The background of the slide is a Romantic-style landscape painting. It depicts a calm body of water, possibly a lake or a wide river, reflecting the warm, golden light of a low sun. The sun is positioned in the upper center, creating a bright, hazy glow that fills the sky and reflects on the water's surface. The water is still, with only a small, dark boat visible in the lower left quadrant. The foreground and middle ground are framed by dense, dark foliage and trees, their leaves catching the light and adding texture to the scene. The overall mood is serene and contemplative, typical of 19th-century landscape art.

Exploratory Text Analysis and The Middle Distance

Marti A. Hearst
U.C Berkeley

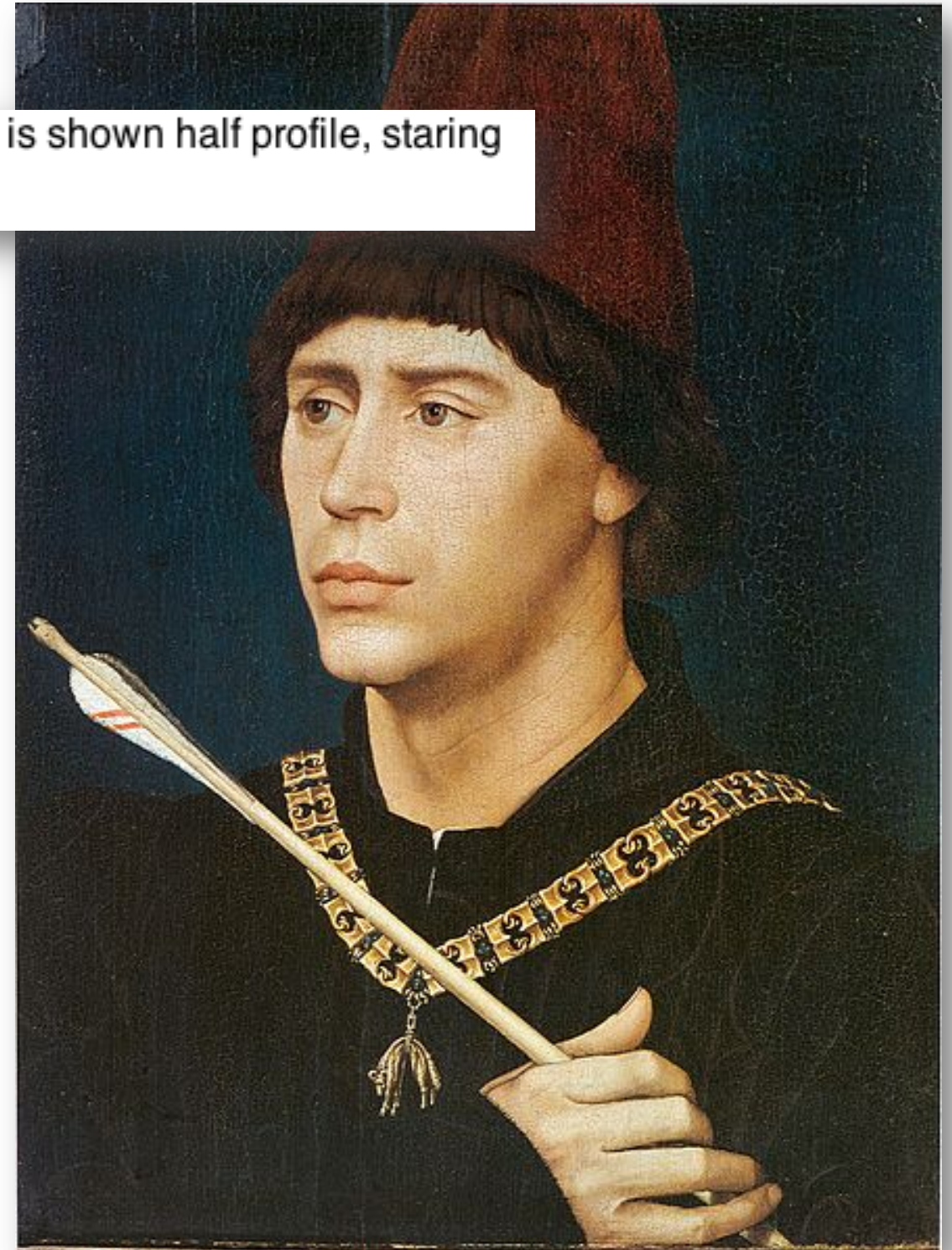
Joint Work with Aditi Muralidharan

Collaborators: Bryan Wagner, Chris Fan, Rex Ganding

Sponsored by NEH HK50011

Usage: “Middle Distance” in Portraiture

In common with most of van der Weyden's male portraits, Antoine is shown half profile, staring aloofly into the **middle distance**.



Portrait of Antoine, 'Grand Bâtard' of Burgundy, van der Weyden (Wikipedia)

Outline

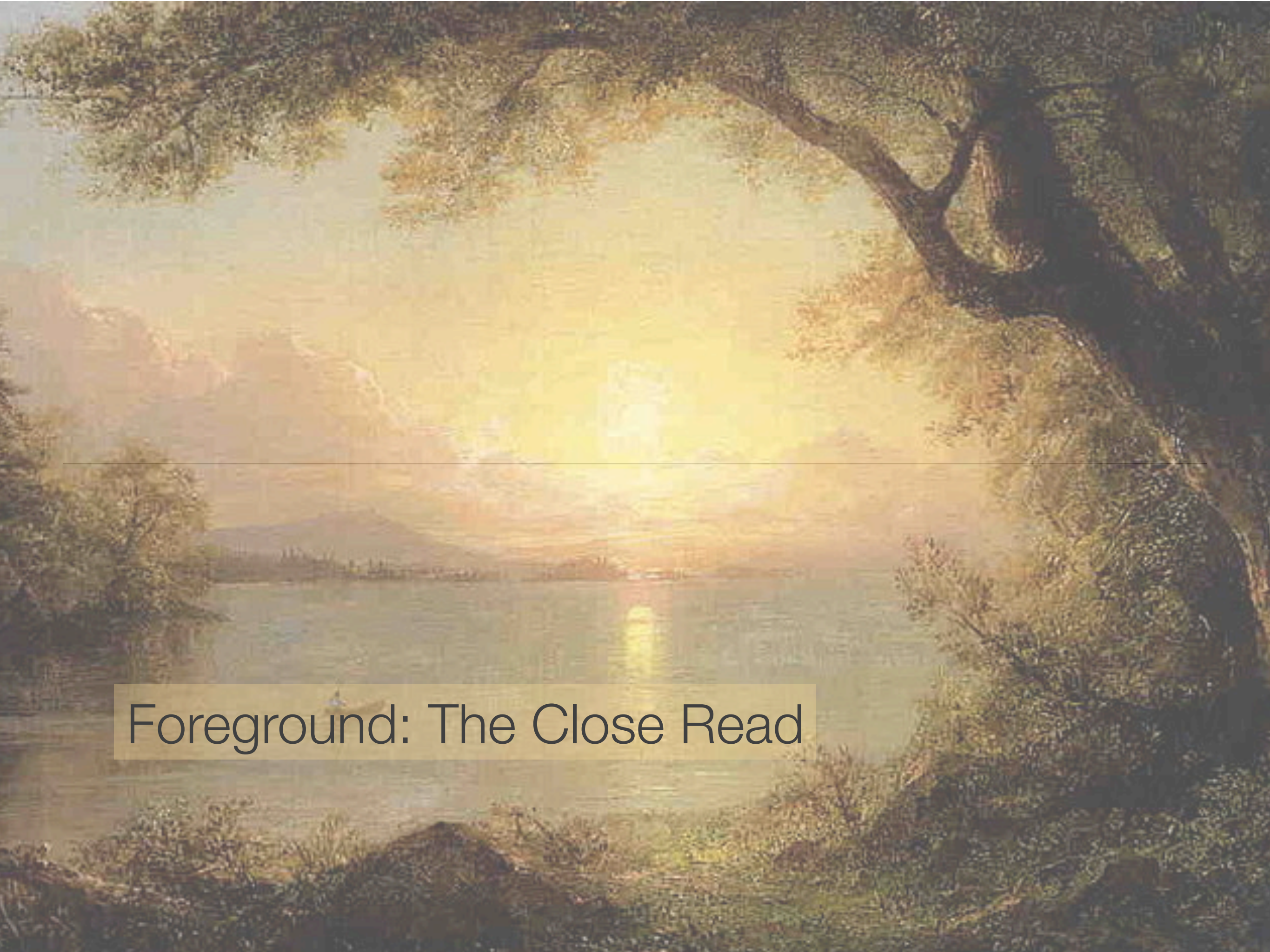
- Close Reading
- Distant Reading / Culturomics / Text Mining
- Exploratory Text Analysis : What is Needed & Related Work
- WordSeer: Case Studies
- What Remains to be Done

A Romantic-style landscape painting. In the foreground, a large, dark, leafy tree stands on the right bank, its branches reaching over the water. The middle ground features a calm lake reflecting the warm, golden light of a low sun. In the background, distant, hazy mountains are visible under a soft, glowing sky. The overall mood is serene and contemplative.

Background: Distant Read (Text Mining)

Middle Distance: Sensemaking

Foreground: The Close Read



Foreground: The Close Read

Definition: “Close Read”

“**Close reading** describes, in literary criticism, the careful, sustained interpretation of a brief passage of text. Such a reading places great emphasis on the particular over the general, paying close attention to individual words, syntax, and the order in which sentences and ideas unfold as they are read.”

-English Wikipedia, 6/4/2012

“Power and Passion in Shakespeare’s Pronouns Interrogating ‘you’ and ‘thou’”

Penelope Freedman, 2007, MPG Books, 280 pp.

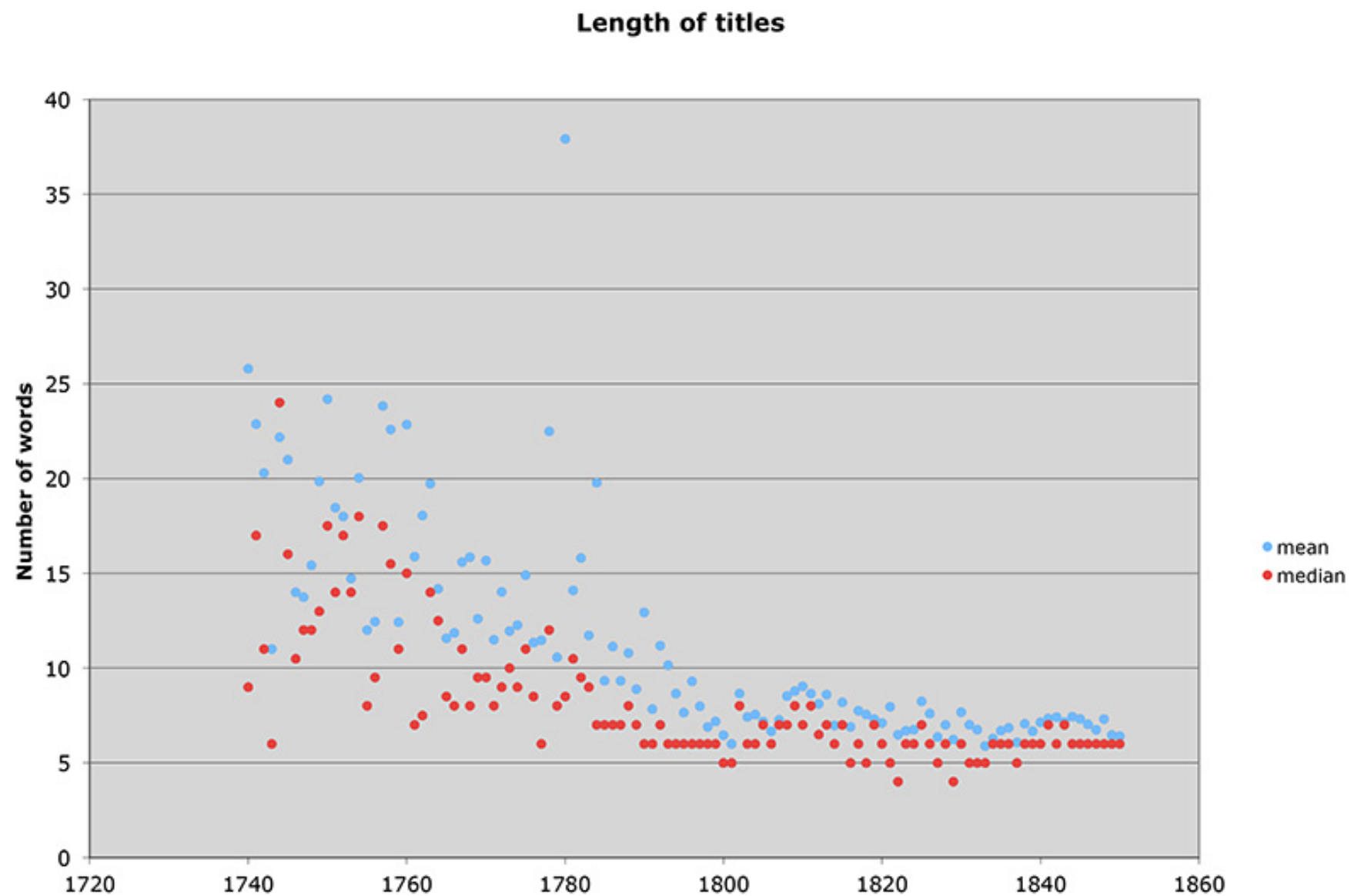
“The subtleties of the use of ‘you’ and ‘thou’
that have emerged ... can seem, at worst,
random or, at best, unfathomable. ...

Far Distance: Distant Read (Text Mining)



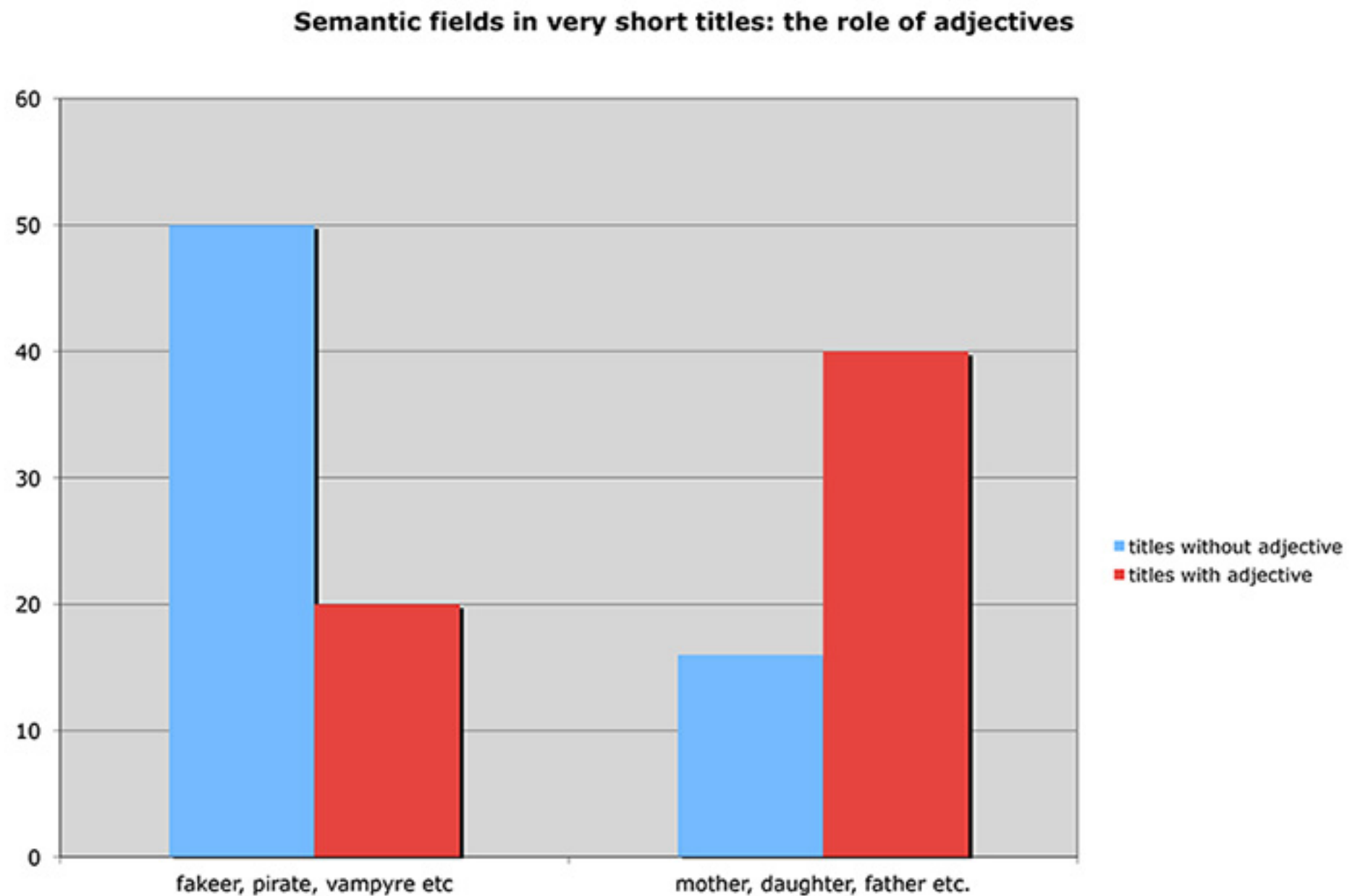
Text Mining in Moretti, “*Distant Reading*”, 2013

Shrinking Book Title Lengths Across the Years



Text Mining in Moretti, “*Distant Reading*”, 2013

Comparing Uses of Adjectives in Book Titles



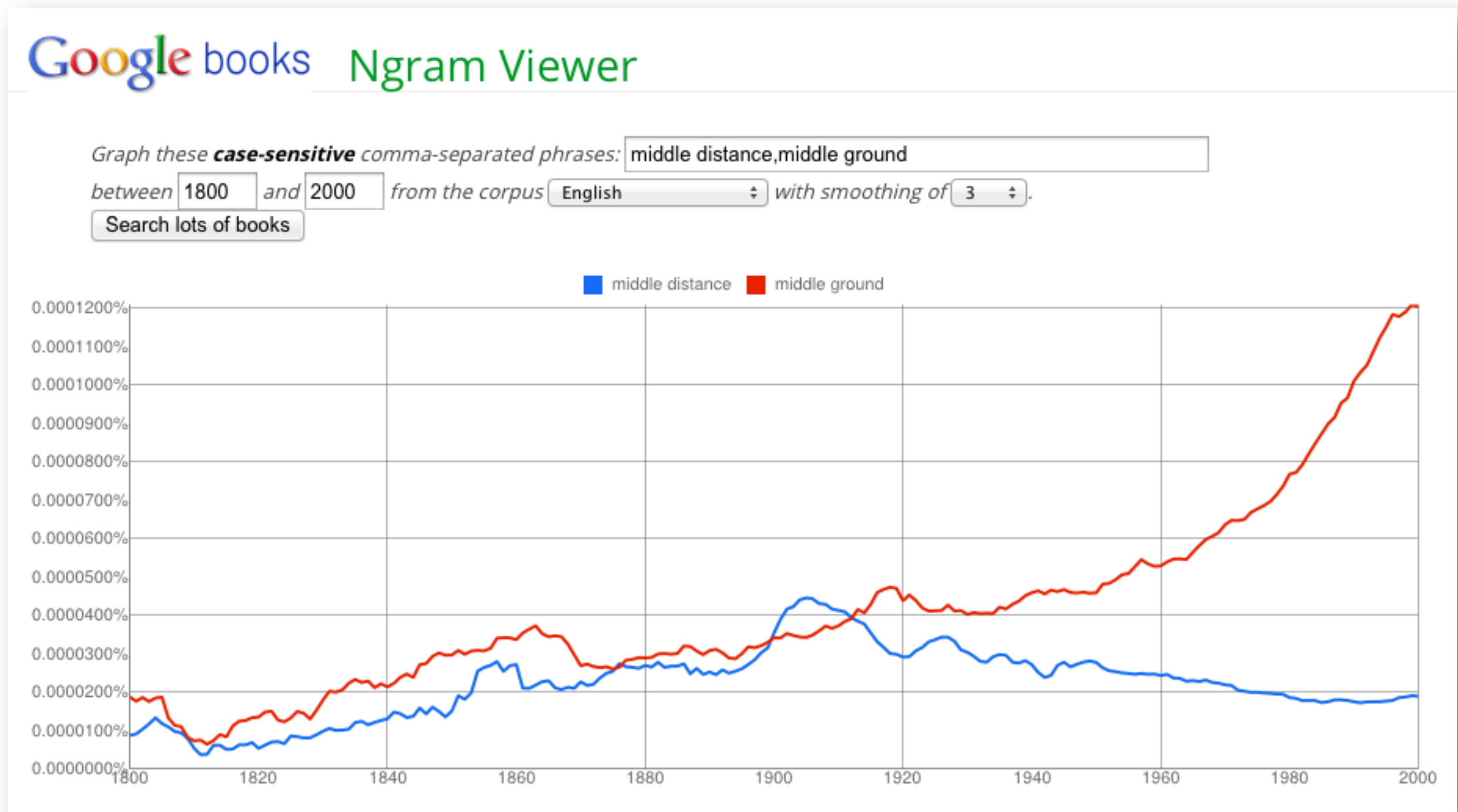


Science

14 January 2011 • \$10

Quantitative Analysis of Culture Using Millions of Digitized Books. Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. *Science* **331** (2011)

“Distant Read” Example: Ngram Viewer applied to *middle distance* vs. *middle ground*



Text Mining:

Statistical Text Analysis since at least 1990

She drank a cup of coffee.

They usually drink coke after work.

Sharon drinks champagne to celebrate.

Text Mining:

Statistical Text Analysis since at least 1990

She drank a cup of coffee.

They usually drink coke after work.

Sharon drinks champagne to celebrate.

Text Mining:

Statistical Text Analysis since at least 1990

She drank a cup of coffee.

They usually drink coke after work.

Sharon drinks champagne to celebrate.

He drank concrete with dinner.

Text Mining:

Statistical Text Analysis since at least 1990

She drank a cup of coffee.

They usually drink coke after work.

Sharon drinks champagne to celebrate.

~~X He drank concrete with dinner.~~

Text Mining:

Statistical Text Analysis since at least 1990

Kenneth Church and Patrick Hanks

Table 5. What Can You Drink?

| Verb | Object | Mutual Info | Joint Freq |
|----------------|---------------------|-------------|------------|
| <i>drink/V</i> | <i>martinis/O</i> | 12.6 | 3 |
| <i>drink/V</i> | <i>cup_water/O</i> | 11.6 | 3 |
| <i>drink/V</i> | <i>champagne/O</i> | 10.9 | 3 |
| <i>drink/V</i> | <i>beverage/O</i> | 10.8 | 8 |
| <i>drink/V</i> | <i>cup_coffee/O</i> | 10.6 | 2 |
| <i>drink/V</i> | <i>cognac/O</i> | 10.6 | 2 |
| <i>drink/V</i> | <i>beer/O</i> | 9.9 | 29 |
| <i>drink/V</i> | <i>cup/O</i> | 9.7 | 6 |
| <i>drink/V</i> | <i>coffee/O</i> | 9.7 | 12 |
| <i>drink/V</i> | <i>toast/O</i> | 9.6 | 4 |
| <i>drink/V</i> | <i>alcohol/O</i> | 9.4 | 20 |
| <i>drink/V</i> | <i>wine/O</i> | 9.3 | 10 |
| <i>drink/V</i> | <i>fluid/O</i> | 9.0 | 5 |
| <i>drink/V</i> | <i>liquor/O</i> | 8.9 | 4 |
| <i>drink/V</i> | <i>tea/O</i> | 8.9 | 5 |
| <i>drink/V</i> | <i>milk/O</i> | 8.7 | 8 |
| <i>drink/V</i> | <i>juice/O</i> | 8.3 | 4 |
| <i>drink/V</i> | <i>water/O</i> | 7.2 | 43 |
| <i>drink/V</i> | <i>quantity/O</i> | 7.1 | 4 |

Word association norms, mutual information, and lexicography
KW Church, P Hanks - Computational linguistics, 1990

Text Mining: Semantic Relation Detection

- Goal: automatically augment a lexical database
- Many potential relation types:
 - ISA (hypernymy/hyponymy)
 - Part-Of (meronymy)
- Idea: find unambiguous contexts which (nearly) always indicate the relation of interest

Text Mining: Semantic Relation Detection

(S1) Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

(1a) NP_0 such as $NP_1 \{, NP_2 \dots, (and \mid or) NP_i\} \quad i \geq 1$

are such that they imply

(1b) *for all $NP_i, i \geq 1$, $hyponym(NP_i, NP_0)$*

Thus from sentence (S1) we conclude

$hyponym(\text{“Gelidium”}, \text{“red algae”})$.

Lexico-Syntactic Patterns

(2) *such NP as {NP ,} * {(or | and)} NP*

... works by such authors as Herrick, Goldsmith, and Shakespeare.

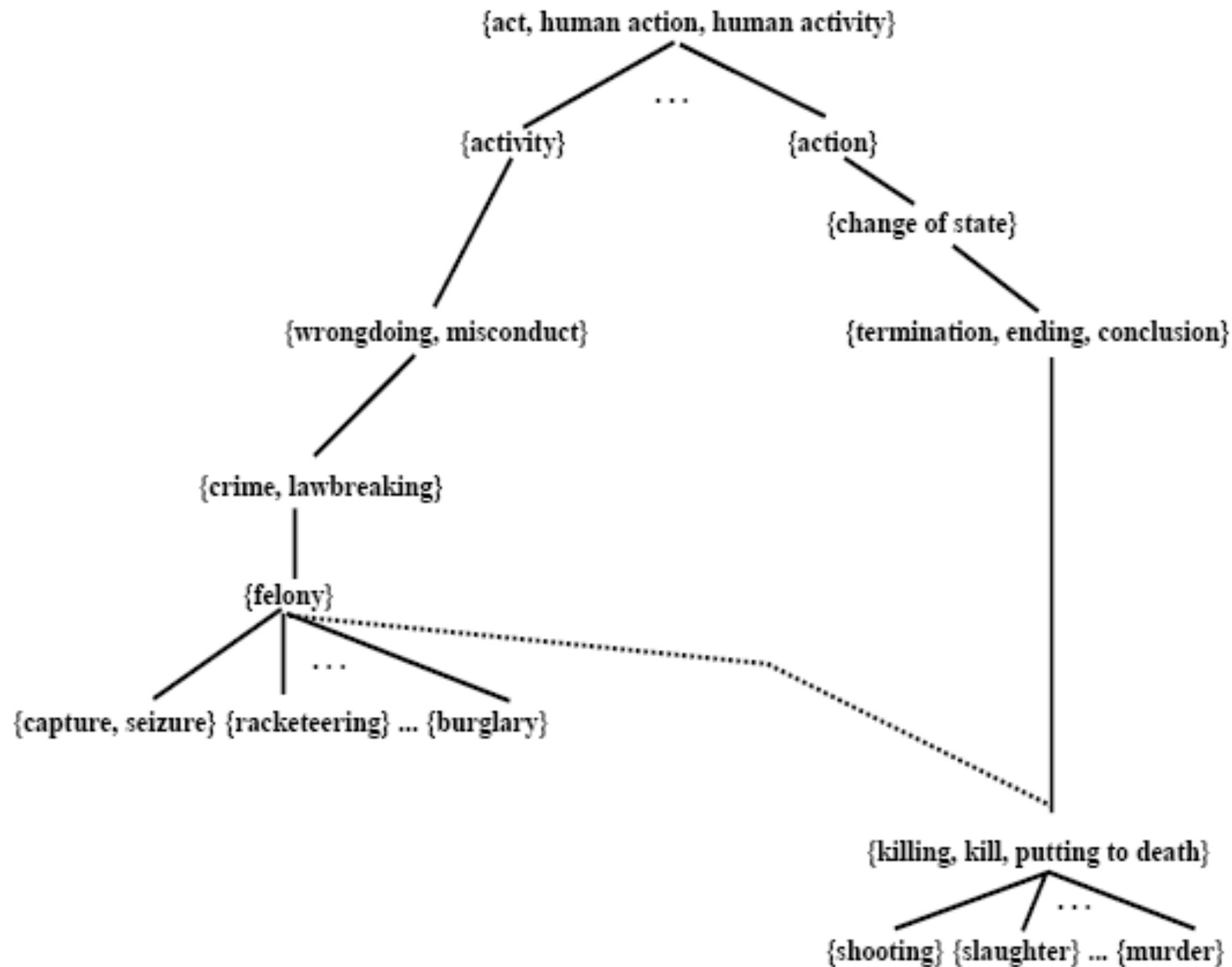
\Rightarrow *hyponym*("author", "Herrick"),
hyponym("author", "Goldsmith"),
hyponym("author", "Shakespeare")

(3) *NP {, NP} * {,} or other NP*

Bruises, ..., broken bones or other injuries ...

\Rightarrow *hyponym*("bruise", "injury"),
hyponym("broken bone", "injury")

Text Mining: Adding a New Relation to a Lexicon



A Romantic-style landscape painting. In the foreground, a large, dark tree with dense foliage frames the right side of the image. The middle ground features a calm body of water reflecting the warm, golden light of a low sun. A small boat with a single figure is visible on the water. In the background, a range of mountains is visible under a hazy, golden sky. The overall mood is serene and majestic.

Middle Distance: Exploratory Text Analysis

In life, we tend to focus on the endpoints, not the middle.

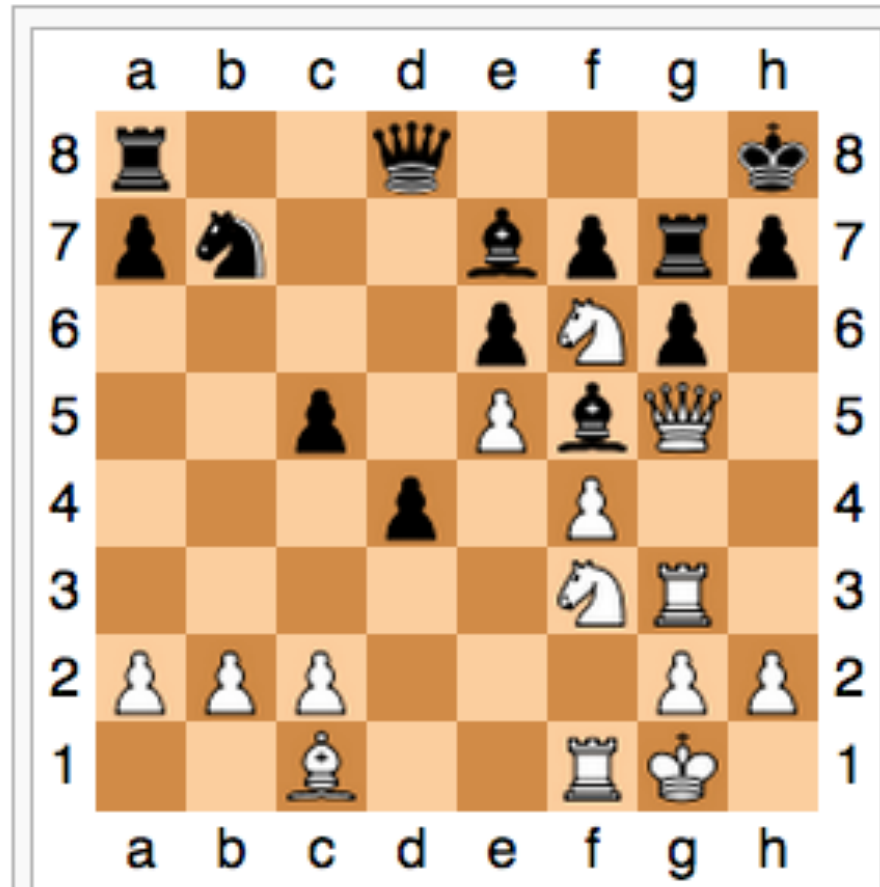
Chess:

Opening

Endgame

Middlegame

?



Middlegame position from the game
**Joseph Henry Blackburne -
Siegbert Tarrasch, Breslau, 1889.**
Last move of White - 26.Qh6-g5,
next move of Black - 26...Nb7-d6.

Ngram Viewer on the Chess Metaphor

Google books Ngram Viewer

Graph these **case-sensitive** comma-separated phrases:

between and from the corpus with smoothing of .

[Search lots of books](#)

[G+ Share](#) 0

[Tweet](#) 0

Ngrams not found: chess middlegame

The Ngram Viewer is case sensitive. Check your capitalization!



WordSeer:

Exploratory Text Analysis at the Middle Distance

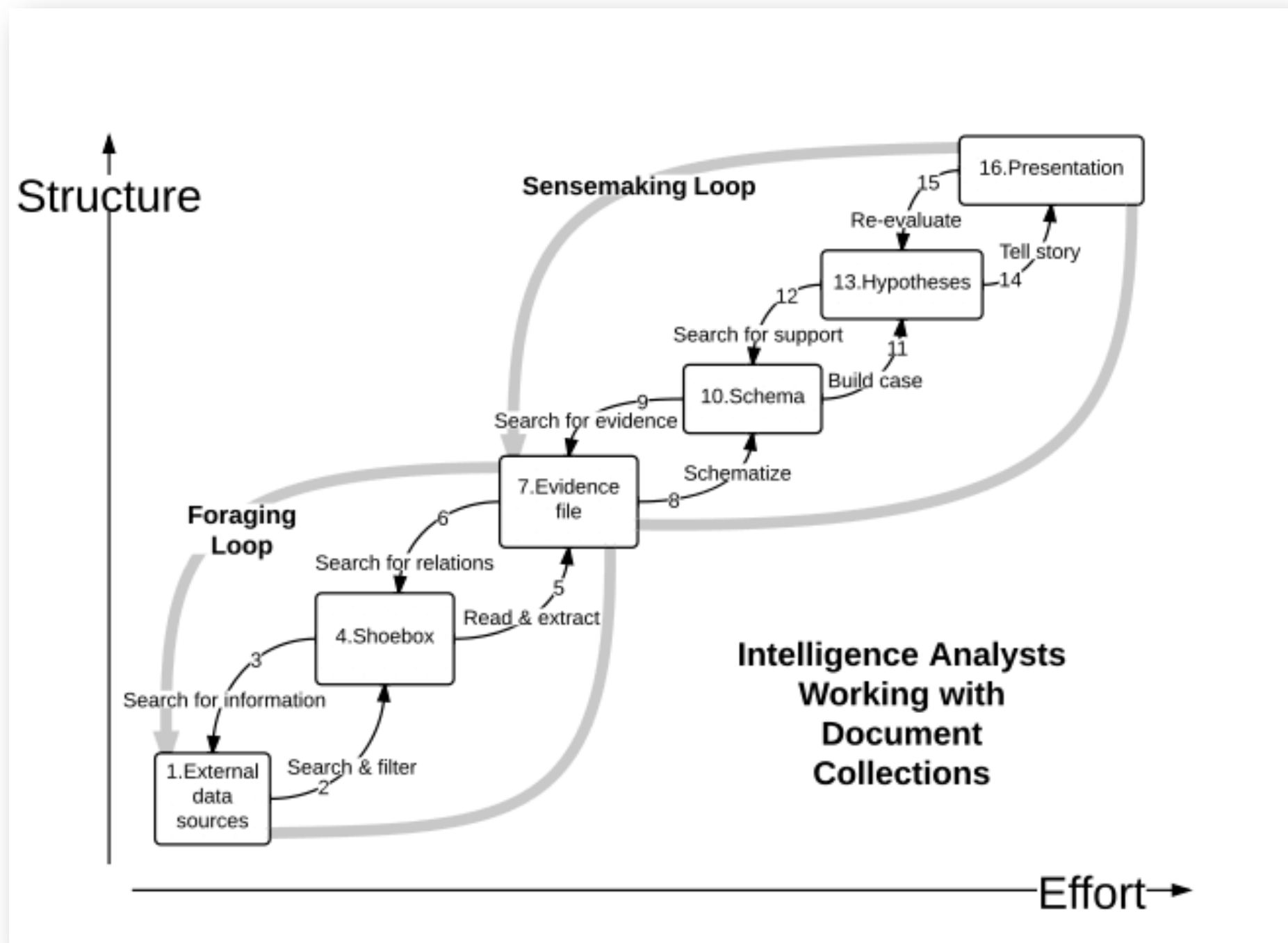
- Goal: help scholars analyze literary text.
- Method: combine natural language processing, information visualization, and search user interface design.
 - Support the “middle game”
 - Midway between close read and distant statistics.
 - Help with hypothesis formulation, verification, and refinement.
- New Goal: Help with Qualitative Coding!

WordSeer Motivation:

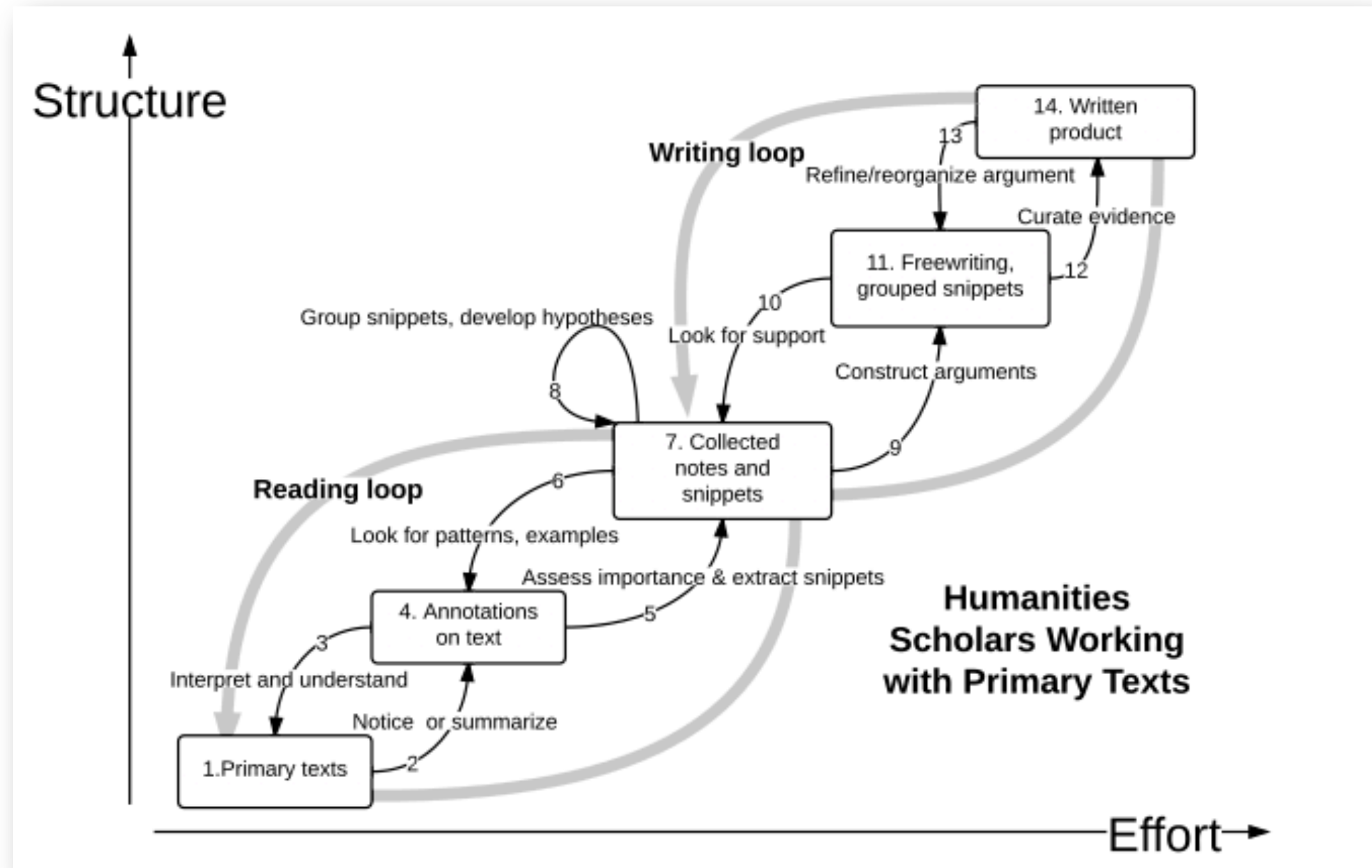
What do Literary PhDs Want to Do?

- Interviewed 12 PhD students in literature and history departments
- Some key requirements:
 - Annotating while reading (12/12)
 - Looking for something specific in the text (12/12)
 - Different ways in which a concept is discussed (7/12),
attractive, slim, tall
 - How common a concept is (3/12)
 - “I’ve been trying to find out if anyone has ever been positive towards the ‘new woman’, being approving of her. I would love to see how often that phrase shows up during the 1960s” – P12
 - Syntactic patterns and regularities (2/12)
 - “I look for grammatical patterns, clauses, or any kind of list. Like, “I went to the post office, the supermarket, and church”. – P1

Intelligence Analysts' SenseMaking Loop (Pirolli, Russell, Card)



Digital Humanities SenseMaking Loop



WordSeer Interaction Paradigm: “Text Sliding”

- A **slice** is a set of sentences, and
- A **view** is a visual representation of the data in a slice
 - Including a list of sentences, a visualization, data charts
- **Text sliding** moves from one view of a slice to another, including:
 - **Showing** a different view of the same slice, or
 - **Narrowing** (by selecting metadata or other filters), or
 - **Broadening** (by removing filters), or
 - **Creating** a new slice (moving laterally) including:
 - Slicing on one of the words from the context
 - Slicing on a related word from a word in the context.
 - Slicing based on grammatical relations of a word in the context.

WordSeer Video Demo

All DocumentsWord Trees

love

negation

▼

not

▼

with stemming: ☐

Grammatical relation

▼

in new tab

▼

Go

test

Shakespeare's Works

Frequent Words

PHRASES

good lord (391) sir john (158) come come (143) let see (124)

fare well (119) give leave (91) enter king (90) know st (88)

good master (82) give hand (81) tis true (78) " say (67)

come let (64) king henry (59) god save (57)

enter messenger (57) come sir (56) make haste (54)

well said (53) draw sword (51)

NOUNS

lord (2089) thy (1904) man (1809) sir (1724) thee (1614)

thou (1344) love (1249) king (1225) father (972) heart (921)

time (908) death (823) art (810) life (797) hand (764)

exit (694) day (664) ti (636) ay (635) lady (625)

Metadata

ACT

Act 4 (37) Act 3 (37) Act 2 (37) Act 1 (37) Act 5 (37)

Induction (1)

LINES

filter

526

0

1

867

SCENE

Act 2, Scene 1 (37) Act 3, Scene 2 (36) Act 5, Scene 1 (37)

Act 1, Scene 1 (37) Act 1, Scene 2 (35) Act 3, Scene 1 (37)

Act 4, Scene 1 (37) Act 5, Scene 2 (31) Act 2, Scene 2 (34)

Act 4, Scene 2 (35) Act 4, Scene 3 (31) Act 2, Scene 3 (29)

Sets

PHRASE SET

{beauty words} (3935) {ki

{Dog like animals} (318) {

SENTENCE SET

{Canines in Shakespeare}

Related Work:

Exploratory Text Analysis Tools

- EDA on text has different demands than on DBMS's
 - Unstructured information
 - Very high dimensionality of text.
- It was big in the '90's, but is less active now.
 - Most systems focus on recognizing entities and showing relations among them
 - Very few focus on the details of language behavior.

Feldman et al. PAKM'98

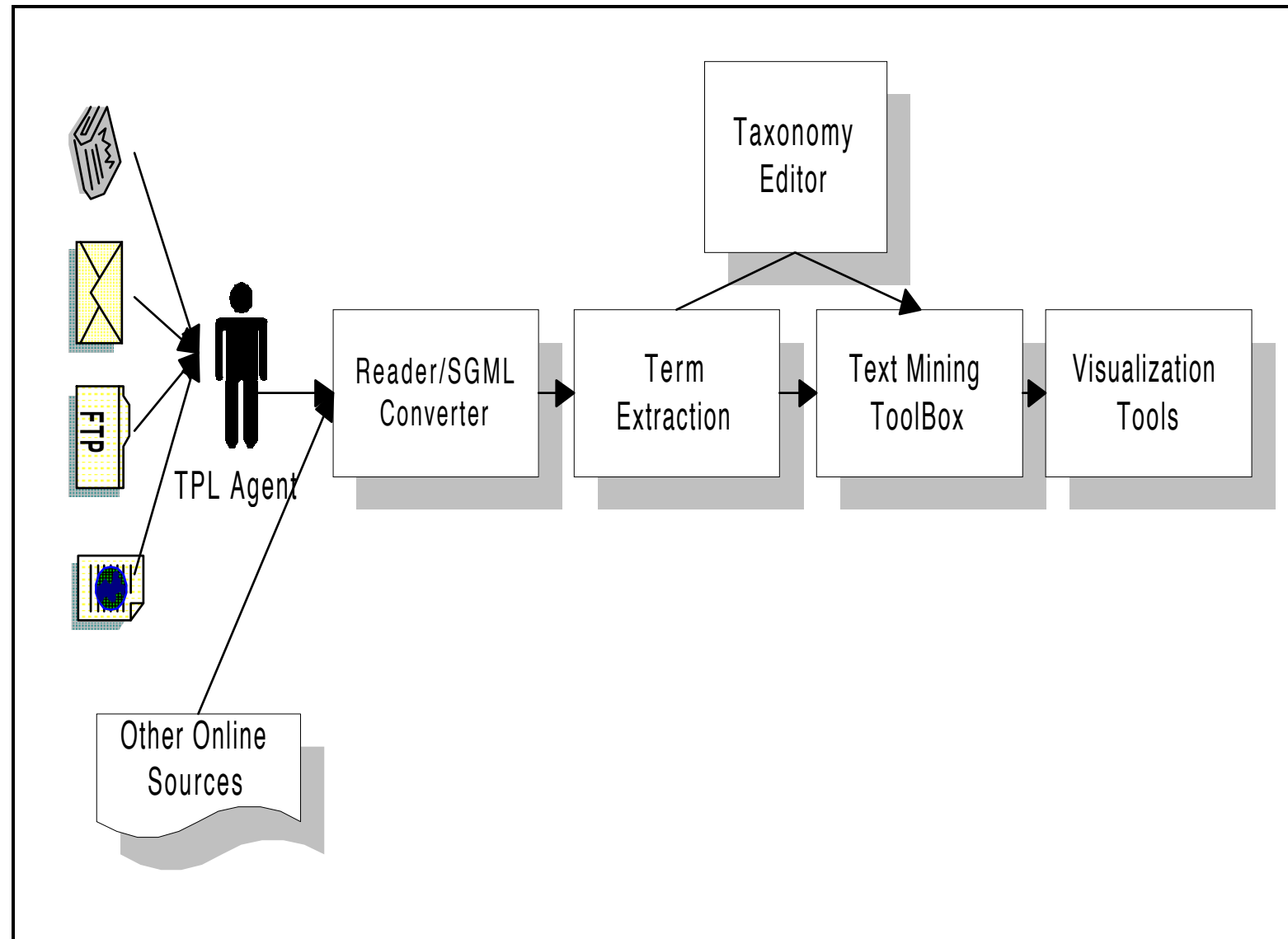
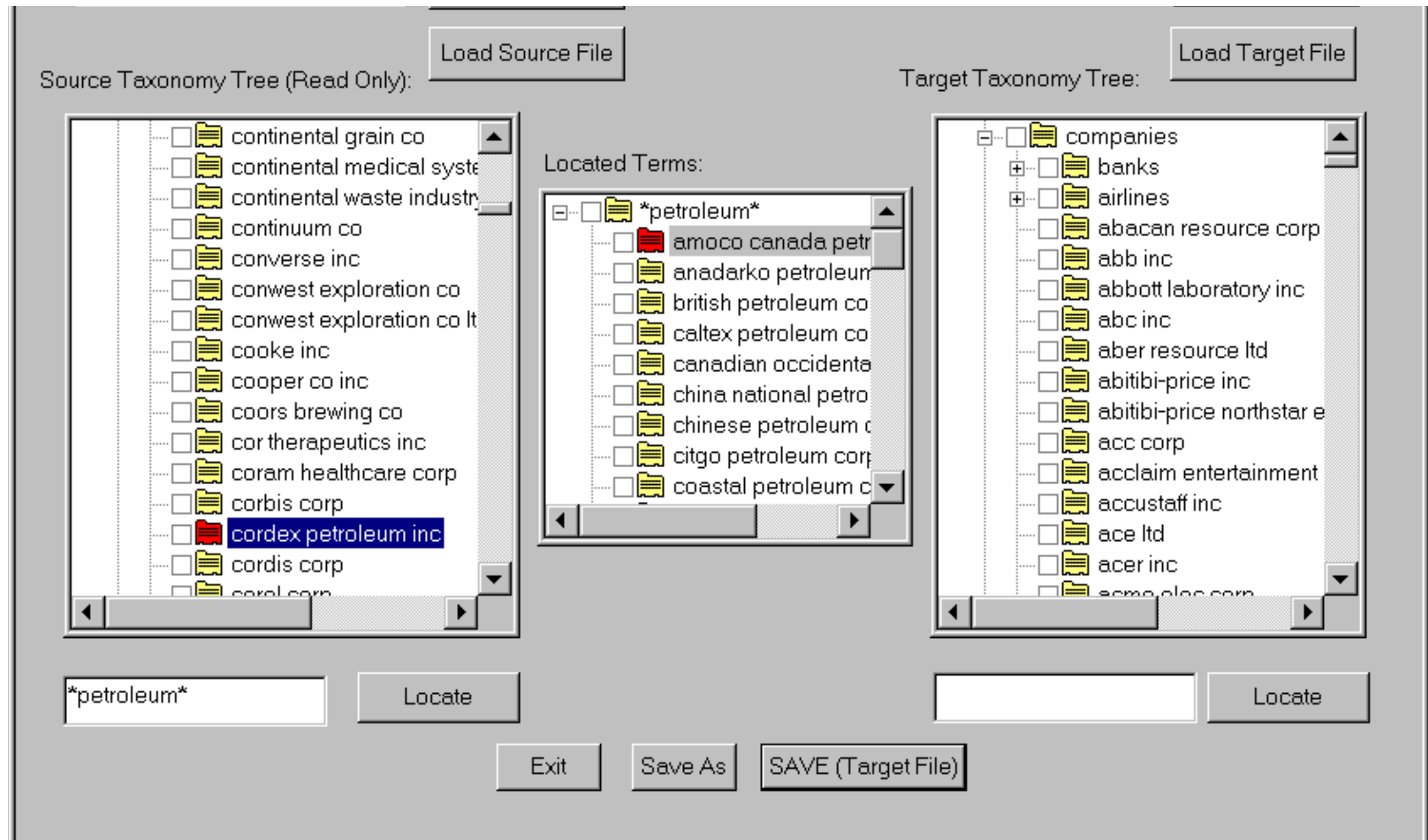
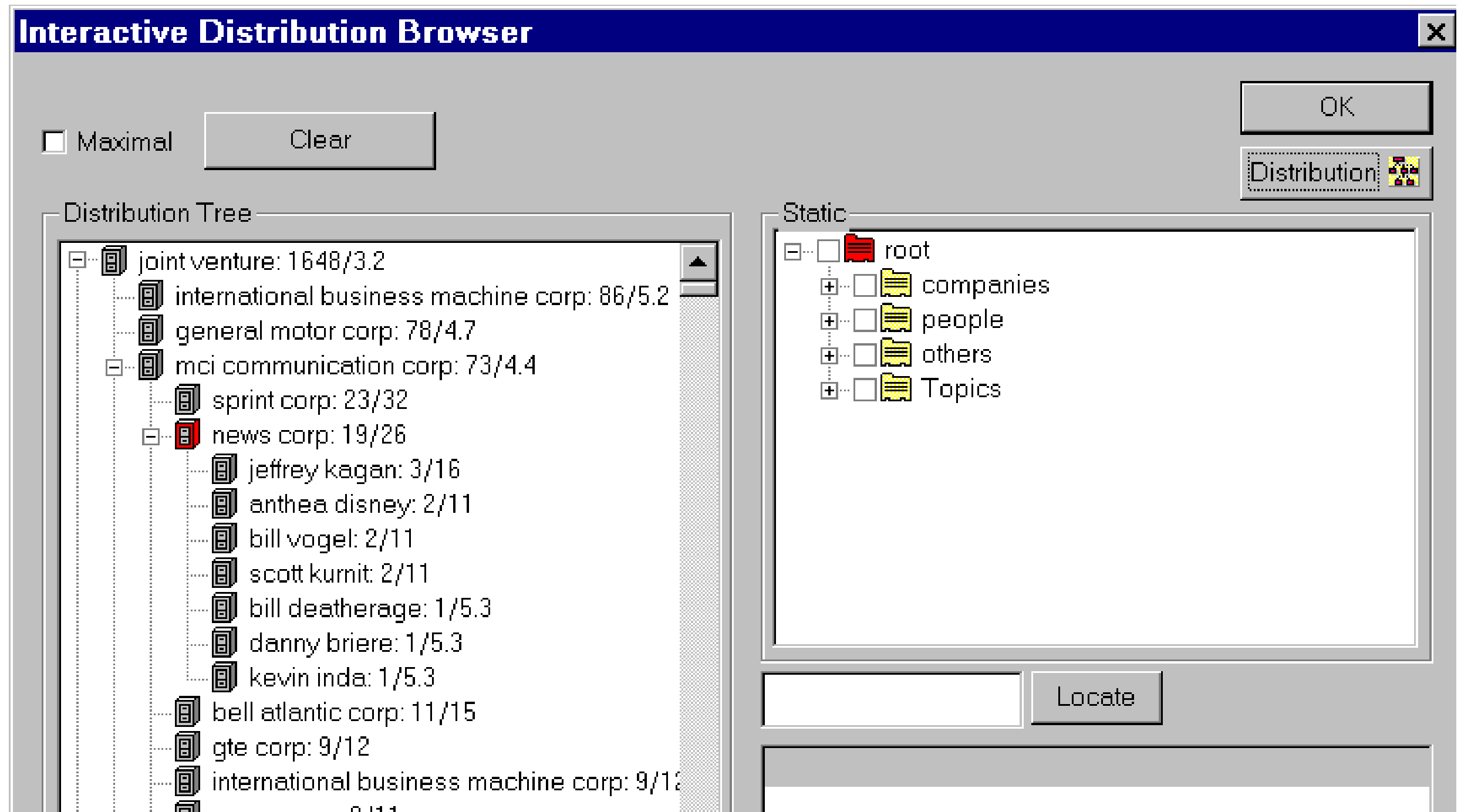


Figure 2. Document Explorer architecture.

Feldman et al. PAKM98



Feldman et al. PAKM98



Feldman et al. PAKM98

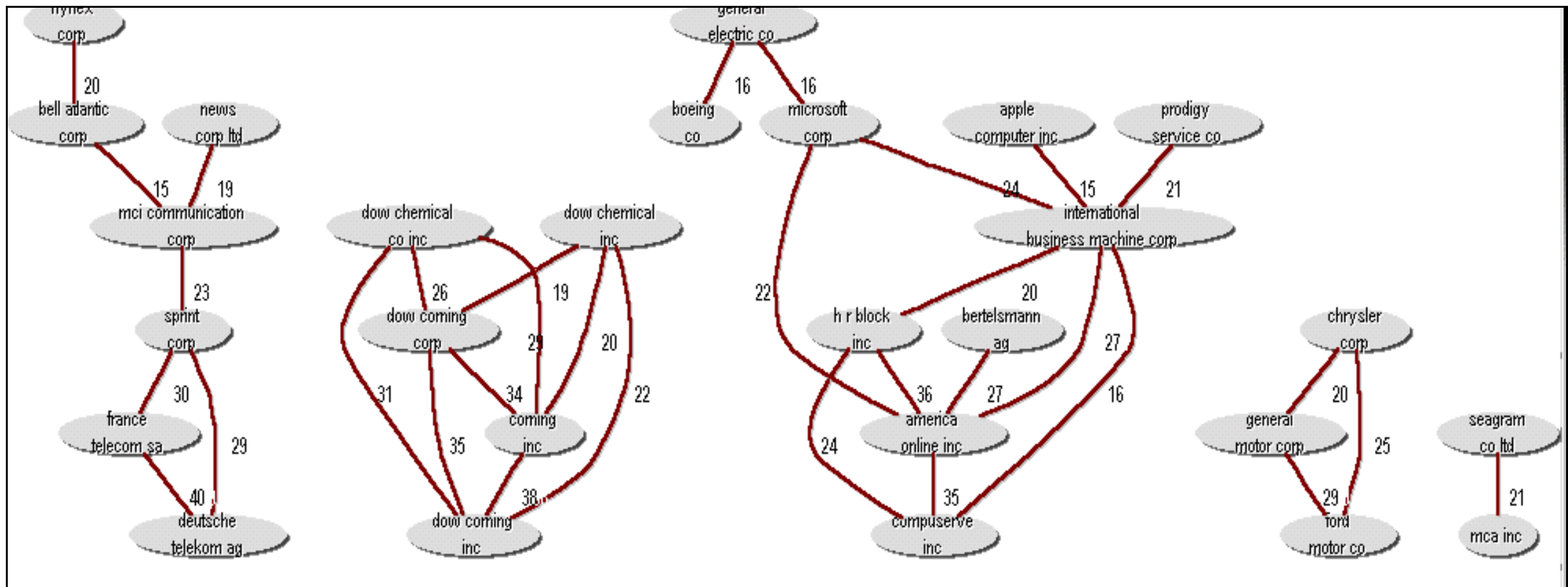


Figure 12 – Context Graph (Companies in Context of “Joint Venture”)

Feldman et al. PAKM98

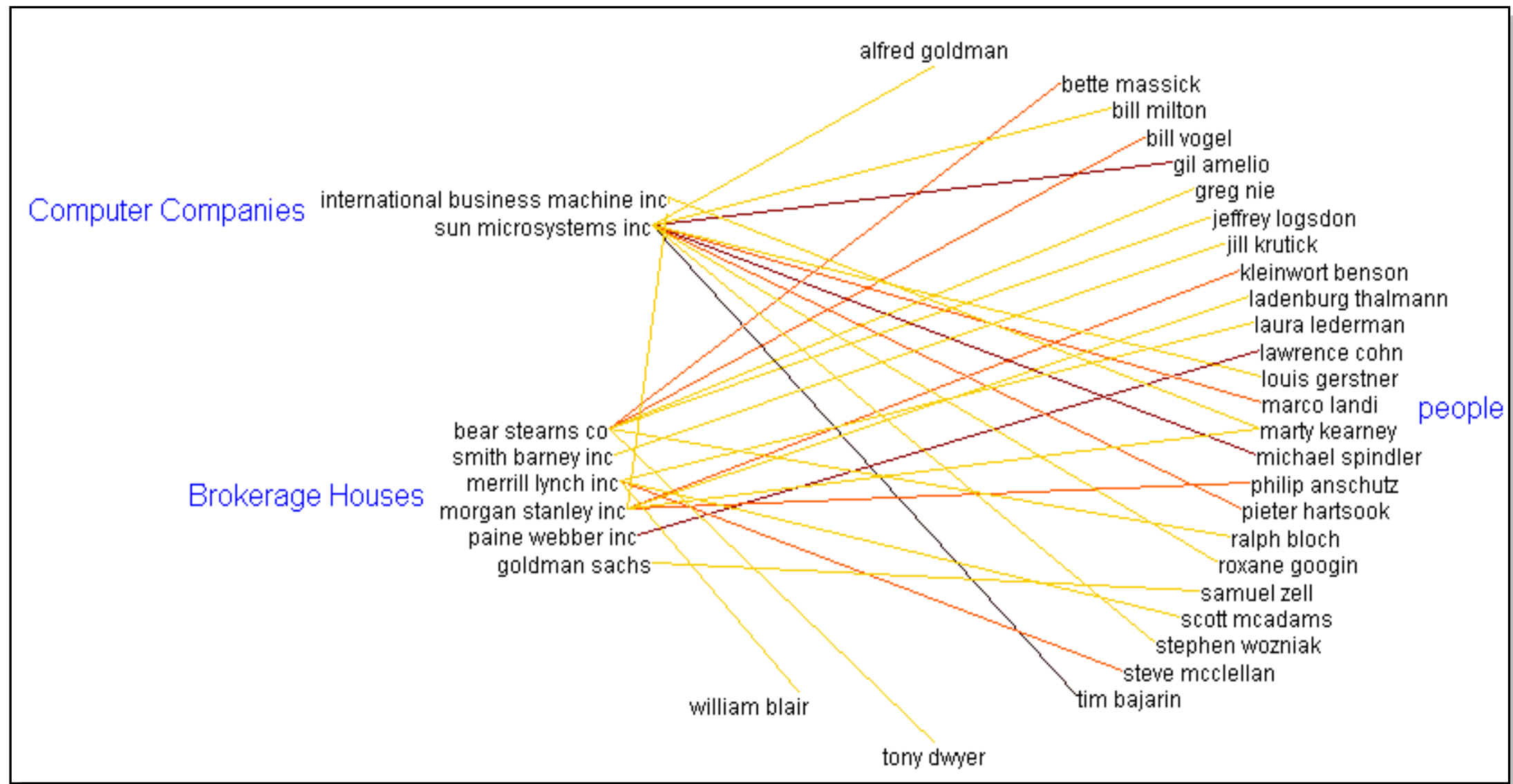
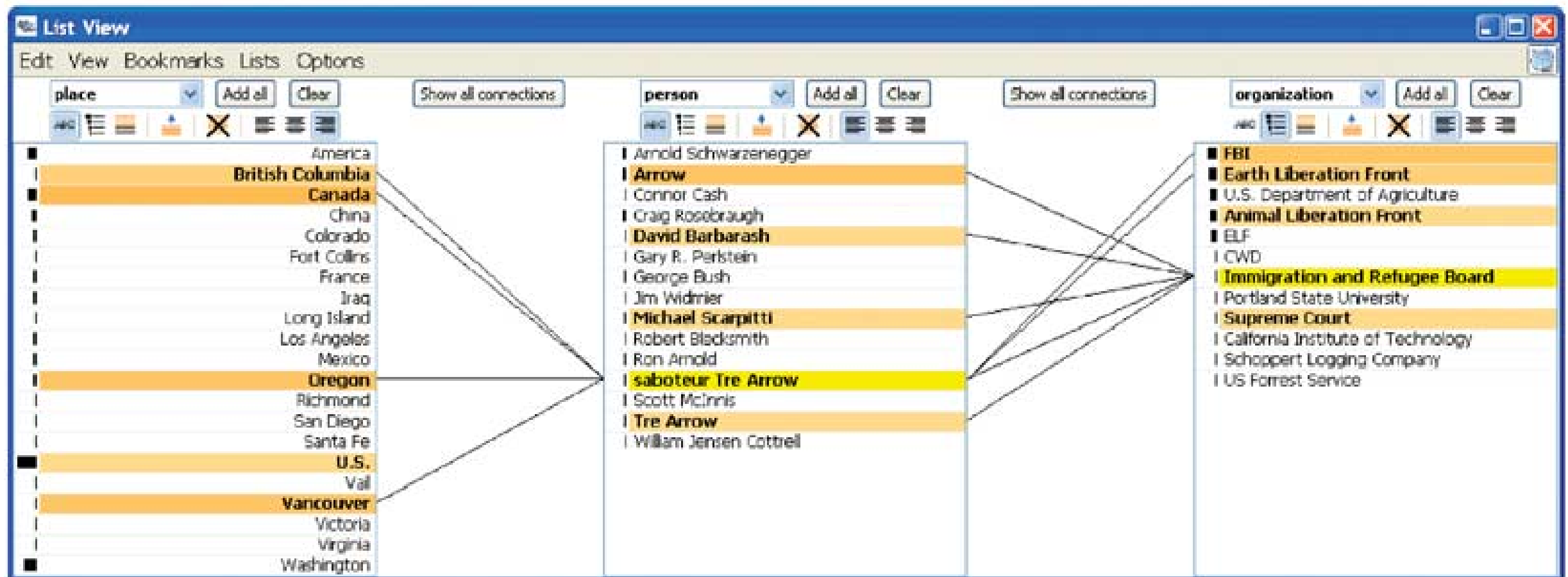


Figure 13 – Category Graph (relationship between People, Brokerage Houses and Computer Companies in Context of “merger”)

Stasko et al., Jigsaw, Information Visualization 2008



Stasko et al., Jigsaw, Information Visualization 2008



Stasko et al., Jigsaw, Information Visualization 2008

arrow arson authorities being canada
instruction elf enforcement environmental federal
known logging members oregon refugee responsibility
terrorist vancouver wanted whether years

Source:
Date: May 14, 2004

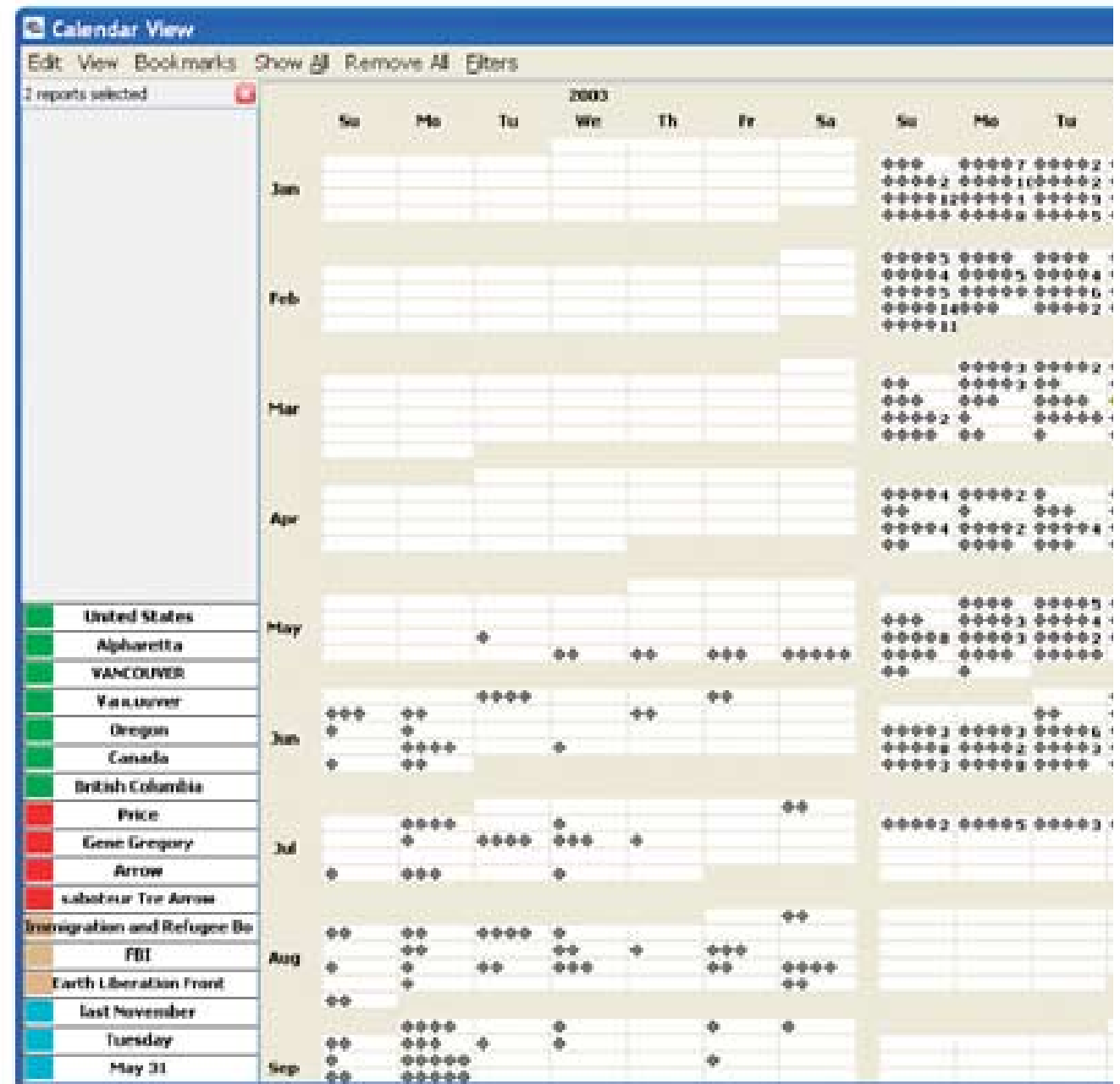
VANCOUVER, British Columbia - A Canadian immigration panel is considering whether accused environmental saboteur Tre Arrow can apply for refugee status in Canada.

Arrow, 30, who is wanted for fire bombing logging and cement trucks in Oregon, asked the Canadian authorities to remain in Canada as a political refugee at a hearing in Vancouver on Tuesday.

A key issue will be whether Arrow is affiliated with a terrorist group, which would immediately disqualify him from receiving refugee status in Canada, authorities said.

The Immigration and Refugee Board is scheduled to decide by May 31 whether Arrow is affiliated with the Earth Liberation Front, a group the FBI considers a terrorist organization responsible for scores of attacks on property over the past dozen years.

☐ All Documents



IBM Content Analytics Version 2.2 (2011)

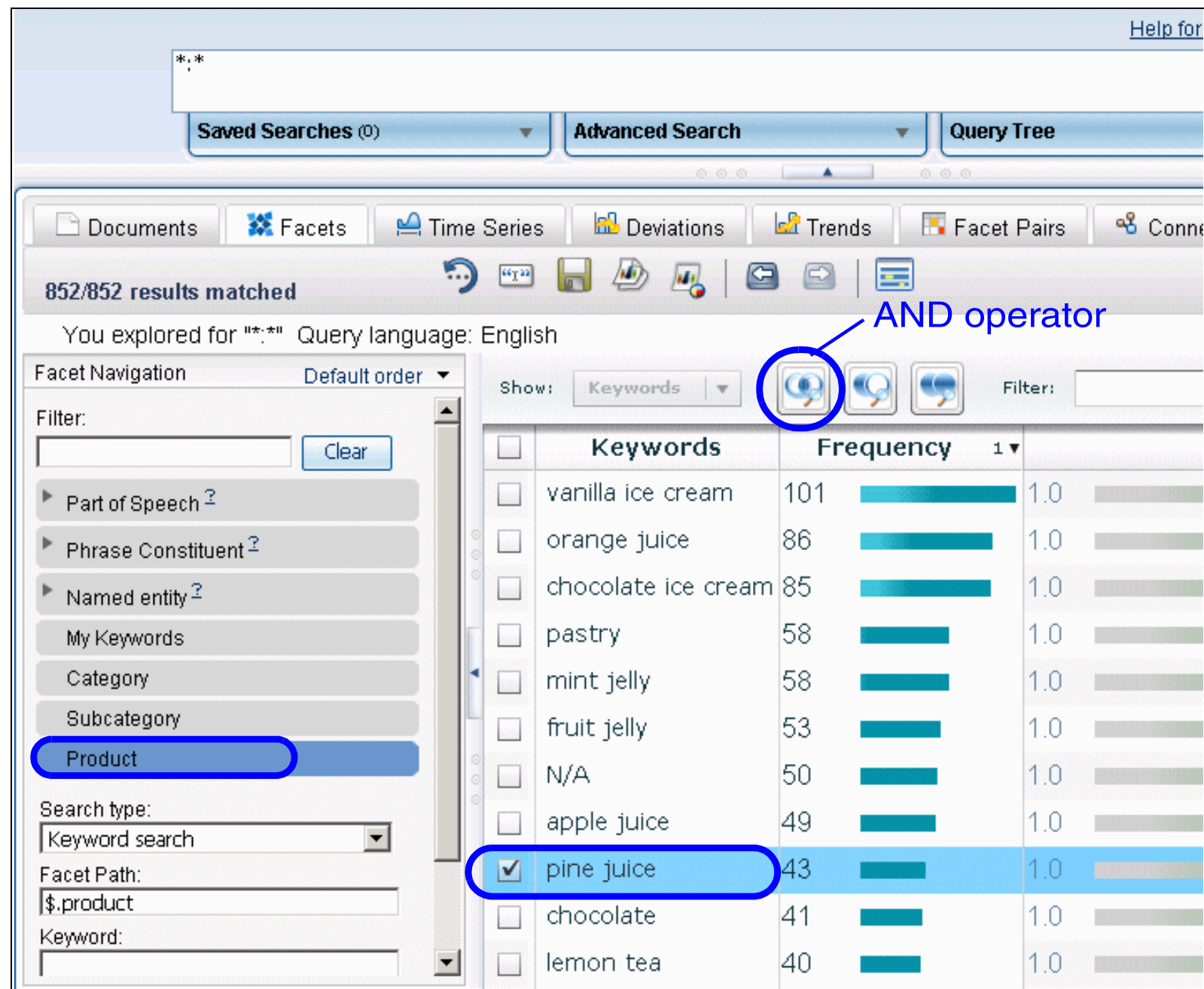


Figure 5-17 Selecting the Product facet and pine juice and clicking the AND operator

IBM Content Analytics Version 2.2 (2011)

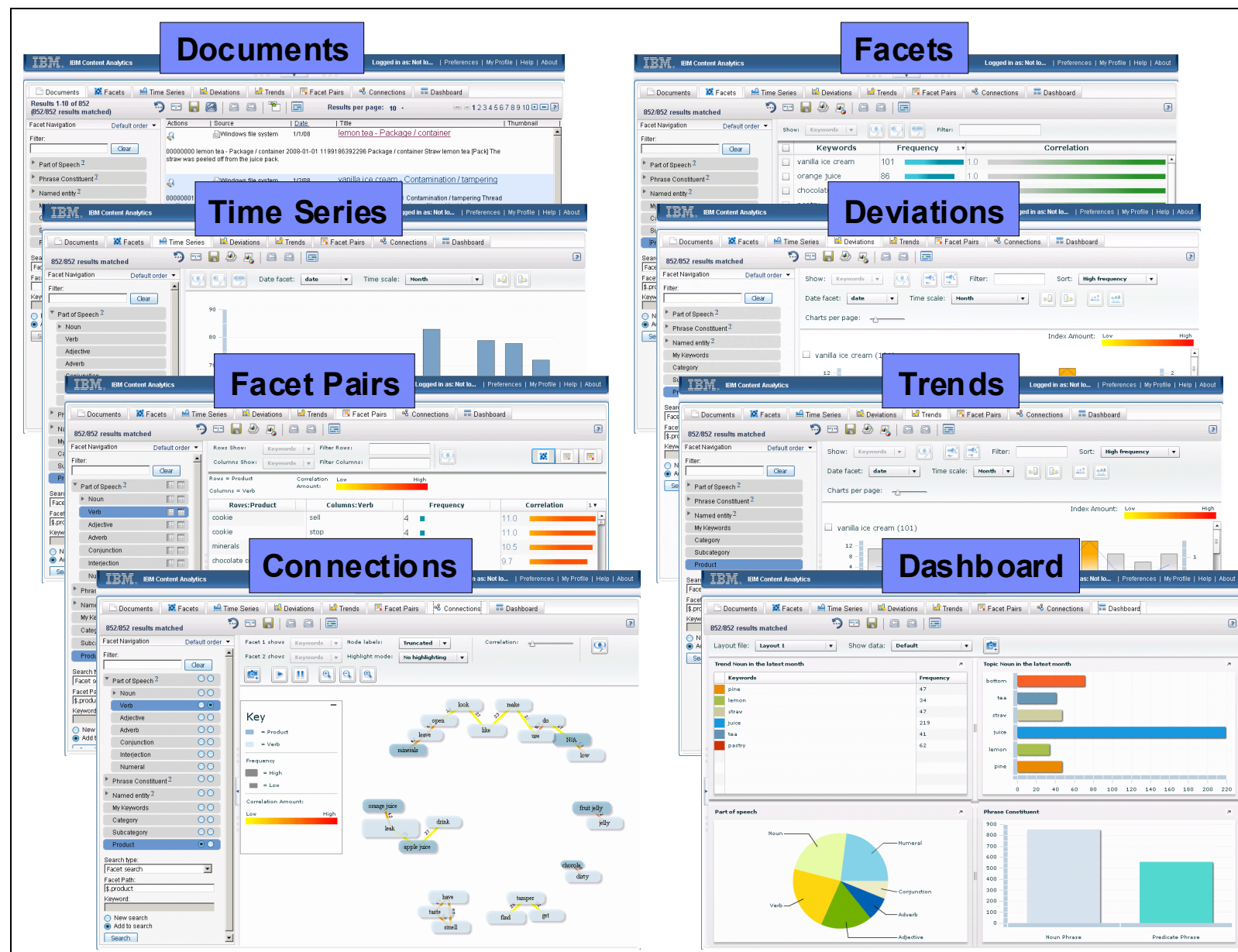


Figure 6-1 All views available in the text miner application

IBM Content Analytics Version 2.2 (2011)

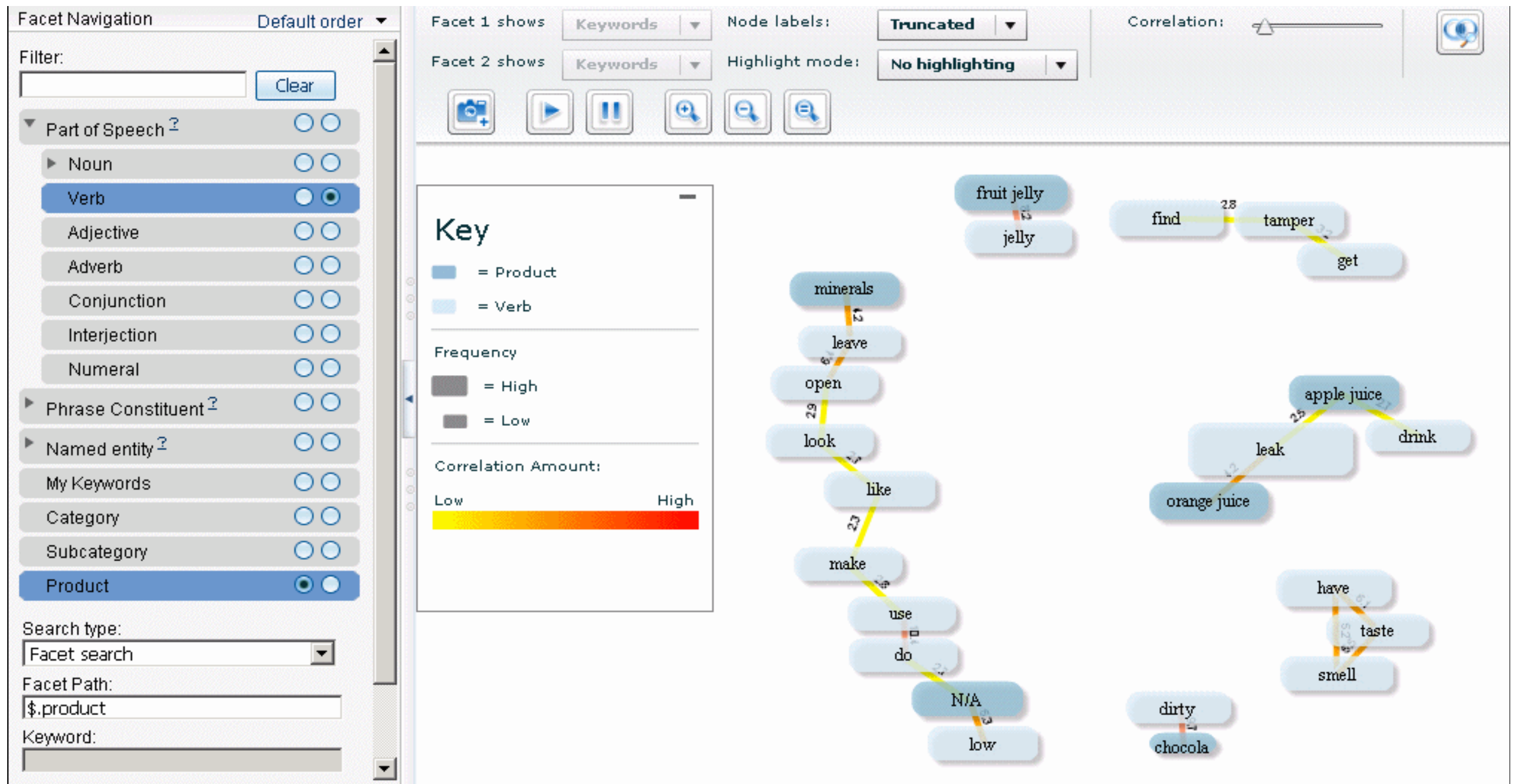


Figure 6-20 Connections view when selecting Product facet and Verb facet

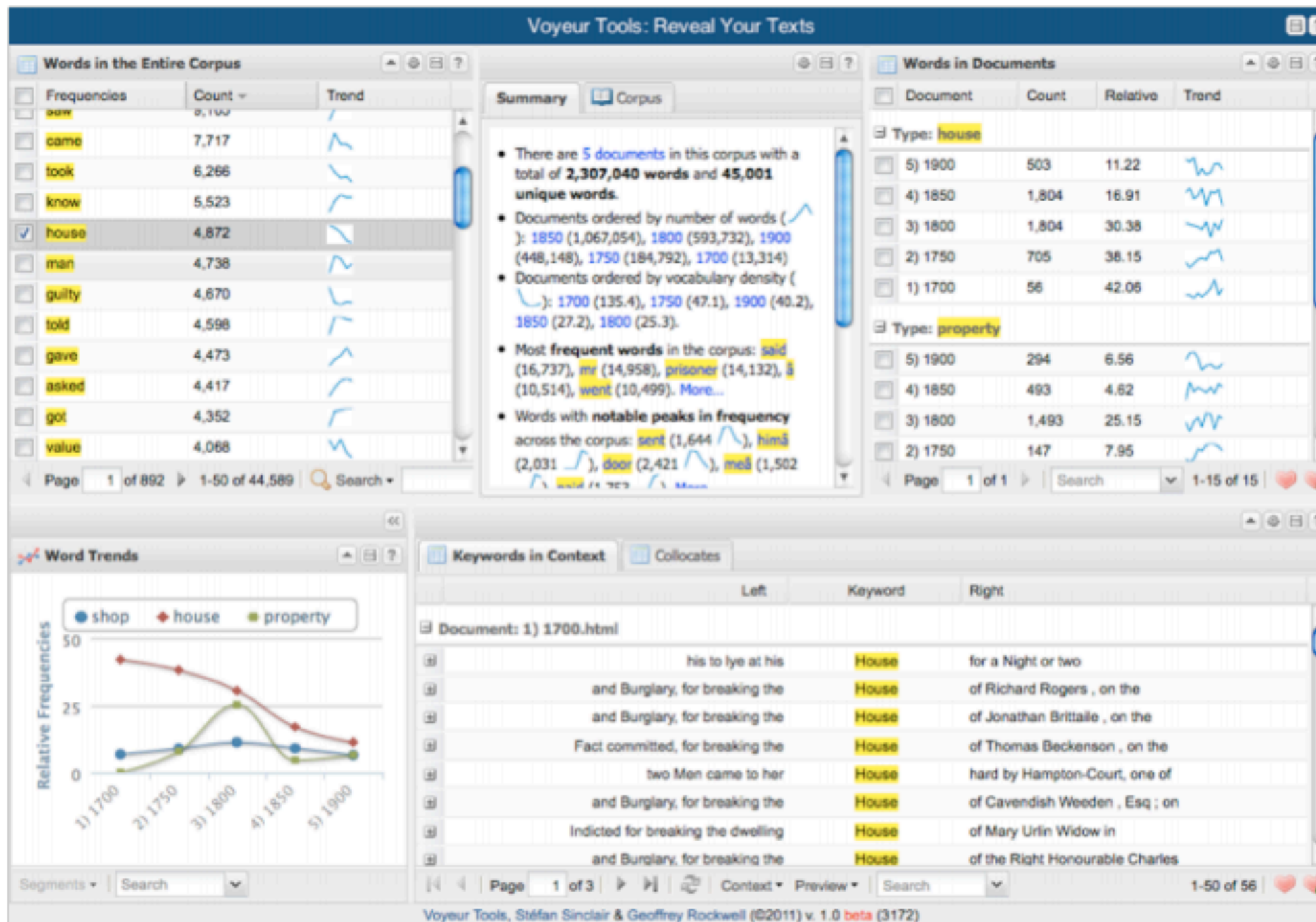
Mueller, WordHoard, Shakespeare, 2008

The screenshot shows a Mac OS-style window titled "Find Words". The menu bar includes "File", "Edit", "Sets", "Find", "Views", "Analysis", "Windows", and "Help". Below the menu bar, a text box instructs the user: "Search for words which satisfy all of the following criteria. Use the plus and minus buttons to add and remove criteria." The main area contains six criteria, each with a plus/minus toggle, a category dropdown, and a value field:

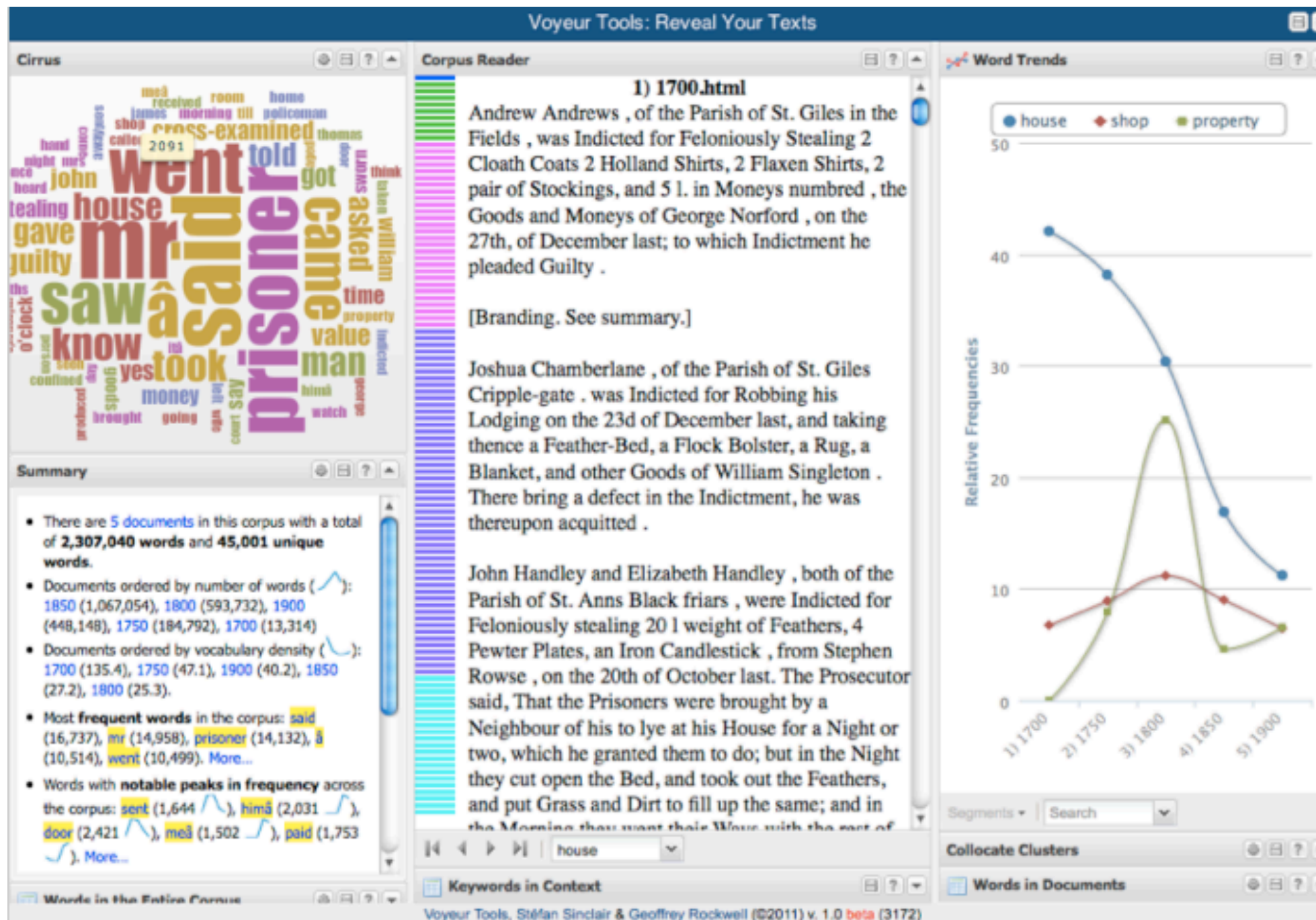
- Corpus is Shakespeare
- Lemma is love (v) [Complete button]
- Speaker gender is female
- Prose or verse is verse
- Work set is Shakespeare Comedies
- Publication year is between 1600 and 1605

At the bottom right, there are "Cancel" and "Find" buttons.

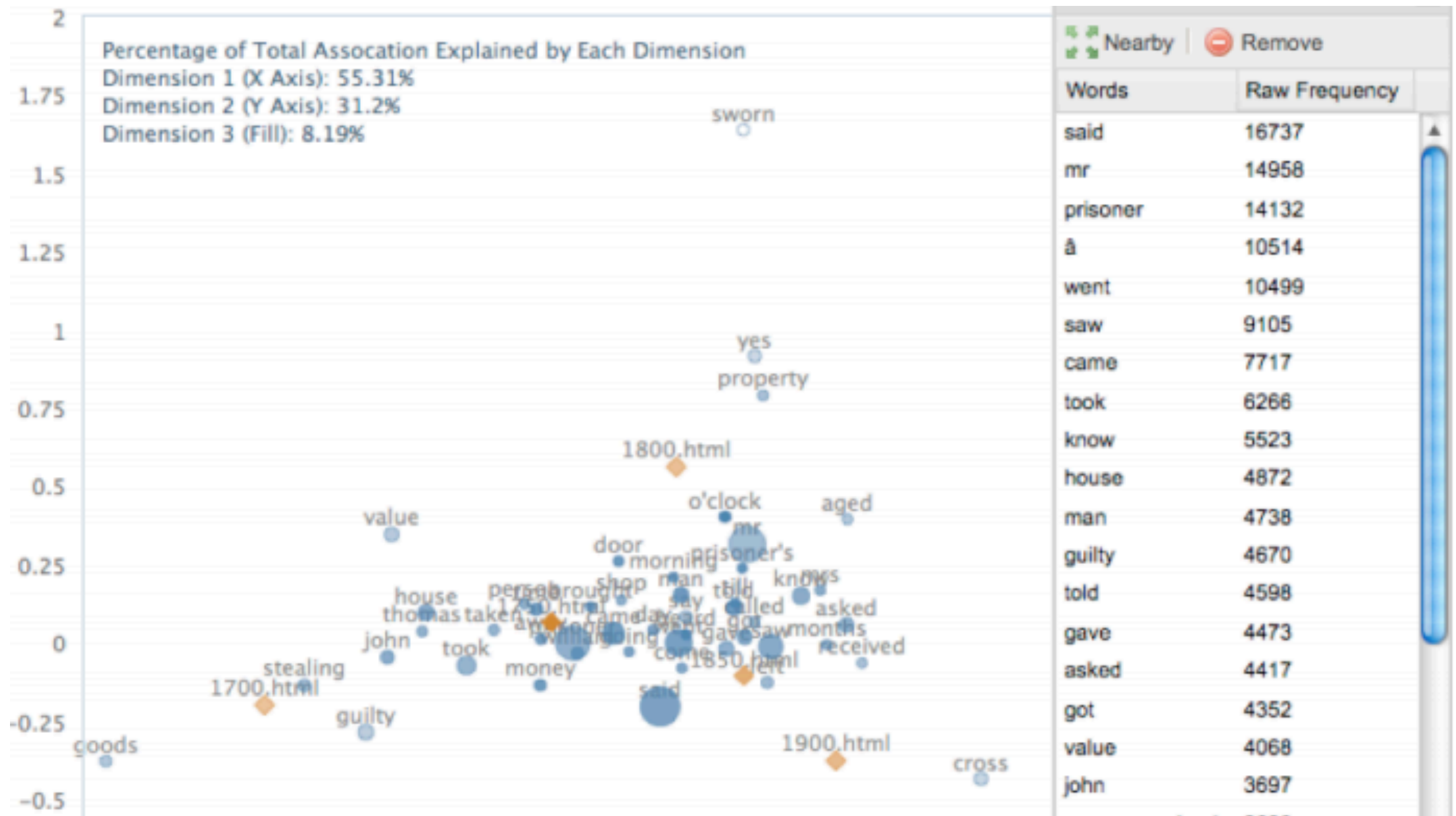
Sinclair & Rockwell, Voyant/Voyeur, TLLO 2009



Cohen et al., Voyeur, criminalintent.org 2011
(only works on one text collection)



Cohen et al., Voyeur, criminalintent.org 2011



Don et al., FeatureLens, CIKM 2007



Applied to Gertrude Stein's *The Making of Americans*

What WordSeer Supports that is Missing from Existing Tools

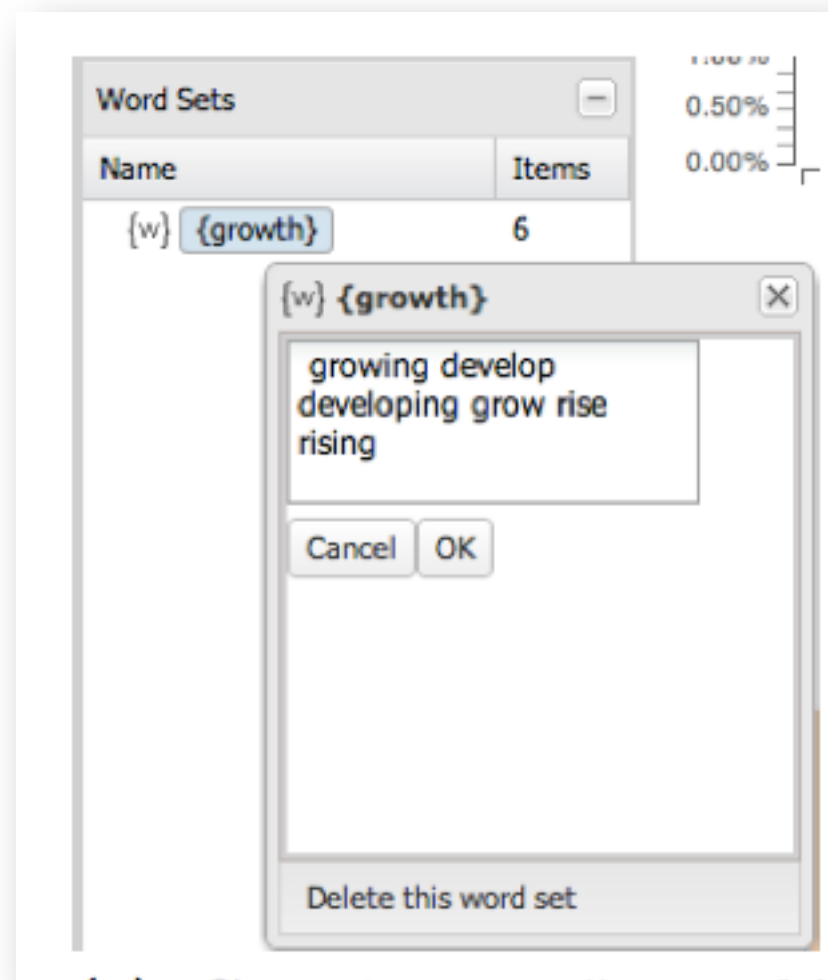
- Seeing analyzed text in context
- Fluidly building rich queries from context
- Easily create new classes of concepts from the text
- Flexible, context-based syntactic analysis
- Flexible, multi-way comparisons

Case Studies

1. Literary scholar studying American literature's reaction to China's rise
2. English Composition Educator analyzing student essays
3. My colleagues and my analysis of use of a MOOC forum

China Scholar: Concept Trend

- Interested in: How U.S. perceptions of China and Japan responded to China's rise over the last 30 years.
- Created a wordset with growth terms



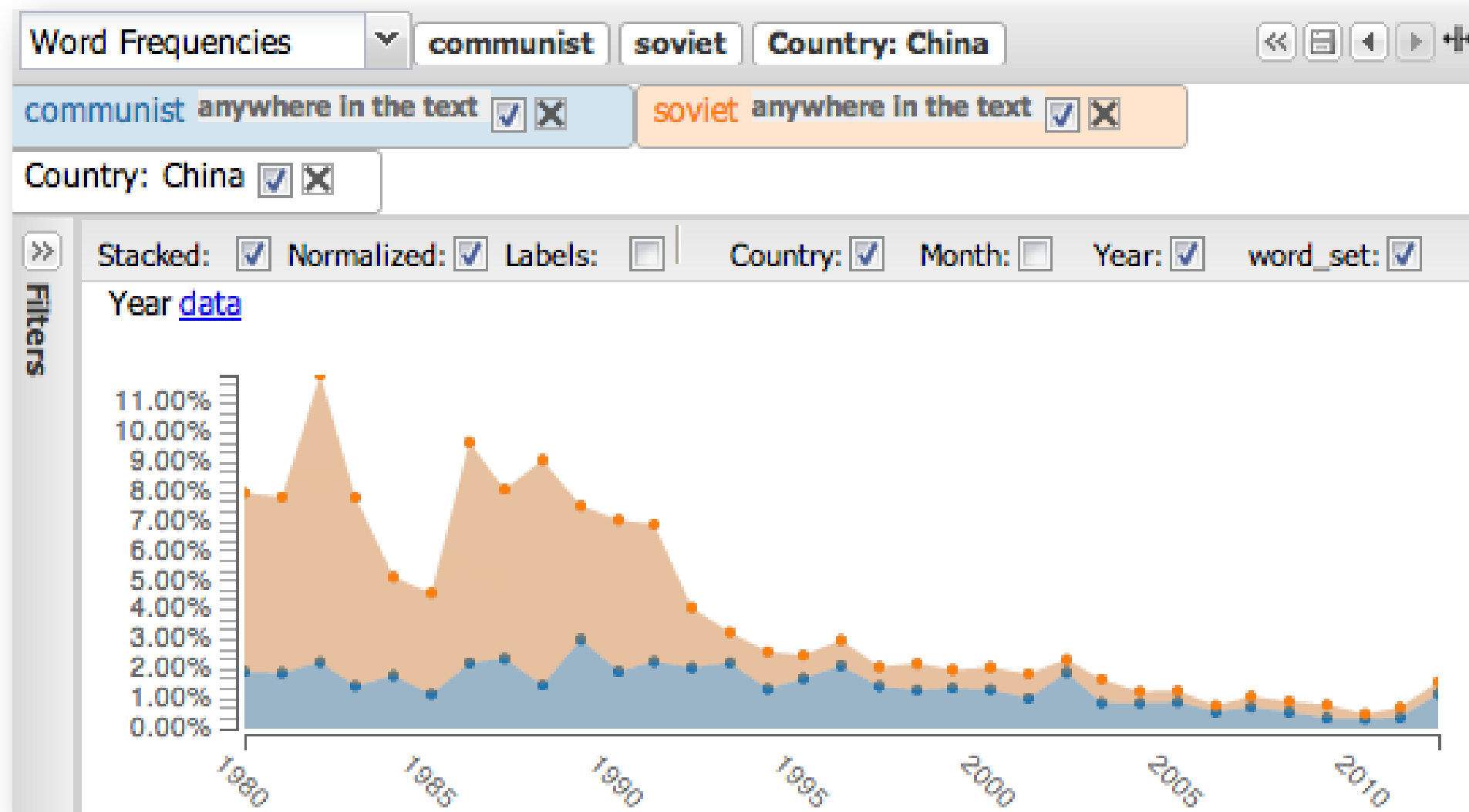
China Scholar: Concept Trend

- Confirmed his intuitions about China's rise by checking whether growth-related words became more frequent over time in editorials about China



China Scholar: Compare Trends

- Interested in: how China talked about over time: cold war rhetoric



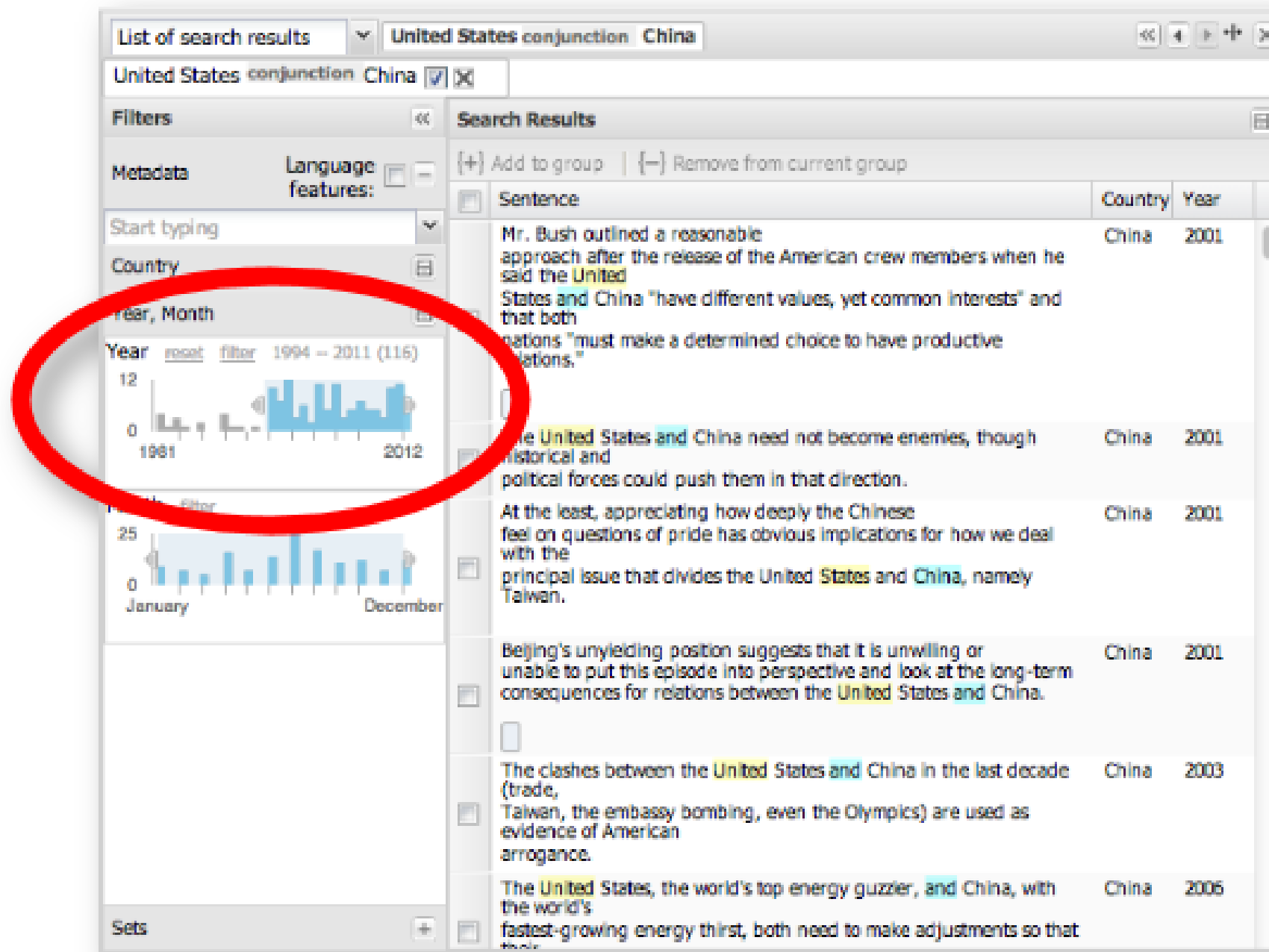
China Scholar: Mid-90s Onward: conjunction(China,U.S.)

- China joined the WTO in 2001; this is when China-US relations are thought to have become more inter-dependent.
- Idea: use a grammatical search to find interdependence: conjunction, e.g.:

“The United States, the world’s top energy guzzler, **and** China, with the world’s fastest-growing energy thirst . . .
” (April 2006). “

China Scholar: Mid-90s Onward: conjunction(China,U.S.)

- Conjunctions much more frequent after 1994



China Scholar: Explore the “Grammatical Neighborhood”

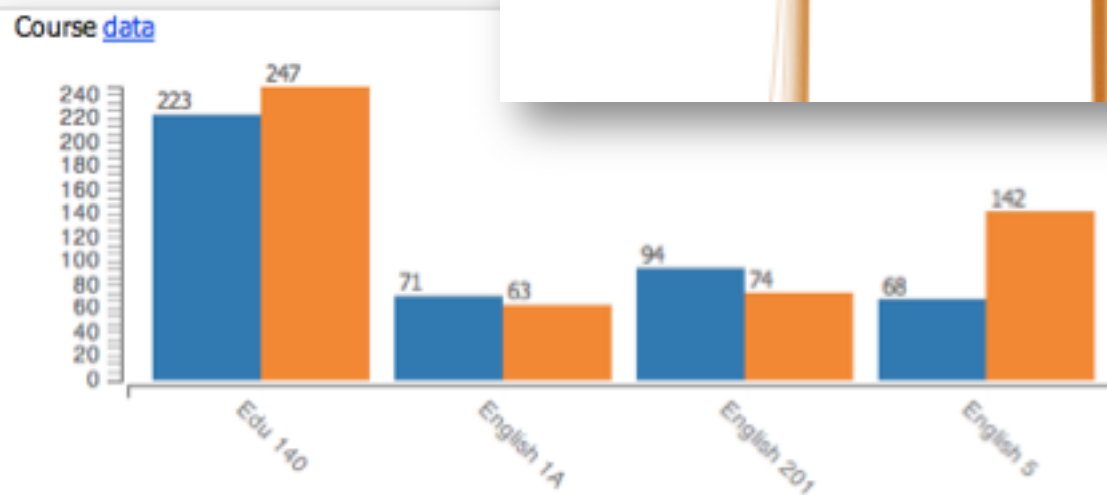
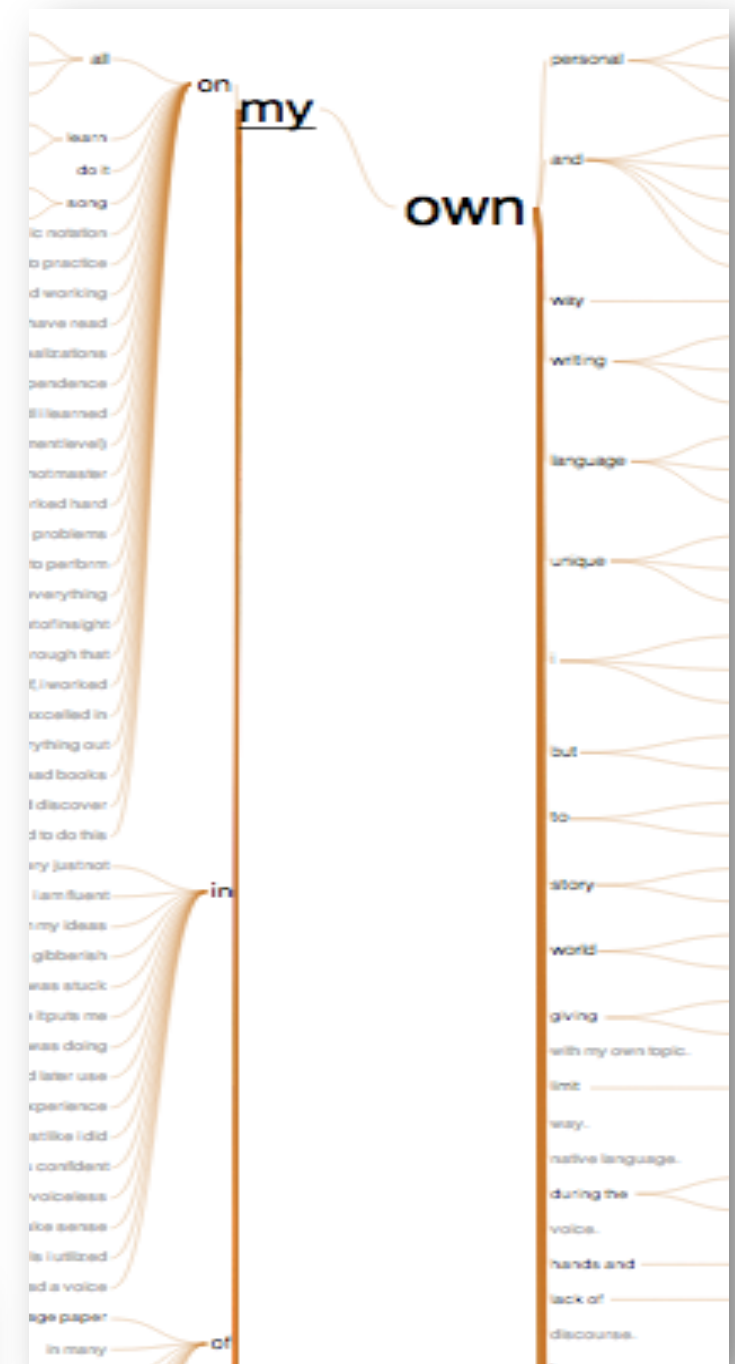
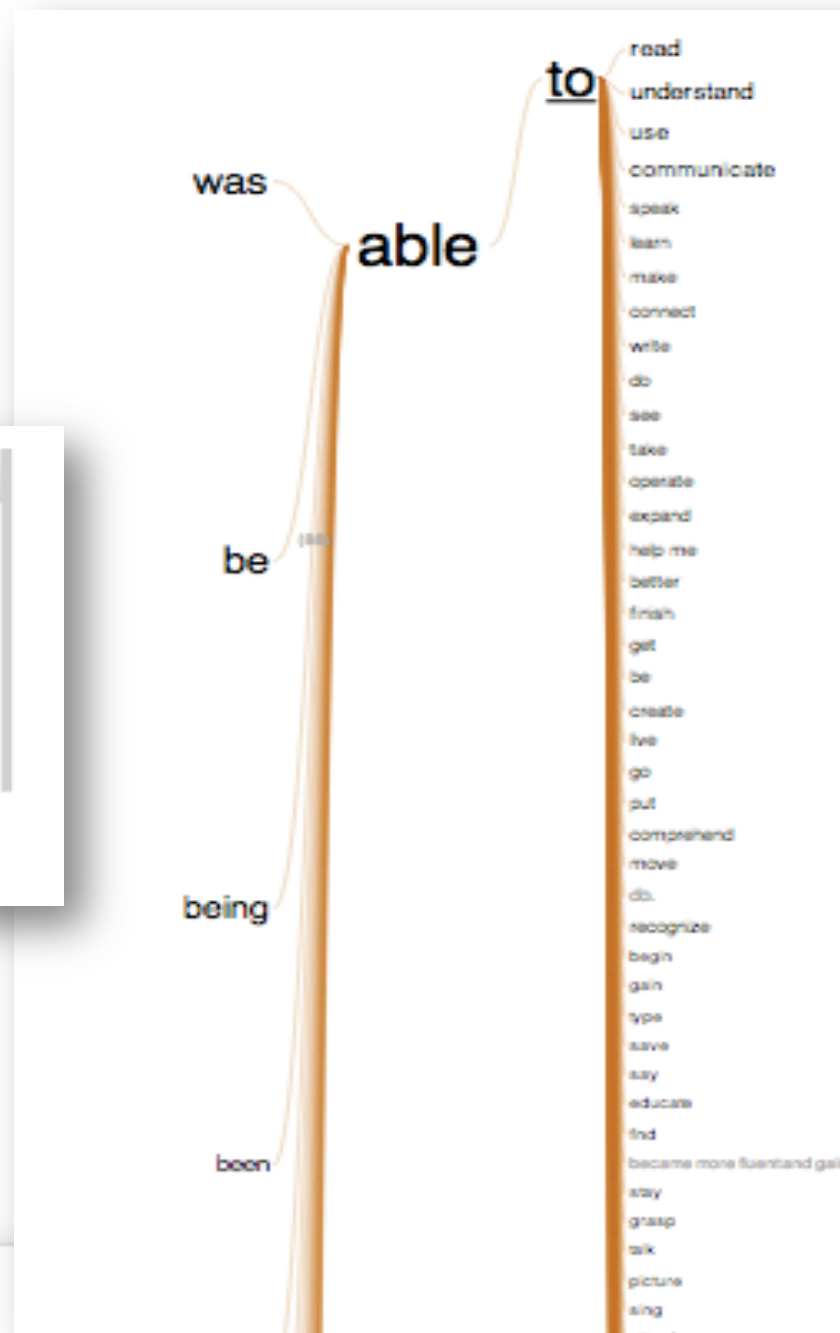
- Interested in: how China is talked about: anything interesting or unexpected?

| | | | | |
|-------------------------|-----------------|--|--|--|
| | | done to (China, ____) 3 | | noun compound modifier (, China) 25 |
| | | clausal complement (China, ____) 4 | | |
| comes | | temporal modifier (China, ____) 2 | | noun compound modifier (card, China) 21 |
| ship between the United | | infinitival modifier (China, ____) 3 | | noun compound modifier ((, China) 20 |
| | | open clausal complement (China, ____) 1 | | noun compound modifier (Corporation, China) 17 |
| Importers in just a | | done to (____, China) 6278 | | noun compound modifier (scholars, China) 10 |
| or China in | | possessive (____, China) 3226 | | noun compound modifier (Connection, China) 9 |
| ed | Sentence | conjunction (____, China) 1184 | | noun compound modifier (Daily, China) 8 |
| av | Edit word set | done by (____, China) 3562 | | noun compound modifier (experts, China) 7 |
| loc | Add to word set | noun compound modifier (____, China) 831 | | noun compound modifier (market, China) 7 |
| is | Related words | appositional modifier (____, China) 119 | | noun compound modifier (Seas, China) 6 |
| i | Search | dependent (____, China) 161 | | noun compound modifier (policies, China) 6 |
| can | | | | noun compound modifier (bill, China) 5 |

Literacy Essays

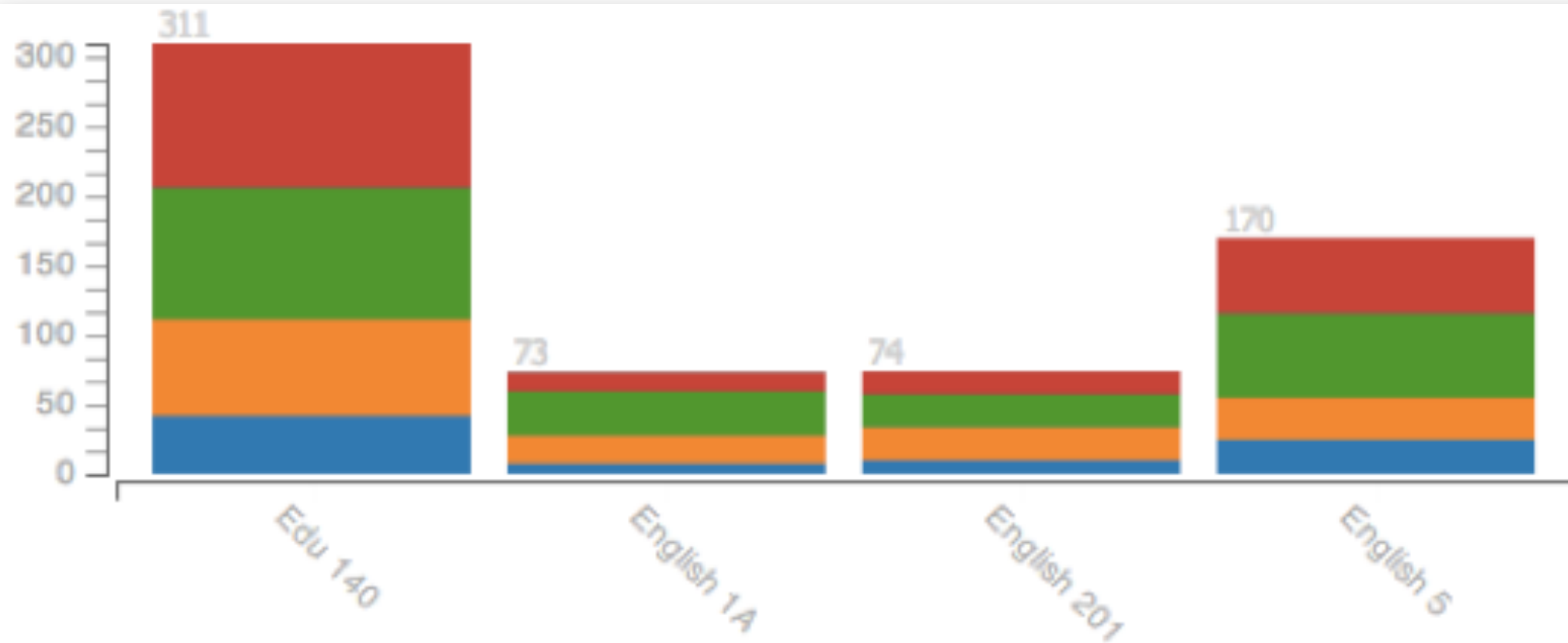
| Frequent Adjectives | group by stem: <input checked="" type="checkbox"/> <input type="checkbox"/> |
|---------------------|---|
| other | 462 |
| first | 382 |
| new | 371 |
| able | 316 |
| english | 286 |
| own | 273 |

Figure 2 – Most Frequent Adjectives



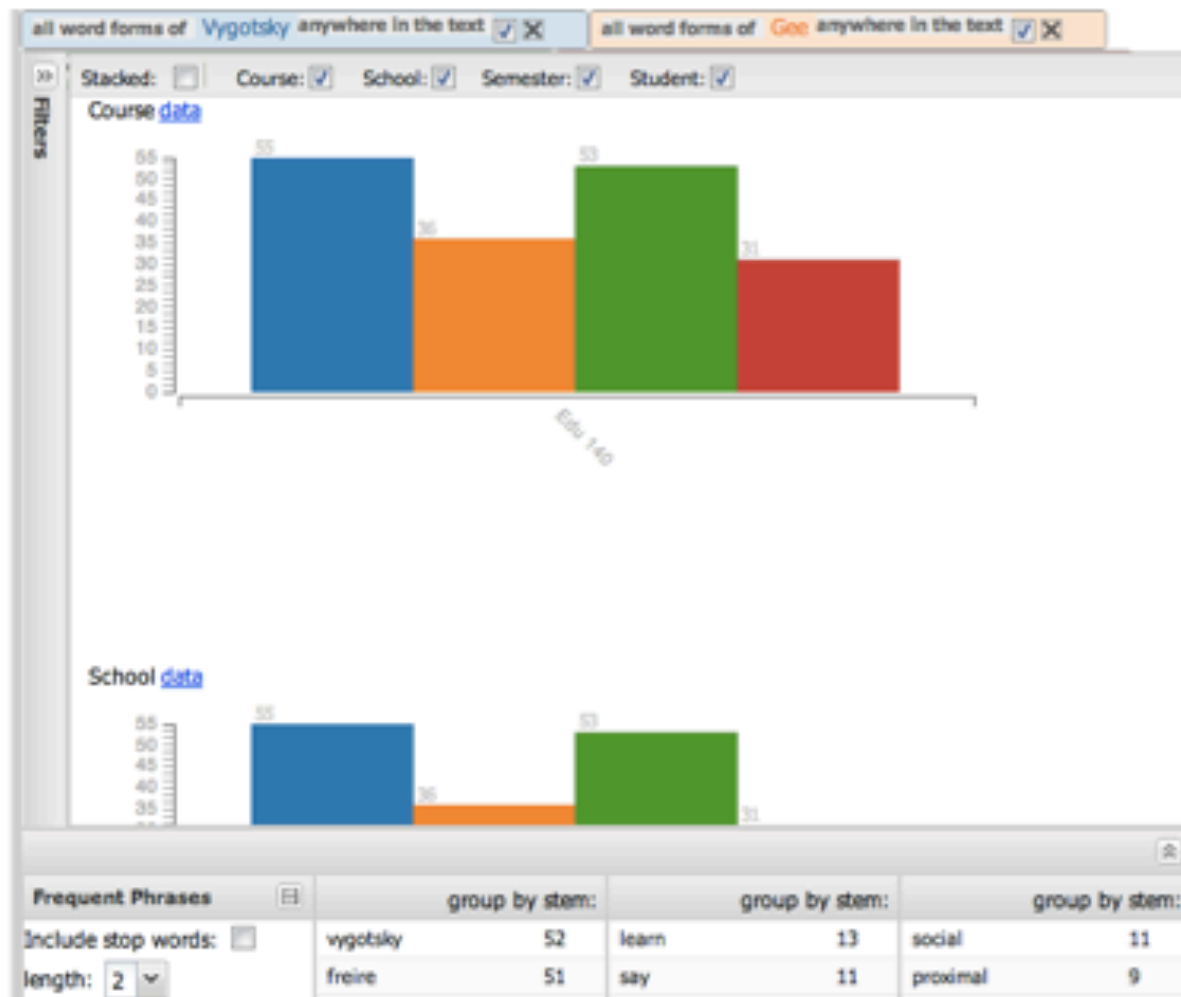
Own Able

Literacy Essays



Although **Though** **While** **However**

Literacy Essays



all word forms of **Rodriguez** anywhere in the text ☒ ☐

Search Results ☐ Add to group ☐ Remove from current group

| | Semester | School | Student | Course |
|---|----------|--------|-----------|---------|
| <input type="checkbox"/> All I know is, if Rodriguez chose of his own accord to become a "scholarship boy," then I grew up in an environment where a new breed of "scholarship boy" was being mass produced by the hundreds. | Fall12 | UCB | Chang | Edu 140 |
| <input type="checkbox"/> Rodriguez , Richard. | Fall12 | UCB | Chang | Edu 140 |
| <input type="checkbox"/> In <i>Hunger for memory: The education of Richard Rodriguez</i> , An autobiography. | Fall12 | UCB | Chang | Edu 140 |
| <input type="checkbox"/> I was "annoyed that I could not get parental help with my homework" and I was also jealous of my peers who had English speaking parents (Rodriguez , 44). | Fall12 | UCB | Gallegos | Edu 140 |
| <input type="checkbox"/> I felt that I had to "cut myself off mentally" so that I could continue surpassing everyone else (Rodriguez , 47). | Fall12 | UCB | Gallegos | Edu 140 |
| <input type="checkbox"/> The kind of allegiance the young student might have given his mother or father only days earlier, he transfers to the teacher, the new figure of authority" (Rodriguez , 49). | Fall12 | UCB | Gallegos | Edu 140 |
| <input type="checkbox"/> (Rodriguez , 58) I was ashamed of my language during my teenage years. | Fall12 | UCB | Gallegos | Edu 140 |
| <input type="checkbox"/> Similar to the way Richard Rodriguez explained in his autobiography, <i>Hunger for Memory</i> . | Fall12 | UCB | Kriegsman | Edu 140 |

Frequent Phrases ☐ Include stop words: ☐ length: 2

| group by stem: | group by stem: | group by stem: |
|----------------|----------------|----------------|
| rodriguez 30 | continue 3 | own 2 |
| richard 14 | say 3 | academic 2 |

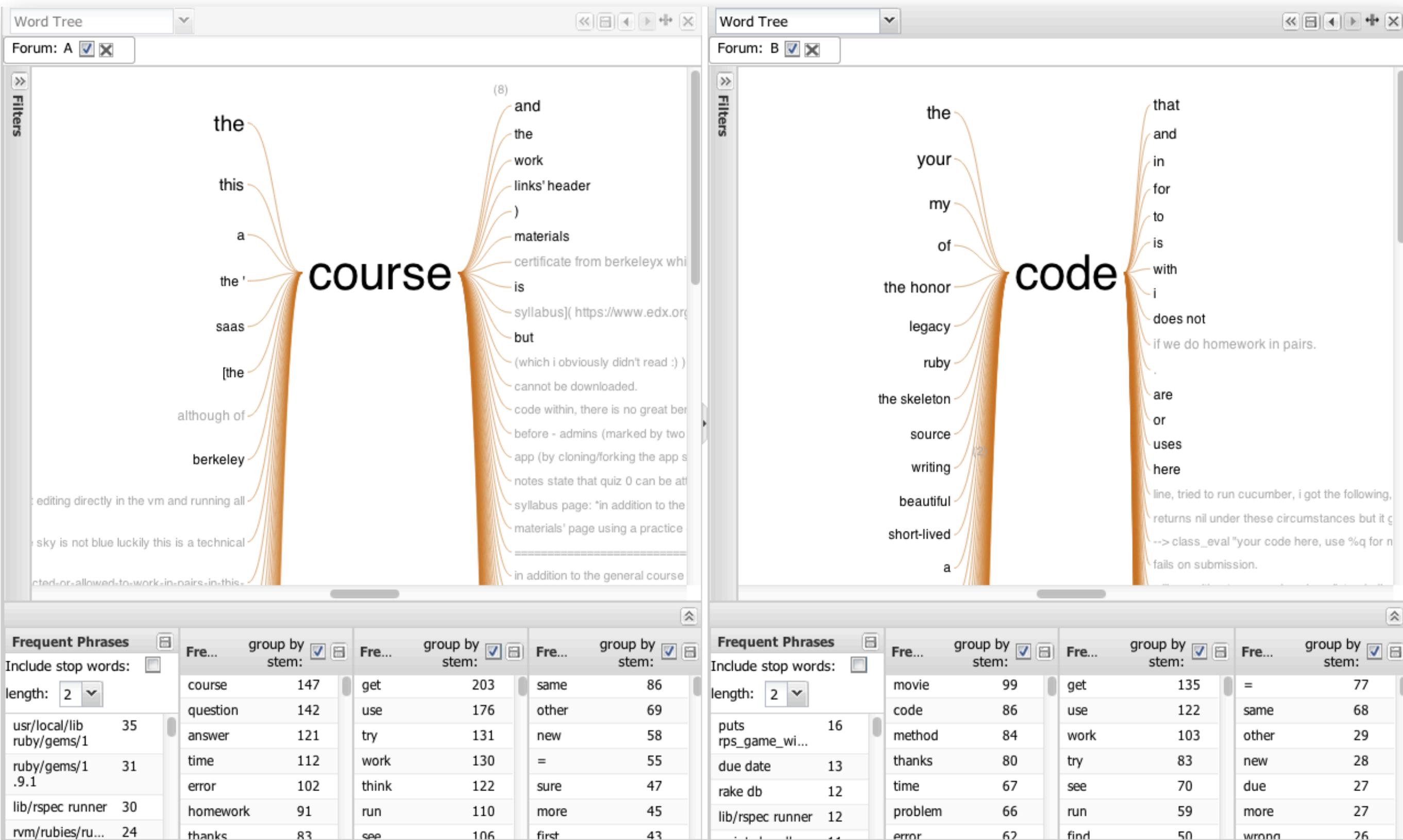
Vygotsky

Gee

Freire

Rodriguez

Forum Analysis



Forum Analysis

List of search results

all word forms of { } anywhere in the text

Filters

Metadata

Language features:

Sets

Document Sets

Sentence Sets

Word Sets

Name

Items

{w} {appeal for help} 7

Search Results

{+} Add to group | {-} Remove from current group

| | Sentence | Author ID | Tags | Type | TotalVotes | Forum | Initiator | Num Respons | Num Answers | Num Comme |
|--------------------------|---|-----------|-------|-------------|------------|-------|-----------|-------------|-------------|-----------|
| <input type="checkbox"/> | o awseome - someone else liked my README - ty Jay | 0100 | quiz0 | answerco... | 1 | A | 0203 | 27 | 7 | 20 |
| <input type="checkbox"/> | i havent looked into this but i will venture to say that berkeleyX is not part us UCberkeley officially - as edX is a non-profit i assume that berkeleyX is as well and o/c UC is not - in any case the certificate is unofficial - it is issued by the professors personally and not by UC anyone ? | 0100 | quiz0 | answerco... | 1 | A | 0203 | 27 | 7 | 20 |
| <input type="checkbox"/> | I don't think anyone is "under the illusion" that they are receiving college credit for this course work nor do I think many of us care. | 0203 | quiz0 | answerco... | 1 | A | 0203 | 27 | 7 | 20 |
| <input type="checkbox"/> | This is a new medium for education so we'll all get smarter with experience (including how to post "general" questions ;-). How's this one, "Has anyone found a good alternative to Wine emulator for getting Kindles to run on Linux?" | 0660 | quiz0 | answerco... | 1 | A | 0203 | 27 | 7 | 20 |
| <input type="checkbox"/> | IMHO, such information is more appropriately placed in the help FAQ for the edX site. | 0203 | quiz0 | answerco... | 1 | | | | | |
| <input type="checkbox"/> | The information that Mr-Jonze and Kwalshy have provided is nowhere to be found in edx.org help section, nor is it easily found anywhere else on the "minimalistic" site. | 0203 | quiz0 | answerco... | 1 | | | | | |
| <input type="checkbox"/> | @Mr-Jonze Please read post with information from my direct e-mail correspondence with edX. | 0203 | quiz0 | answerco... | 1 | | | | | |
| <input type="checkbox"/> | as a rule we should try not to reveal the answers to assionments but as this particular quiz has unlimited | 0100 | quiz0 | answer | 1 | | | | | |

Refresh New Subset Delete

Frequent Phrases

Include stop words:

length: 2

| | |
|----------------------|----|
| Hope helps | 10 |
| puts rps_game_winner | 9 |
| best answer | 8 |
| Thanks help | 7 |
| new forum | 6 |

Frequent Nouns

group by stem:

| | |
|----------|----|
| anyone | 66 |
| help | 57 |
| someone | 39 |
| answer | 38 |
| question | 28 |
| time | 20 |

Frequent Verbs

group by stem:

| | |
|------------|----|
| help | 7 |
| please | 6 |
| get | 32 |
| appreciate | 24 |
| ask | 24 |
| post | 22 |

Frequent Adjectives

group by stem:

| | |
|------|----|
| best | 10 |
| more | 9 |
| such | 8 |
| own | 8 |

{w} {appeal for help}

someone

anyone

help

please

anybody

appreciate

somebody

Delete this word set

What's Missing?

- Integration of thesauri / taxonomies
- More sophisticated syntactic pattern matching
- Tools to learn from examples
- Tools to check “the rest” of the hypothesis

Conclusions

- “Big data” is often not explored in a sophisticated way
 - Instead, perhaps we should focus on medium-sized, motivated slices to gain more insight.
- Interfaces and analysis should support the “middle game”
 - Midway between close read and distant statistics.
 - Help with hypothesis formulation, verification, and refinement.

A Romantic-style landscape painting depicting a serene scene at sunset. A large, gnarled tree with dense foliage frames the right side of the image. In the center, a calm body of water reflects the bright, golden light of the setting sun, which is partially obscured by soft, hazy clouds. A small boat with a single figure is visible on the water in the lower left. In the distance, a small town or village is nestled at the foot of a mountain range. The overall atmosphere is peaceful and majestic, with a warm color palette dominated by yellows, oranges, and greens.

Thank you!

Exploratory Text Analysis and The Middle Distance

Marti A. Hearst
U.C Berkeley

Joint Work with Aditi Muralidharan

Collaborators: Bryan Wagner, Chris Fan, Rex Ganding

Sponsored by NEH HK50011