

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

Search Form

AnnisQL: `case="genitive" & LEMMA="virtus" & #1 _=#2`

Show Result Query Builder History

Result: 1

More Corpora

Name	Texts	Tokens
<input type="checkbox"/> Aeschylus	3	22113
<input checked="" type="checkbox"/> Cicero	1	6229
<input type="checkbox"/> Hesiod	3	18866
<input type="checkbox"/> Homer_Iliad_1-12	12	61464
<input type="checkbox"/> Homer_Iliad_13-24	12	66638
<input type="checkbox"/> Homer_Odyssey_1-12	12	53287
<input type="checkbox"/> Homer_Odyssey_13-24	12	50950

Search Export

Context Left: 5

Context Right: 5

Results Per Page: 10

Search Result - case="genitive" & LEMMA="virtus" & #1 _=#2 (5, 5)

Page 1 of 1

Token Annotations Show Citation URL Document Path

Displaying Results 1 - 1 of 1

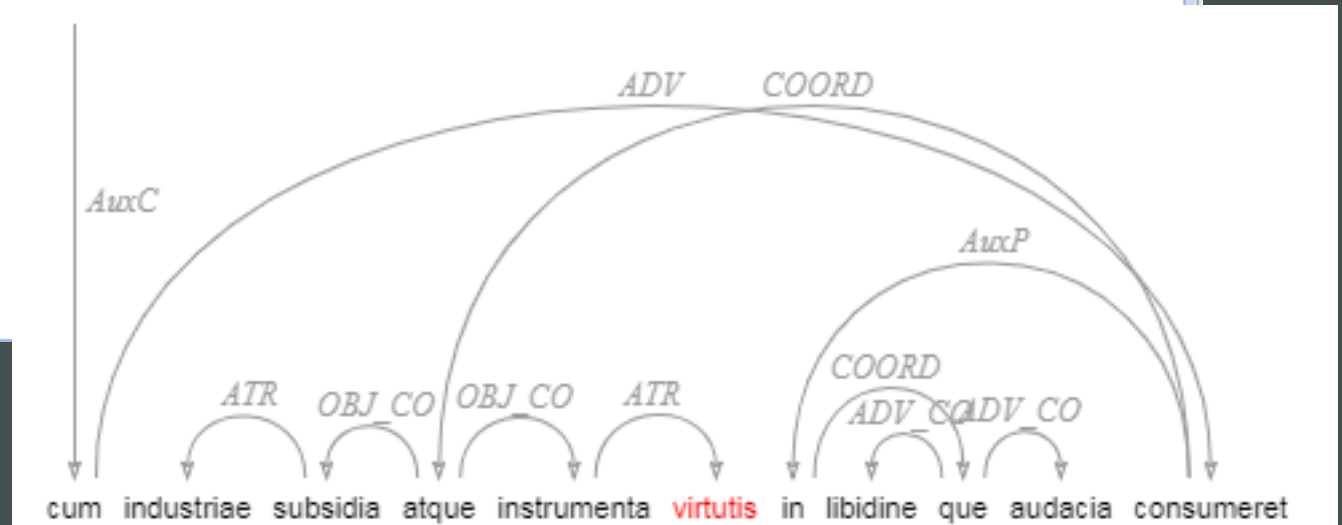
Path: Cicero > um:cts:latinLit:phi0474.phi013.perseus-lat1

cum	industri	subsidia	atque	instrumenta	virtutis	in	libidine	que	audacia	consumeret
1	1	1	1	1	1	1	1	1	1	1
conjunction	genitive	accusative	conjunction	accusative	genitive	preposition	ablative	conjunction	ablative	verb
c-----	n-s--fg-	n-p--na-	c-----	n-p--na-	n-s--fg-	r-----	n-s--fb-	c-----	n-s--fb-	v3sisa---
										imperfect
										subjunctive
										active
	feminine	neuter		neuter	feminine		feminine		feminine	
conjunction	noun	noun	conjunction	noun	noun	preposition	noun	conjunction	noun	verb
cum	industria	subsidiu	atque	instrumentu	virtus	in	libido	que	audacia	consumo
cum	industri	subsidia	atque	instrumenta	virtutis	in	libidine	que	audacia	consumeret
										third
										singular
	singular	plural		plural	singular		singular		singular	singular

Arch Dependency
 paula
 paula text

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



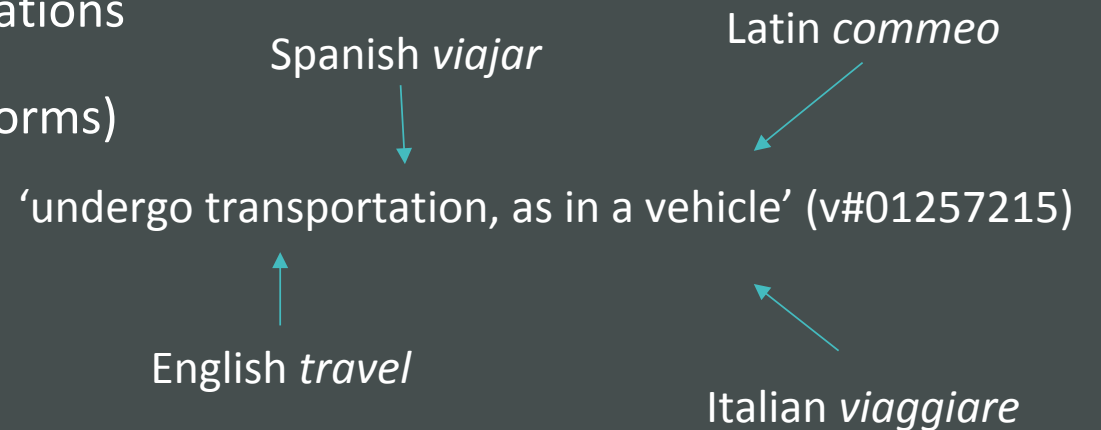
Latin WordNet 1.0

- 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. *ērectiō*, *ōnis*, f. l. *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