Cultural Heritage Language Technologies

Jeff Rydberg-Cox
Director, Classical Studies Program
Department of English and Religious Studies
University of Missouri at Kansas City
rydbergcoxj@umkc.edu

Cultural Heritage Language Technologies

- A collaborative project to create computational tools for the study of Ancient Greek, Early Modern Latin, and Old Norse texts in a network of affiliated digital libraries.
- Project funding provided by the National Science Foundation and the European Union International Digital Library Collaborative Research Program
- http://www.chlt.org

United States Partners

- Classical Studies Program and Department of English, University of Missouri at Kansas City
- The Perseus Project, Tufts University
- The Stoa Consortium, University of Kentucky
- Scandinavian Section, University of California at Los Angeles

European Partners

- The Newton Project and the Department of Computer Science, Imperial College
- Faculty of Classics, Cambridge University
- Istituto di Linguistica Computazionale del CNR, Pisa
- Arnamagnaean Institute, University of Copenhagen

Project Goals

- Adapt techniques from fields of computational linguistics, information retrieval and visualization, and data mining for students and scholars in the humanities.
- Lower the barriers to reading Greek, Latin and Old Norse texts in their original languages.
- Establish an international framework for metadata sharing and interoperability between affiliated digital libraries

Core Technology

- Core digital library technology will be provided under GNU license by Perseus
  - Proven DL system that delivers ~8.5 million pages a month over the web
  - System already in use by three American collaborators
- Applications will be integrated into the production system and made widely available
- System design allows for modularity so that applications can be used in other DL environments or on their own
  - Tufts University group is working to port core technologies to Fedora
  - Many components are being integrated with Greenstone Digital Library
  - We are also exploring integration with Cocoon

Testbeds

- Greek and Latin texts from Perseus (6 million words of Greek, 4 million words of Latin, parallel English translations)
- Works of Isaac Newton from the Newton Project at Imperial College
- Early modern Latin texts from the Stoa Consortium at the University of Kentucky
- Old Norse texts from the University of California at Los Angeles and the Arnamagnaean Institute
- Texts from the Archimedes Project (DFG/NSF funded DL for the history of mechanics)
- Rare books from the History of Science Collection at the Linda Hall Library of Science and Technology in Kansas City

Language Technologies

- NLP technology is mature with a focus on commercial and national security applications.
  - TREC
  - TIDES
  - CLEF
- Which of these technologies are language dependent and, therefore, need to be optimized for cultural heritage languages.
- Which of these technologies are most useful for users in the humanities?

Parsers

- The extension or development of morphological analysis facilities for early modern Latin and Old Norse is fundamental for these applications.
  - Simple stemming techniques (e.g. Porter's algorithm) are not precise enough.
  - In highly inflected languages, lexical normalization is required in order to have enough data to obtain statistically significant results.
- Perseus Project will provide a parser for Classical Greek and Latin.
- The Istituto di Linguistica Computazionale del CNR, Pisa will develop a system for early modern Latin.
- The University of California at Los Angeles and the Arnamagnaean Institute and will create a parser for Old Norse
- We have also extended beyond our original plans with a parser for Old English

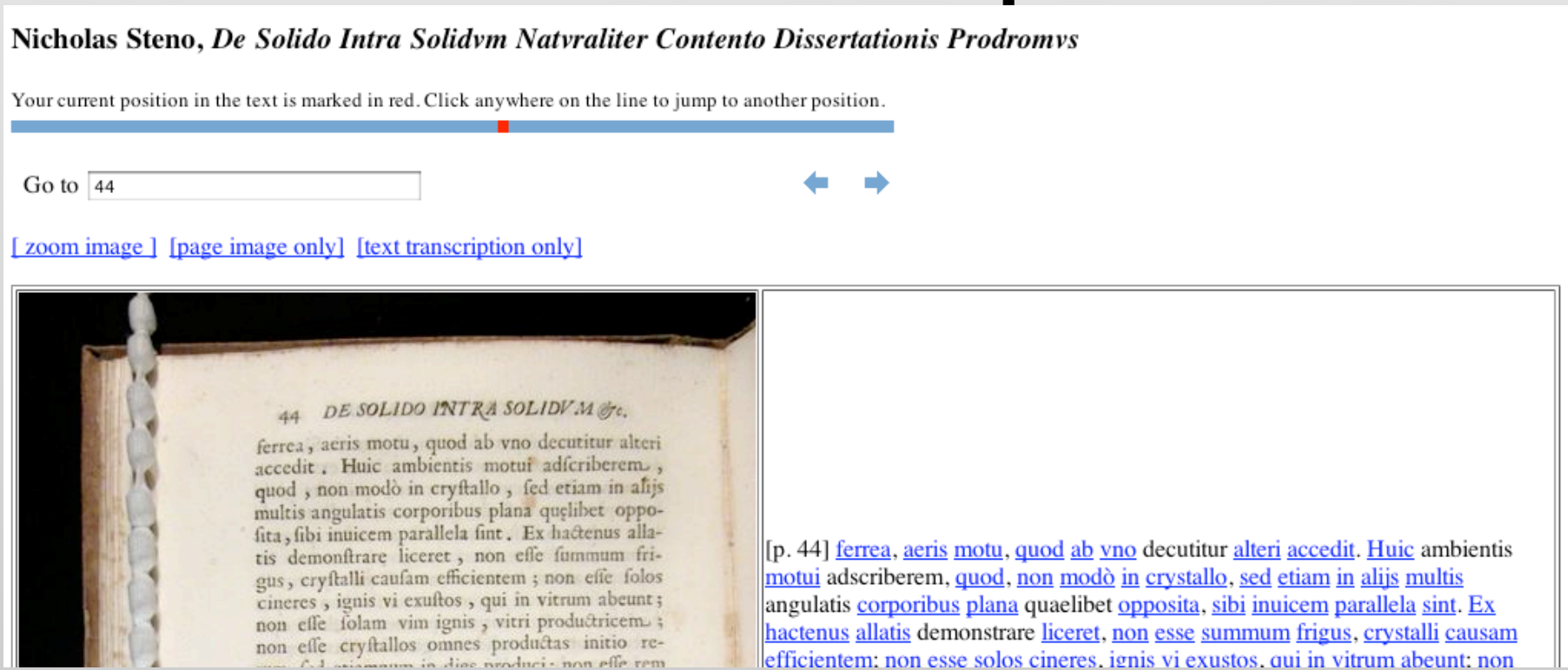
Three Approaches to Parsing

- Parsers are expensive and difficult to develop. We have explored three methods with two goals in mind:
  - Help students and lifelong learners read texts in original languages
  - Provide accurate data for NLP and IR applications
- Most Expensive: Neolatin Parser group is extracting data from lexica and hand coding morphological types according to EAGLES standards
- Medium: More automated approach where morphological data is automatically extracted and rules based parser generates parses or a database of possible lexical forms
- Least Expensive: Mining detailed text specific lexica for citation information and using this data to leverage pattern matching in the original text.
- Least expensive method good for reading support but not IR or NLP.
- Medium level provides good precision/recall/expense balance

Integrated Reading Environment

- Parser is used to automatically generate hypertext leading to word study tool
- Word study tool shows lexical form, links to dictionaries, grammars, frequency, and search tools
- Perseus text display technology initially built for Greek, Latin, and English. Has been generalized for Old Norse, Old English, and other languages.

Page Images of Rare Books & Manuscripts



- Zoomable, high resolution page images are displayed as part of reading environment
- Infrastructure will be used for Old Norse manuscript images

Old Norse Reading Environment

Link to translations and other versions of the text (link can be generated to versions in federated DLs via OAI)

Word Study Tool with extracted short definition of word, morphological analysis, frequency data and search links

Volsunga Saga

Editions and translations: Old Norse | English (ed. William Morris and Eirikr Magnusson)
Your current position in the text is marked in red. Click anywhere on the line to jump to another position.

Table of Contents Go to 1

Volume 1

I. [p. 3] HÉR hefr upp ok segir frá þeim manni, er Sigi er nefndr ok kallaðr, at héti sonr Óð
var ríkr ok mikill fyrir sér, en þó var Sigi þeira enn ríkari ok ættstærri, at því er menn mæltu
við söguna. Hann hét Breði. Hann er fróðr við þat, er hann skyldi at hafast. Hann hafði íþróttir ok atgervi jafnframt hinum, er meira þóttu verðir, eða umfram nokkura. Þat er at segja eitthvert sinn, at Sigi ferr á dýraveiði, ok með honum þrællinn, ok veiða dýr um daginn allt til aptans. En er þeir bera saman veiði sína um aptaninn, þá hafði Breði veitt miklu fleira ok meira en Sigi, en honum líkaði stórilla ok segir, at sik undri, at einn þræll skuli sik yfirbuga í dýraveiði, hleypr því at honum ok drepr hann; dysjar síðan líkit í snjófönn. Nú ferr hann heim um kveldit ok segir, at Breði hafi riðit frá honum á skóginn, -- "ok var hann senn ór augliti mér, ok veit ek ekki til hans." Skaði grunar sögn Siga ok getr, at vera munu svik hans ok mun Sigi hafa drepit hann, [p. 4] fær menn til at leita hans, ok lýkr svá leitinni, at þeir fundu hann í skafli einum, ok mælti Skaði, at þann skafl skyldi kalla Breðafönn heðan af, ok hafa menn nú þat eptir síðan ok kalla svá hverja fönn, er mikil er. Þá kemr upp, at Sigi hefir drepit þrælinn ok myrðan. Þá kalla þeir hann varg í véum, ok má hann nú eigi heima vera með feðr sínum. Óðinn fylgir honum nú af landi brott, svá langa leið, at stóru bar, ok eigi létti hann fyrr en hann kom honum til herskipa. Nú tekr Sigi at leggjast í hernað með þat lið, er faðir hans fekk honum, áðr þeir skildu, ok varð hann sigrsæll í hernaðinum. Ok svá kemr hans máli, at hann fekk herjat sér land ok ríki um síðir. Ok því næst fekk hann sér göfugt kvánfang, ok gerðist hann ríkr konungr ok mikill fyrir sér ok réð fyrir Húnalandi ok er inn mesti hermaðr. Hann á son við konu sinni, er hét Rerir. Hann vex þar upp með feðr sínum ok gerist brátt mikill vexti ok gerviligr.

Word Study Tool

mikill	great, tall, of stature (m vexti, maðr m ok sterkr);	
mikill	strong nom sg	
Frequency by Authors		**Old Norse Word Search**

Corpus	Words	Max. Inst.	Freq./10K	Min. Inst.	Freq./10K

Old Norse text with automatically generated hypertext with links to word study tool

Digital Library Federation

- Federation of links to different versions and translations of works already possible via OAI
- Original vision was to share high-granularity data like morphological analysis via OAI
- This approach would 'break' OAI services that are expecting work level metadata
- In place of OAI, we have established a morphology SOAP service that allows for submission of documents for morphological annotation (http://www.chlt.org/cgi-bin/morph/parser)

Federation of Higher Resolution Metadata

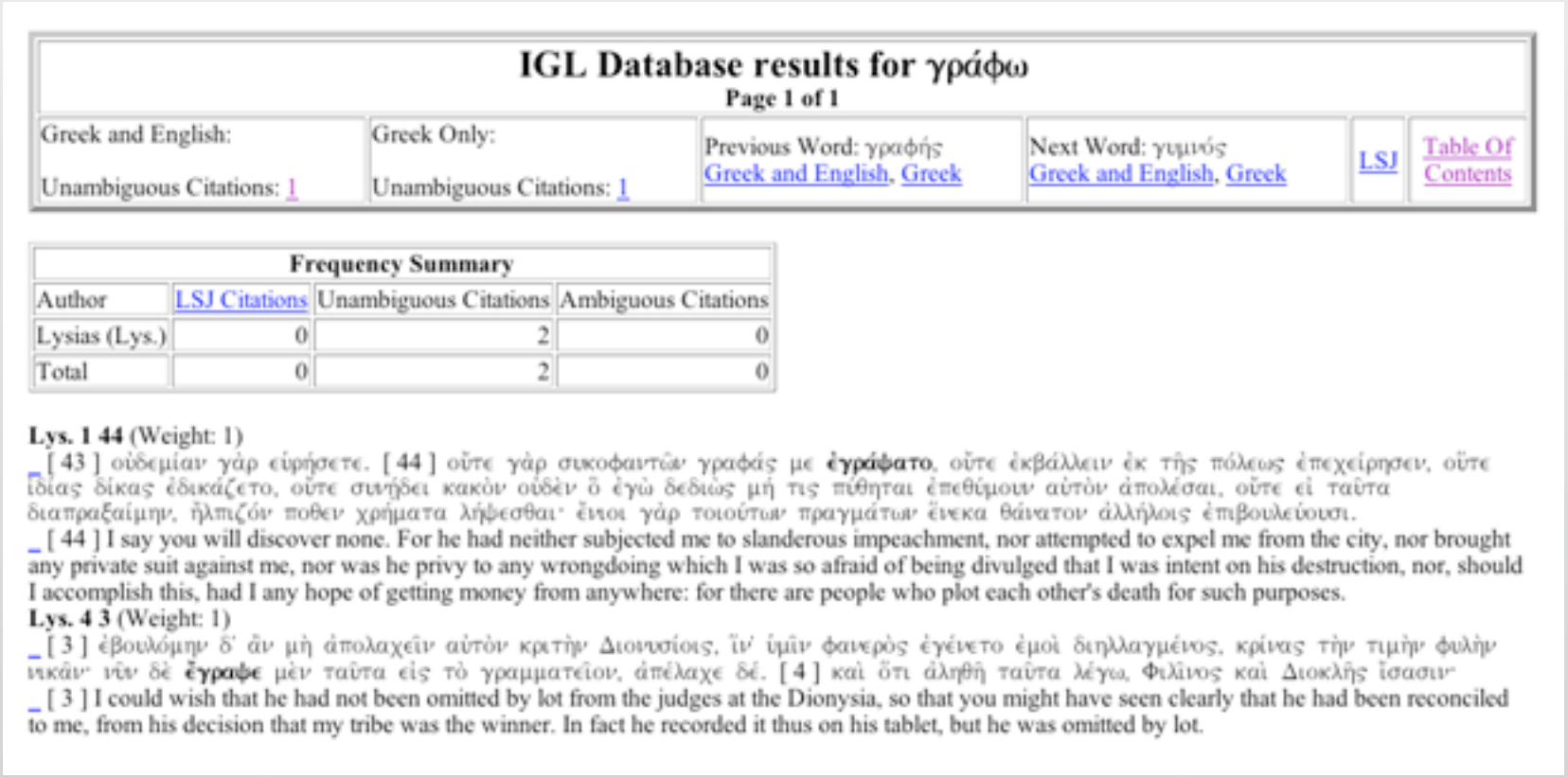
- In current practice, OAI data provides a very low resolution look at documents in the digital library.
- Harper's *Dictionary of Classical Antiquities* would have one OAI record for a 12 Mb structured document made up of 50,000 individual entries.
- Other unabridged dictionaries are even more dense
- OAI records publicize the existence of the document, but offers no help to the harvester in making use of it.

Case Study: PlanetMath.org

- Hosts an encyclopedia of mathematical terminology.
- The PlanetMath OAI data provider presents itself as a collection of single-page documents, rather than a monolithic document.
- The implication is that a harvester can extract the headwords of the encyclopedia directly from the OAI provider.
- Direct access to the sub-documents can give harvesters detailed access to the document.
- This increases the potential for linking between sites and opens the possibility for uses of the data that the original document maintainers would never have envisioned.
- The question is whether this level of granularity is appropriate for the OAI or if this level of access is like morphological analysis and so should be provided as a SOAP service.
- We are testing both approaches with Harper's Encyclopedia and Zoega's Lexicon of Old Icelandic

Vocabulary Profiles & Reintegration of Expert Knowledge

- Even extremely simple applications such as word lists can be useful for non-specialist users
- Integration with other measures such as frequency, relative frequency, *tf x idf* scores, and measures of lexical richness can make these tools even more useful
- Profiles are being used at Cambridge University to write the first new Greek-English lexicon in 120 years
- 'Born Digital' lexicon will be integrated into DL reading environment



IGL Database results for γράφω

Page 1 of 1

Greek and English:	Greek Only:	Previous Word: γραφής	Next Word: γυμνός	LSJ	Table Of Contents
Unambiguous Citations: 1	Unambiguous Citations: 1	Greek and English, Greek	Greek and English, Greek		

Frequency Summary

Author	LSJ Citations	Unambiguous Citations	Ambiguous Citations
Lysias (Lys.)	0	2	0
Total	0	2	0

Lys. 1 44 (Weight: 1)

_ [43] οὐδεμίαν γὰρ εἰρήσετε. [44] οὔτε γὰρ συκοφαντῶν γραφάς με **ἐγράψατο**, οὔτε ἐκβάλλειν ἐκ τῆς πόλεως ἐπεχείρησεν, οὔτε ἰδίας δίκας ἐδικάζετο, οὔτε συνῄδει κακὸν οὐδὲν ὃ ἐγὼ δεδιὼς μή τις πύθηται ἐπεθύμουν αὐτὸν ἀπολέσαι, οὔτε εἰ ταῦτα διαπραξαίμην, ἤλπιζόν ποθεν χρήματα λήψεσθαι· εἰσὶ γὰρ τοιούτων πραγμάτων ἕνεκα θάνατον ἀλλήλοις ἐπιβουλεύουσι.

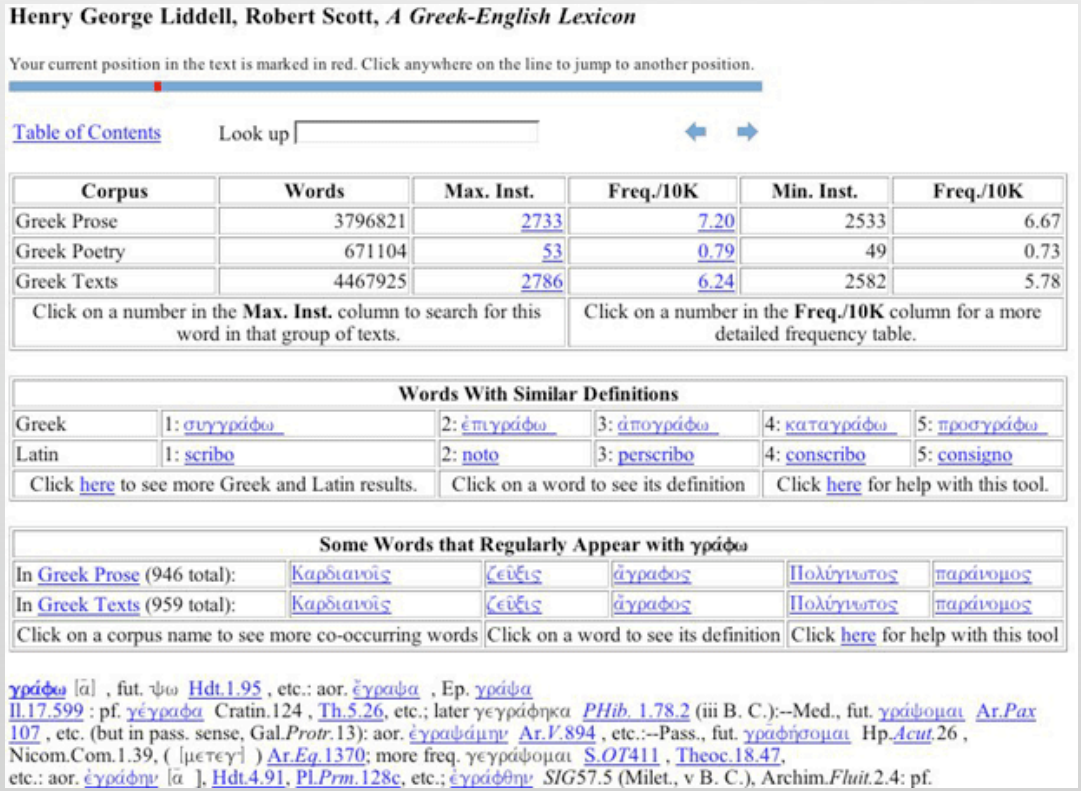
_ [44] I say you will discover none. For he had neither subjected me to slanderous impeachment, nor attempted to expel me from the city, nor brought any private suit against me, nor was he privy to any wrongdoing which I was so afraid of being divulged that I was intent on his destruction, nor, should I accomplish this, had I any hope of getting money from anywhere: for there are people who plot each other's death for such purposes.

Lys. 4 3 (Weight: 1)

_ [3] ἐβουλόμην δ' ἂν μὴ ἀπολαχεῖν αὐτὸν κριτὴν Διονυσίοις, ἵν' ὑμῖν φανερὸς ἐγένετο ἐμοὶ διηλλαγμένος, κρίνας τὴν τιμὴν φυλὴν νικᾶν· νῦν δὲ **ἔγραφε** μὲν ταῦτα εἰς τὸ γραμματεῖον, ἀπέλαχε δέ. [4] καὶ ὅτι ἀληθῆ ταῦτα λέγω, Φιλῖνος καὶ Διοκλῆς ἴσασιν·

_ [3] I could wish that he had not been omitted by lot from the judges at the Dionysia, so that you might have seen clearly that he had been reconciled to me, from his decision that my tribe was the winner. In fact he recorded it thus on his tablet, but he was omitted by lot.

- Key-word-in-context display for every occurrence of a word in the corpus along with English translations from the Perseus Corpus where possible.
- Passages are presented in LSJ and then chronological order
- Accompanied by an author-by-author frequency summary.



Henry George Liddell, Robert Scott, *A Greek-English Lexicon*

Your current position in the text is marked in red. Click anywhere on the line to jump to another position.

Table of Contents Look up

Corpus	Words	Max. Inst.	Freq./10K	Min. Inst.	Freq./10K
Greek Prose	3796821	2733	7.20	2533	6.67
Greek Poetry	671104	53	0.79	49	0.73
Greek Texts	4467925	2786	6.24	2582	5.78

Click on a number in the **Max. Inst.** column to search for this word in that group of texts. | Click on a number in the **Freq./10K** column for a more detailed frequency table.

Words With Similar Definitions

Greek	1: συγγράφω	2: ἐπιγράφω	3: ἀπογράφω	4: καταγράφω	5: προσγράφω
Latin	1: scribo	2: noto	3: perscribo	4: conscribo	5: consigno

Click here to see more Greek and Latin results. | Click on a word to see its definition | Click here for help with this tool.

Some Words that Regularly Appear with γράφω

In Greek Prose (946 total):	Καρδιαινοῖς	ζεῦξις	ἄγραφος	Πολύγνωτος	παράνομος
In Greek Texts (959 total):	Καρδιαινοῖς	ζεῦξις	ἄγραφος	Πολύγνωτος	παράνομος

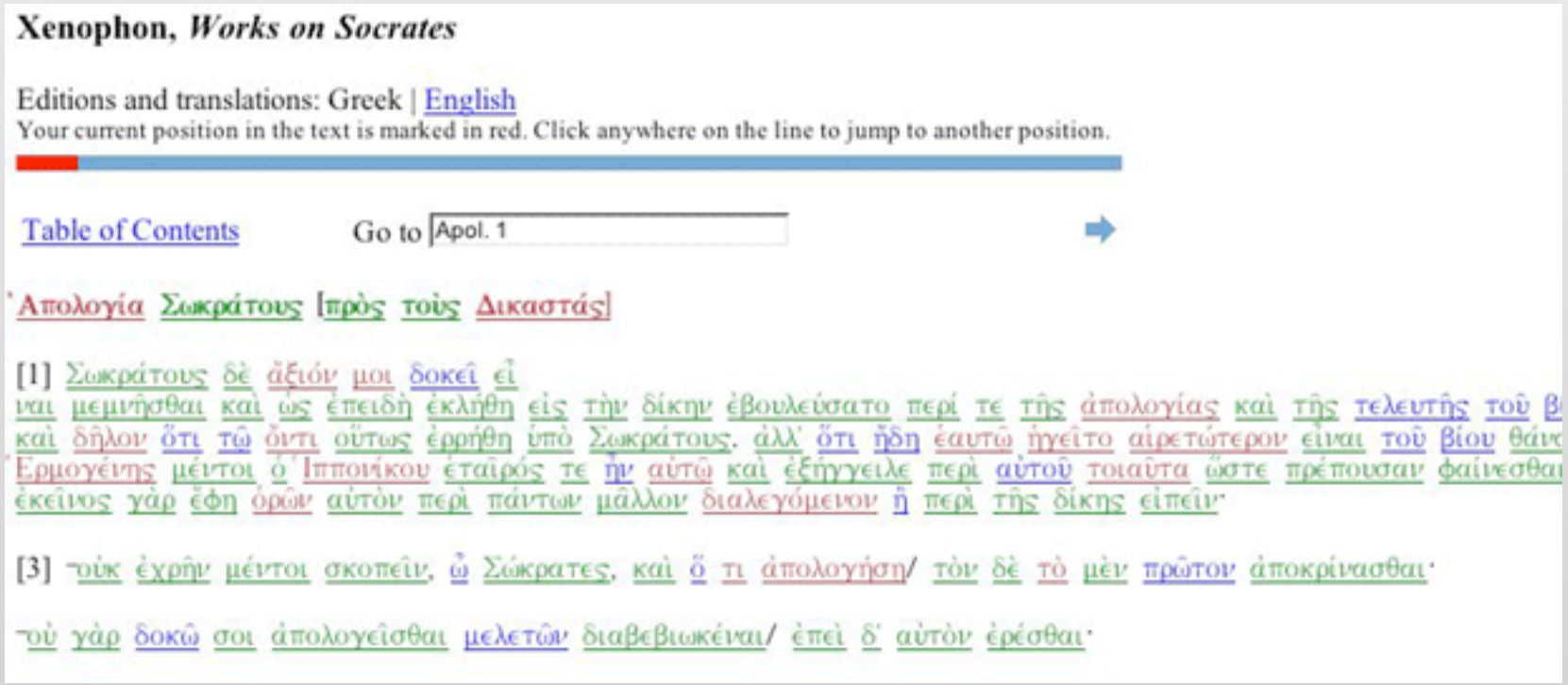
Click on a corpus name to see more co-occurring words | Click on a word to see its definition | Click here for help with this tool

γράφω [ᾰ] , fut. ψω Hdt.1.95 , etc.: aor. ἔγραψα , Ep. γράψα Il.17.599 : pf. γέγραφα Cratin.124 , Th.5.26, etc.; later γεγράφηκα *PHib.* 1.78.2 (iii B. C.):--Med., fut. γράψομαι Ar.*Pax* 107 , etc. (but in pass. sense, Gal.*Protr.*13): aor. ἐγραψάμην Ar.*V.*894 , etc.:--Pass., fut. γραφήσομαι Hp.*Acut.*26 , Nicom.Com.1.39, ([μετεγ]) Ar.*Eq.*1370; more freq. γεγράψομαι S.*OT*411 , Theoc.18.47, etc.: aor. ἐγράφην [ᾰ], Hdt.4.91, Pl.*Prm.*128c, etc.; ἐγράφθην *SIG*57.5 (Milet., v B. C.), Archim.*Fluit.*2.4: pf.

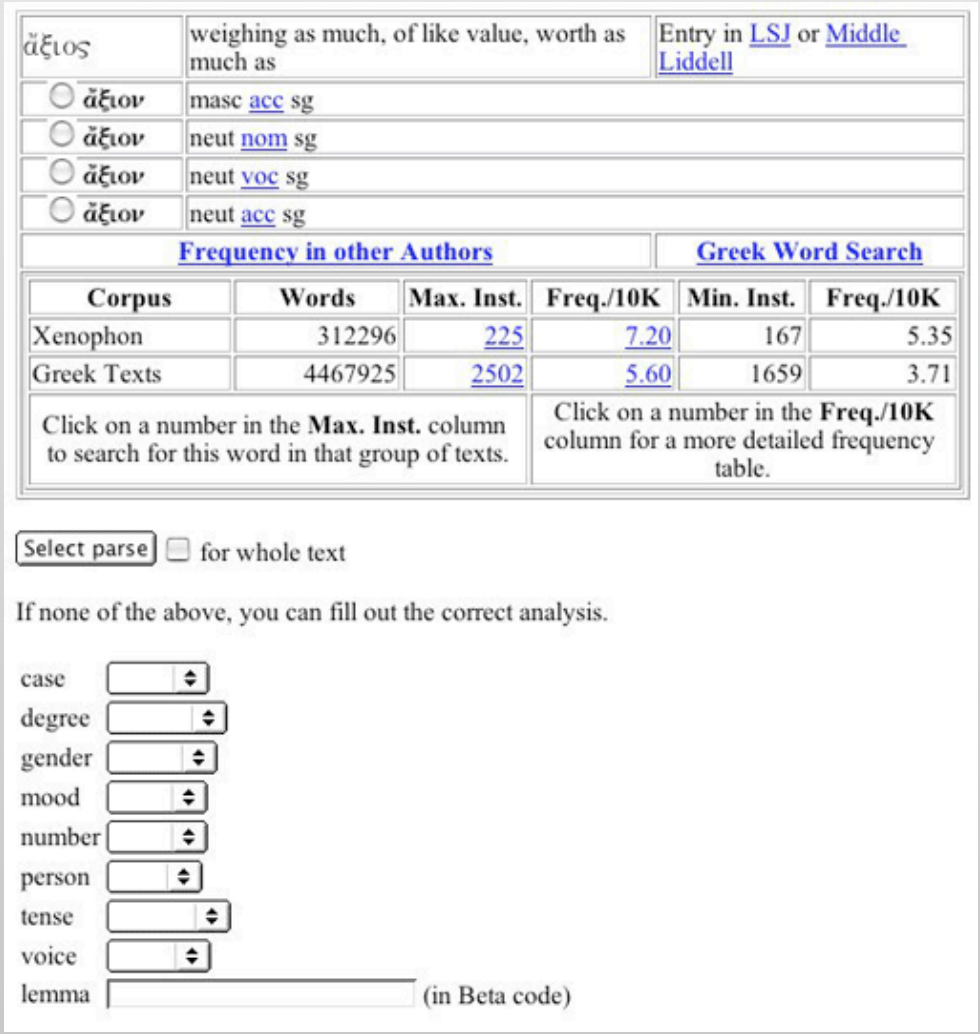
- Links are provided to the Online Edition of the Liddell, Scott, Jones Greek English Lexicon .
- Including both the lexicon entry and statistical information
  - comparative frequency data,
  - word collocation information
  - automatically extracted list of words with similar definitions.

Integration of Expert Knowledge

- While automatic processes can provide a great deal of useful information, scholars will want to correct, annotate, and extend automatic results
- Usual practice is for scholars to print 'screen dumps' and hand annotate
- We need systems to capture this knowledge about morphology and syntax and reintegrate it into the DL system



- Text display system marks ambiguous forms
  - Unambiguous forms are green
  - Morphologically ambiguous forms are red
  - Lexically ambiguous forms are blue.



ἄξιος	weighing as much, of like value, worth as much as	Entry in LSJ or Middle Liddell
○ ἄξιον	masc acc sg	
○ ἄξιον	neut nom sg	
○ ἄξιον	neut voc sg	
○ ἄξιον	neut acc sg	

Frequency in other Authors | Greek Word Search

Corpus	Words	Max. Inst.	Freq./10K	Min. Inst.	Freq./10K
Xenophon	312296	225	7.20	167	5.35
Greek Texts	4467925	2502	5.60	1659	3.71

Click on a number in the **Max. Inst.** column to search for this word in that group of texts.

Click on a number in the **Freq./10K** column for a more detailed frequency table.

Select parse ☐ for whole text

If none of the above, you can fill out the correct analysis.

case
degree
gender
mood
number
person
tense
voice
lemma (in Beta code)

- Ambiguous word links brings users to a form where they can see an interface with all of the possible parses and a simple web form where they can indicate the correct parse.

Cross-Lingual Information Retrieval

- Cross-lingual information retrieval allows users to identify documents of interest written in languages in which the user can not form an intelligible query.
- At best, it only identifies documents that need further study or professional translation
- Do students and scholars using digital resources in the humanities have different needs?

Approaches

- Essentially a miniature machine translation problem.
- Approaches have focused on dictionary translation and generation of translation equivalents from parallel and comparable corpora
- Main problems are the introduction of ambiguity in the translation process

Dictionary Translation

- Simple dictionary translation of a query (mechanical replacement of source language query terms with target language translations) introduces ambiguity
  - Adds extra senses
  - Doesn't cover technical terminology
  - Doesn't deal with phrases

Parallel and Comparable Corpora

- Statistical measures are used to generate similarity thesauri from parallel or comparable that are then used to translate query terms
- Many of the same ambiguity problems of dictionary translation
- Parallel corpora are expensive
- Most effective in a specific, restricted domains

Approaches for Greek, Latin, and Old Norse

- We can use both approaches
  - Study in these fields is by nature domain specific, so parallel / comparable corpora can be used
  - Corpus is constrained and stable
    - Cost of parallel corpora is finite
    - Parallel corpora once constructed have lasting value
- Presence of canonical and 'comprehensive' dictionaries allows for dictionary translation approach

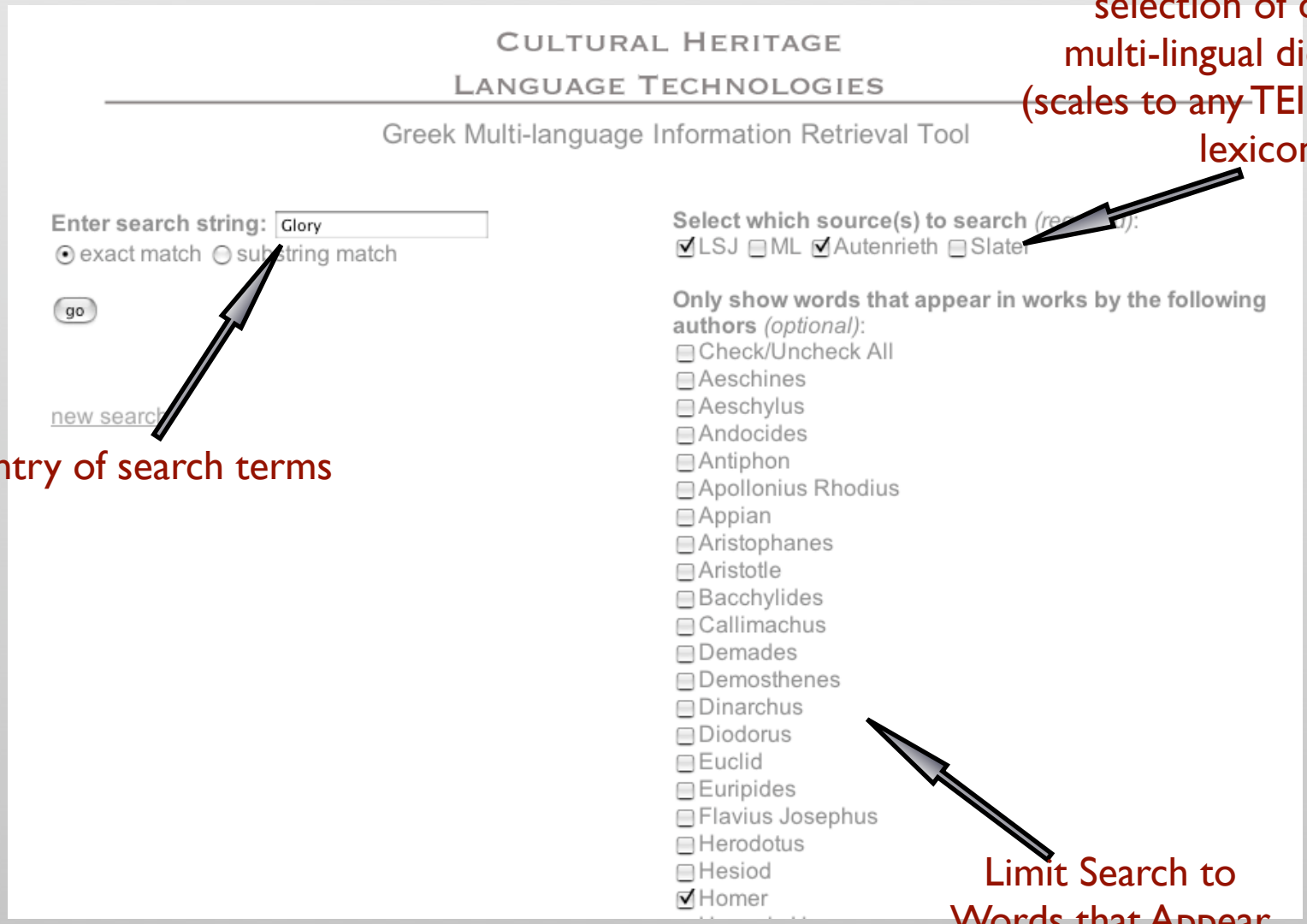
Are Readers of Greek/ Latin/Old Norse Different?

- 'Hyper-fit' for the general profile of a user of Cross-Lingual Information Retrieval tools
- Even very few experts speak/write the languages
- Highly specialized close readers
- Accustomed to philological approach (i.e. *metis* -> *Cunning Intelligence in Ancient Greece*)
- Well versed in reference works
  - Martin Mueller "Very few readers know ancient Greek well enough to read it without frequent recourse to a dictionary or grammar, and because of their highly specialized interests, the few readers who can do so are likely to be particularly intensive users of such reference works." (*Ariadne* 25)

Approaches for Humanists

- Salton in 1972 argued that a carefully constructed translation thesaurus could be almost as effective as mono-lingual IR
- This has been rejected in IR community because of demands of domain independence and scalability
- This objection doesn't apply in the our domains; it is reasonable to ask Greek/Latin/Old Norse readers to construct their own ad hoc translation thesauri (really just 'relevance feedback')
- Relevance feedback can be enhanced with statistical measures

Query Entry



Allow for the selection of different multi-lingual dictionaries (scales to any TEI conformant lexicon)

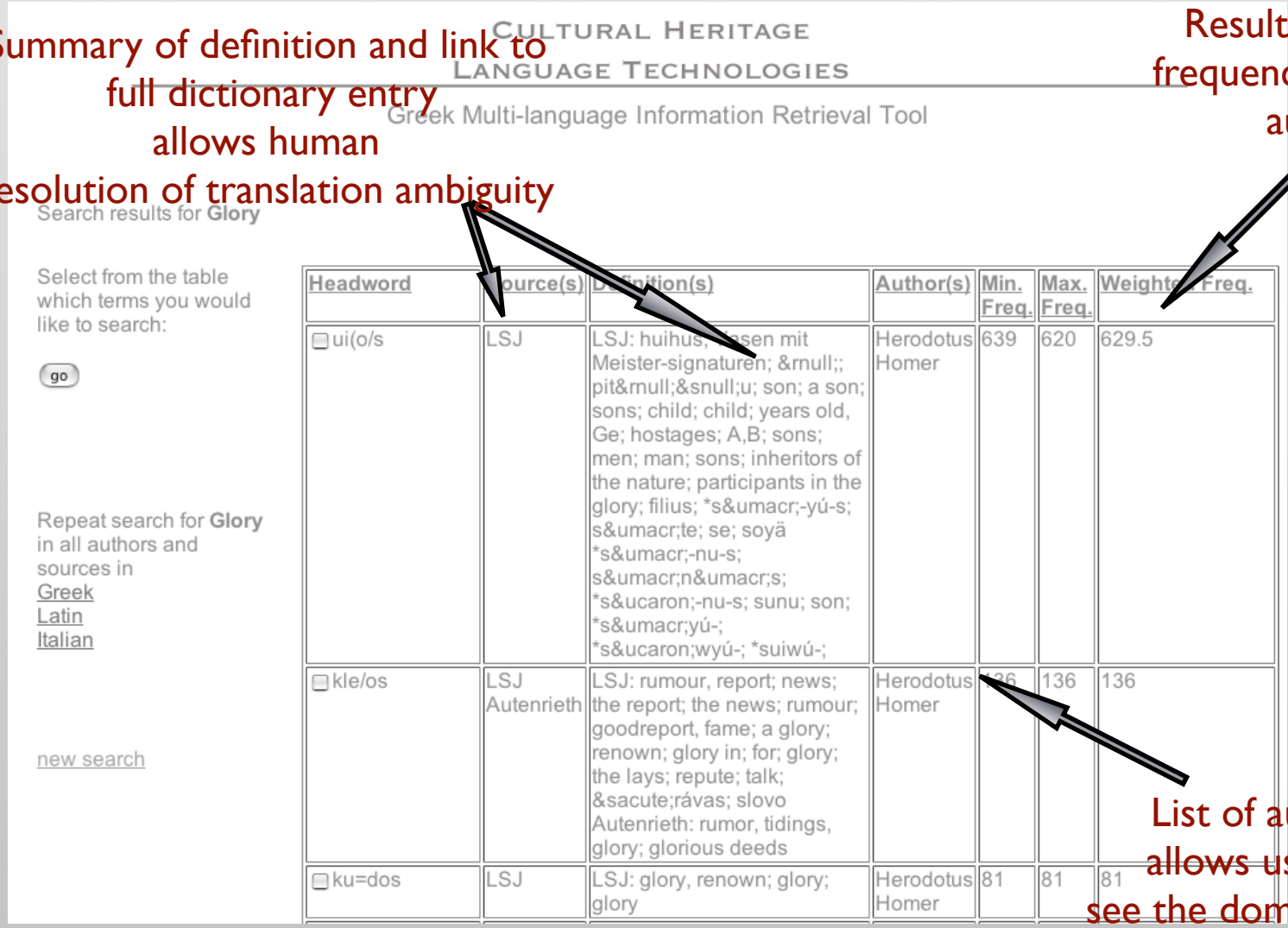
Entry of search terms

Limit Search to Words that Appear in Works by Specific Authors

User Feedback Step

- Interface provide detailed information to help user build the translation thesaurus
- User needs can determine level of disambiguation, from simply removing obvious extraneous terms to careful consideration of each term
- After user refinement, query is passed to monolingual search tool

User Feedback Interface for Greek



To Do List

- Integrate automatically extracted translation equivalents as data source alongside dictionaries
- Add support for multi-word searching and phrases
  - Extend the interface
  - Suggest phrases based on collocation data
- Integrate clustered definitions to help users identify semantic ranges

Visualization of Search Results

- After user provides feedback to select Greek, Latin, or Old Norse query terms, query is passed to search engine (currently MG) and visualization facilities.
- Visualization tool developed for CHLT and integrated with Perseus text display system
- Also will be part of release 3 of the Greenstone Digital Library

Traditional List

Link to full text in DL

List sorted by quality of match and then by document name

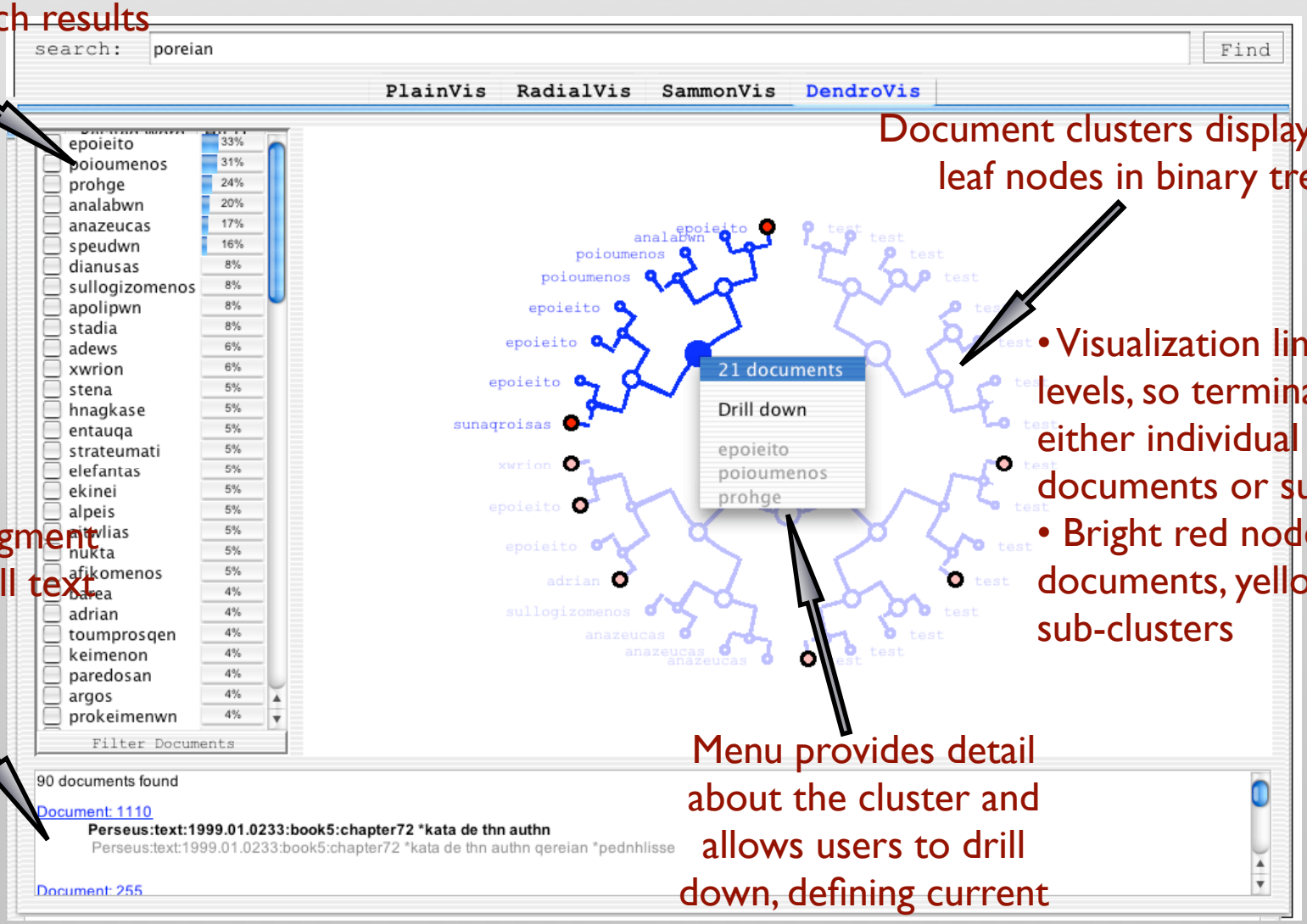
search: logos

PlainVis | RadialVis | SammonVis | DendroVis

Document Number	Sample Text
462	Perseus:text:1999.01.0233:book22:chapter2 *oti kata thn ogdohn kai m# olumpiada
1186	Perseus:text:1999.01.0233:book6:chapter5 *akribesteron men oun isws o peri ths k
121	Perseus:text:1999.01.0233:book12:chapter26d *to d' auto kai *timaiwi sumbebhke p
43	Perseus:text:1999.01.0233:book10:chapter48 VIII. Res Asiae*oi d' *apasiakai kato
461	Perseus:text:1999.01.0233:book22:chapter19 *oti *filopoimhn pros *arxwna ton str
1025	Perseus:text:1999.01.0233:book4:chapter85 *makedonas polemon. to men oun prwton
235	Perseus:text:1999.01.0233:book16:chapter39 VII. Res Asiae*marturei toutois hmwn
345	Perseus:text:1999.01.0233:book1:chapter5 *upoqhsomeqa de tauths arxhn ths bublou
153	Perseus:text:1999.01.0233:book14:chapter12 *isws de tines epaporhsousi pws hmeis
777	Perseus:text:1999.01.0233:book36:chapter10 *peri men oun *rwmaiwn kai *karxhdoni
779	Perseus:text:1999.01.0233:book36:chapter12 *ou xrh de qaumazein ean pote men twi
574	Perseus:text:1999.01.0233:book29:chapter25 *oti ews men tinos oi peri ton *andrw
776	Perseus:text:1999.01.0233:book36:chapter1 A. Olymp. 157, 3. I. Bellum Punicum Te
142	Perseus:text:1999.01.0233:book13:chapter2 *oti *skopas *aitwlwn strathgos apotux
743	Perseus:text:1999.01.0233:book33:chapter16 II. Bellum Rhodiorum Cum Cretensibus*
296	Perseus:text:1999.01.0233:book18:chapter7 *tauta de dialexqeis pros tous allous
61	Perseus:text:1999.01.0233:book11:chapter19 *tis ouk an epishmhnaito thn hgemonia
808	Perseus:text:1999.01.0233:book38:chapter22 *o de *skipiwn polin orwn ... tote ar
339	Perseus:text:1999.01.0233:book1:chapter44 *oi d' en thi *karxhdoni toutwn men ou
297	Perseus:text:1999.01.0233:book18:chapter8 *ths d' hmeras hdh proagoushs epi polu
1036	Perseus:text:1999.01.0233:book5:chapter105 *to men oun *eqelous toiauta dialexqei

Tree View

Interface to filter words out of search results



Document clusters displayed as leaf nodes in binary tree

- Visualization limited to 5 levels, so terminal nodes can be either individual documents or sub-clusters
- Bright red nodes are documents, yellow nodes are sub-clusters

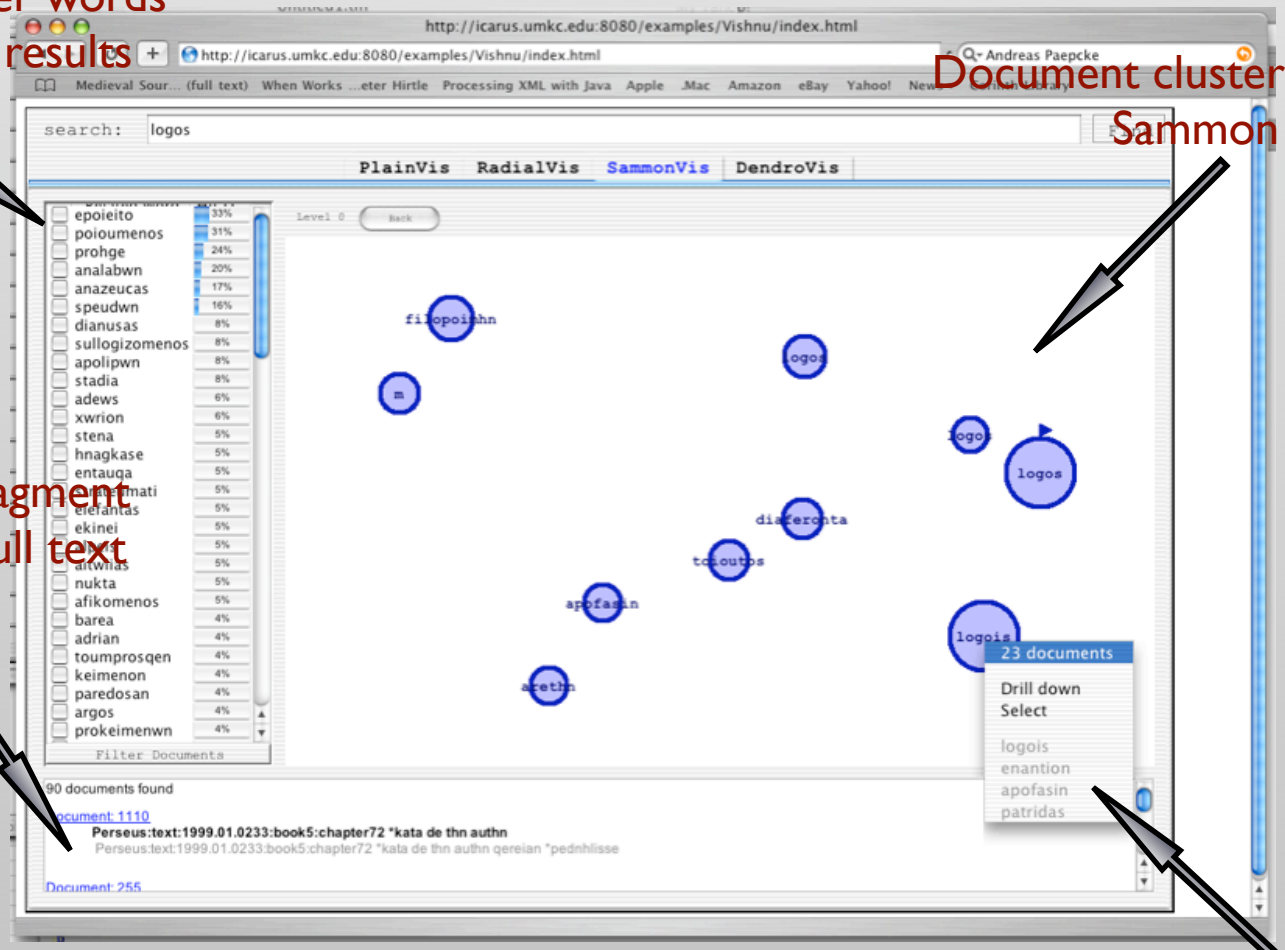
Document fragment with link to full text in DL

Menu provides detail about the cluster and allows users to drill down, defining current node as the center of the new cluster

Sammon Map

Interface to filter words out of search results

Document clusters displayed on Sammon Map



Document fragment with link to full text in DL

Menu provides detail about the cluster and allows users to drill down, defining current node as the center of the new cluster

Where Next?

- Tighter integration of tools
- User evaluations of clustering tools and OAI federation
- Parse Trees and phrase discovery