Alignment with LiLa
  - universal resource identifier for each lemma, currently using LEMLAT unique IDs for disambiguation
- RESTful API for programmatic access
- etymological information
- discrimination of literal, metonymic, and metaphor sense of words
- large-scale figurative groupings that capture relations between sets of concepts (*Lexicon Translaticium Latinum)*

Query types

form matching
- simple
- complex: adjacency
  - contextual

lemma-based
- with morphological annotation

meaning-based
- with morphological annotation
- in different languages

relation-based
- lexical
- semantic

semfield-based

annotation-based