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

Background: Distant Read (Text Mining)

Middle Distance: Sensemaking

Foreground: The Close Read

Foreground: The Close Read

Definition: “Close Read”

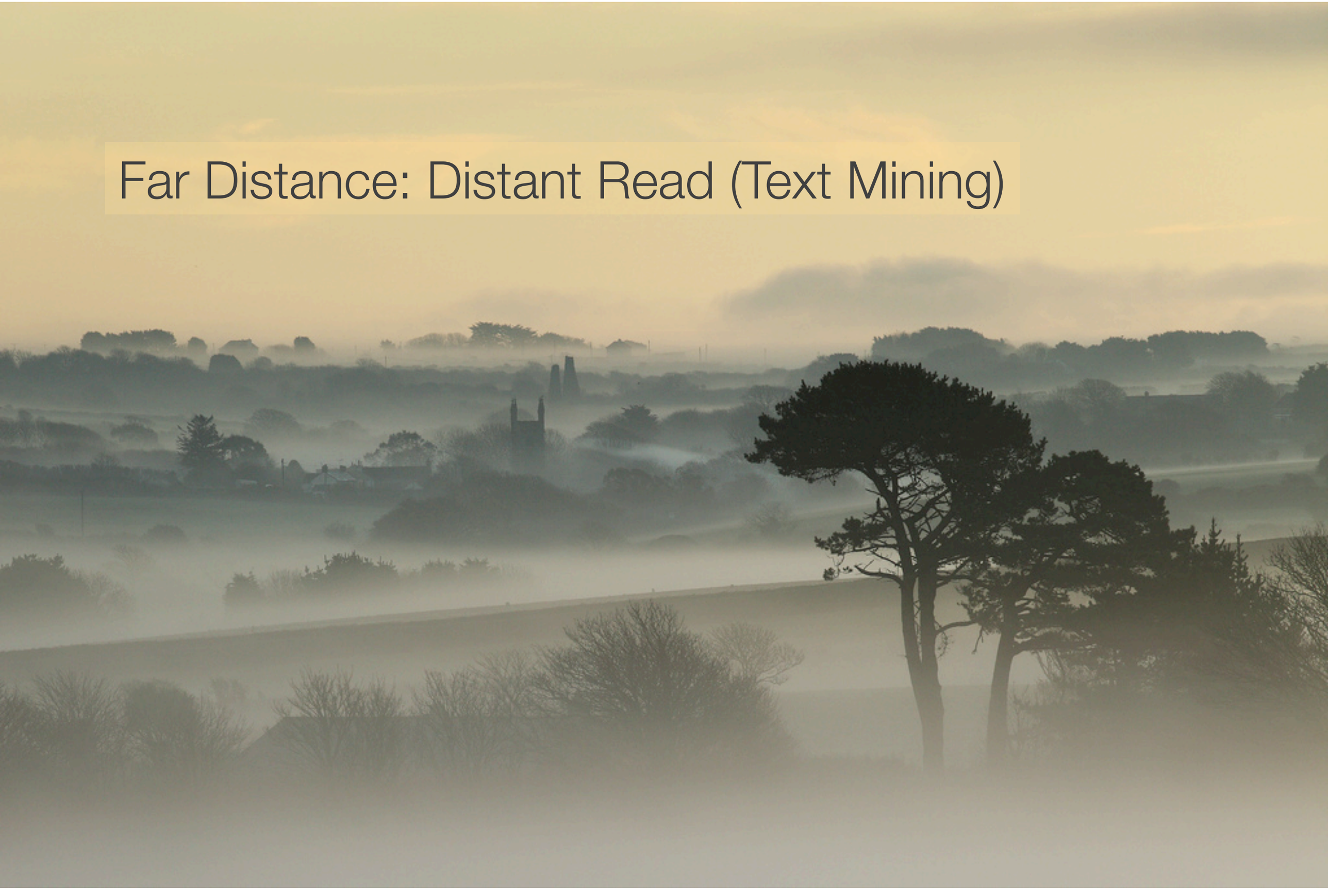
“Close reading describes, in literary criticism, the careful, sustained interpretation of a brief passage of text. Such a reading places great emphasis on the particular over the general, paying close attention to individual words, syntax, and the order in which sentences and ideas unfold as they are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

-English Wikipedia, 6/4/2012

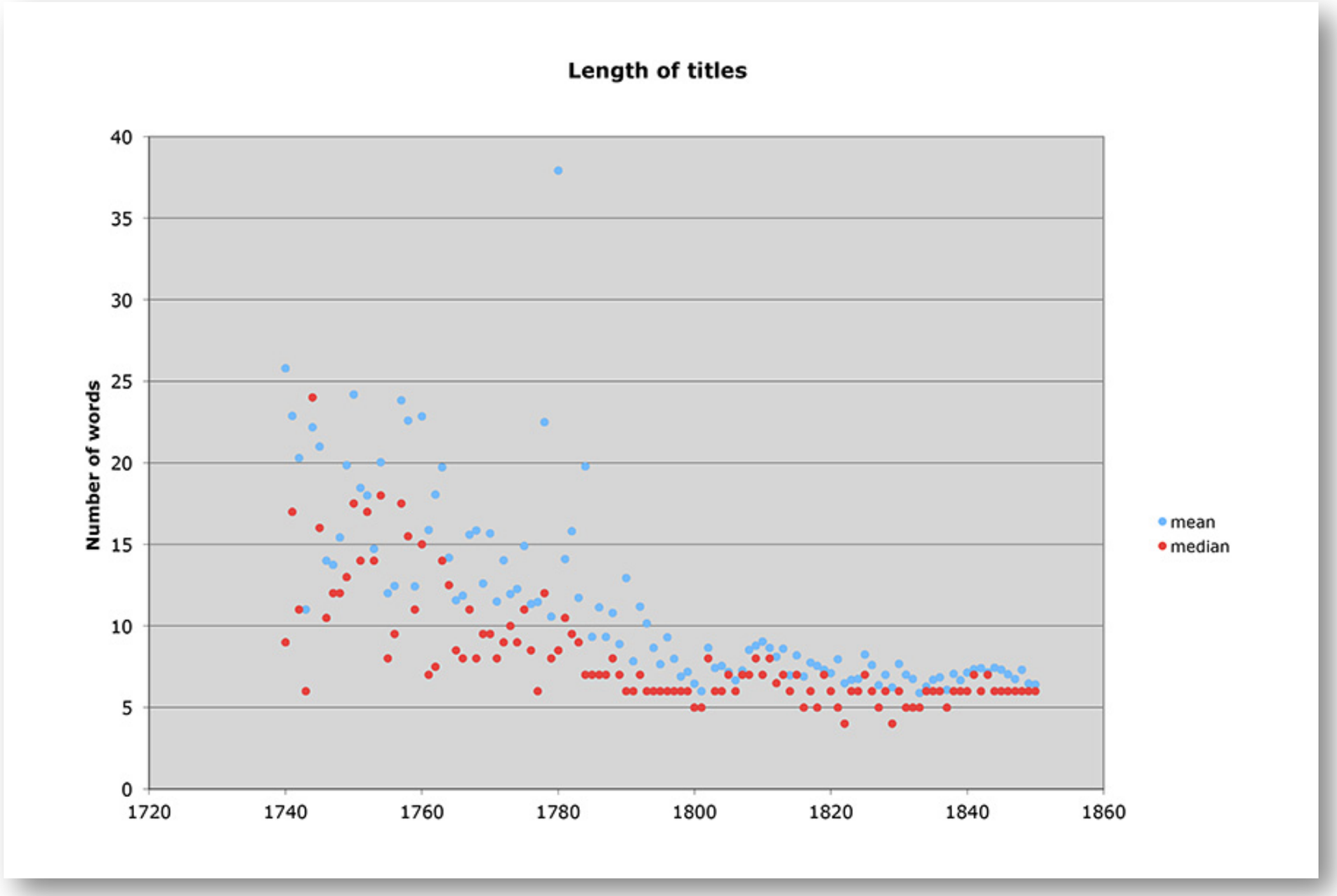
“Power and Passion in Shakespeare’s Pronouns
Interrogating ‘you’ and ‘thou’”
Penelope Freedman, 2007, MPG Books, 280 pp.

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

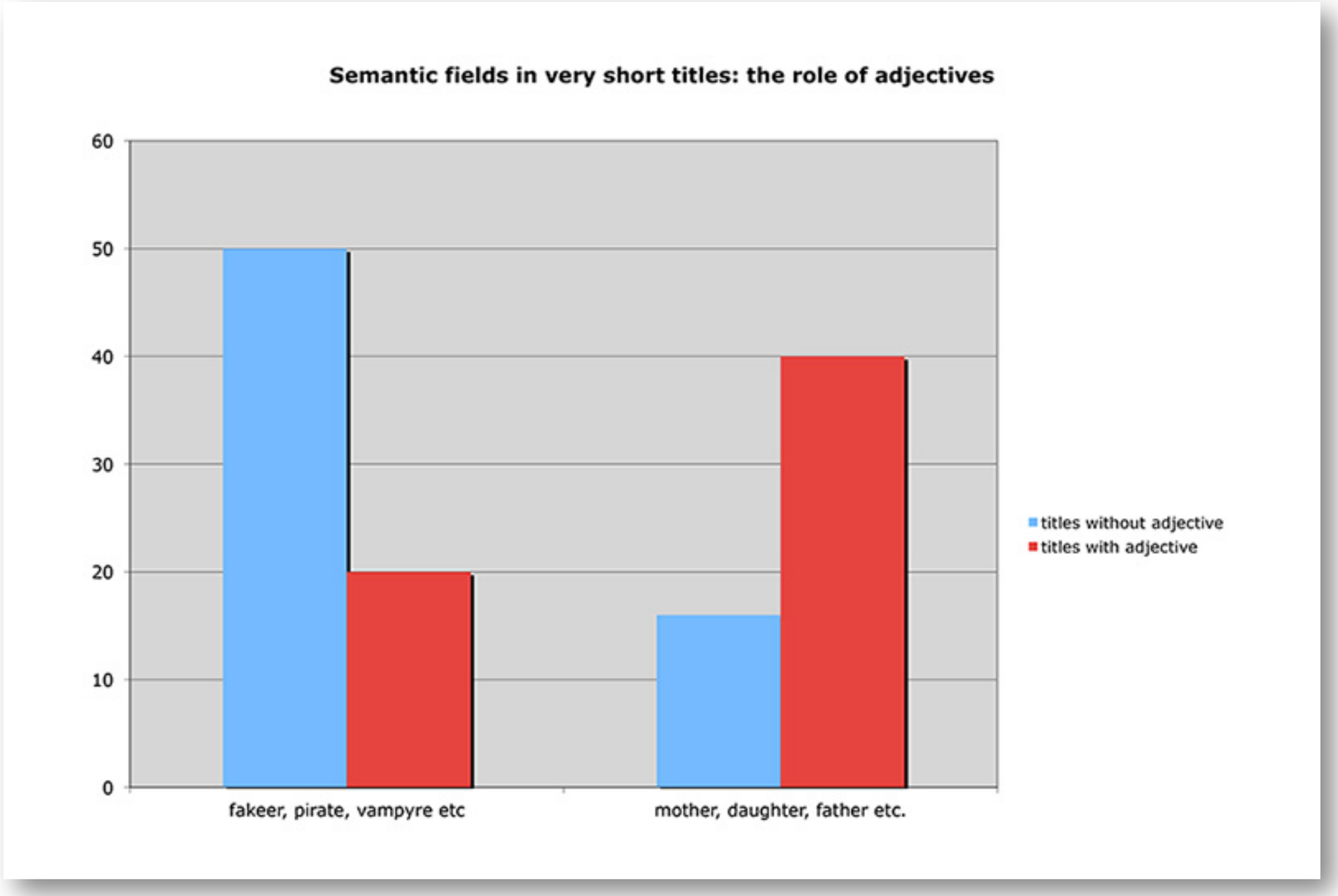
Far Distance: Distant Read (Text Mining)



Text Mining in Moretti, "*Distant Reading*", 2013
Shrinking Book Title Lengths Across the Years

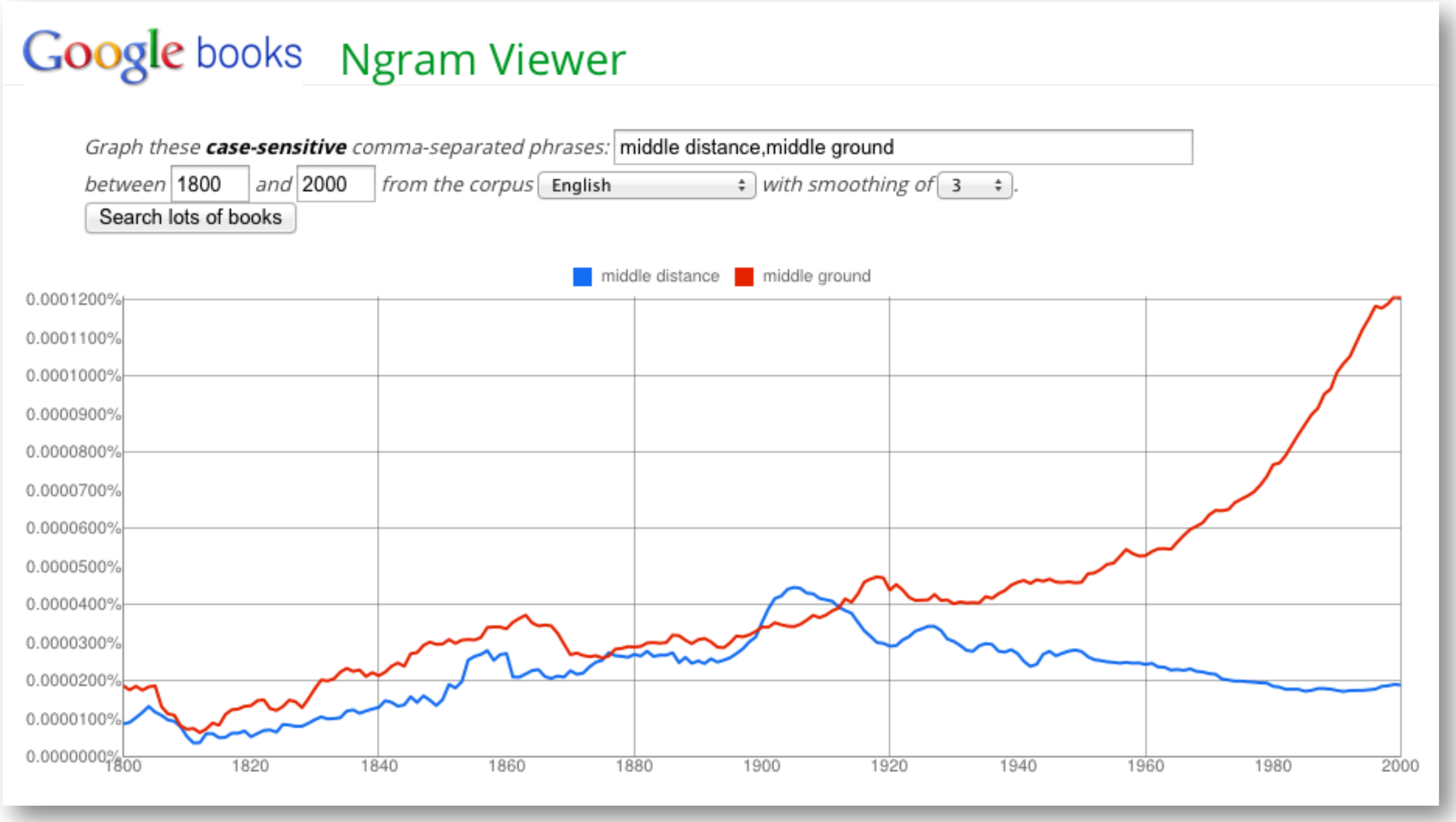


Text Mining in Moretti, "*Distant Reading*", 2013
Comparing Uses of Adjectives in Book Titles



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.

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

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.

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

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.

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

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~~.

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

Text Mining:
Statistical Text Analysis since at least 1990

Table 5. What Can You Drink?

Verb	Object	Mutual Info	Joint Freq
drink/V	*martinis/O*	12.6	3
drink/V	*cup_water/O*	11.6	3
drink/V	*champagne/O*	10.9	3
drink/V	*beverage/O*	10.8	8
drink/V	*cup_coffee/O*	10.6	2
drink/V	*cognac/O*	10.6	2
drink/V	*beer/O*	9.9	29
drink/V	*cup/O*	9.7	6
drink/V	*coffee/O*	9.7	12
drink/V	*toast/O*	9.6	4
drink/V	*alcohol/O*	9.4	20
drink/V	*wine/O*	9.3	10
drink/V	*fluid/O*	9.0	5
drink/V	*liquor/O*	8.9	4
drink/V	*tea/O*	8.9	5
drink/V	*milk/O*	8.7	8
drink/V	*juice/O*	8.3	4
drink/V	*water/O*	7.2	43
drink/V	*quantity/O*	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 $\{, NP_2 \ldots , (and \mid or)\ NP_i\}$ $\quad i \geq 1$

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("*Gelidium*","*red algae*").

Automatic acquisition of hyponyms from large text corpora
MA Hearst - Proceedings of the 14th conference on Computational Linguistics, 1992

Lexico-Syntactic Patterns

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

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

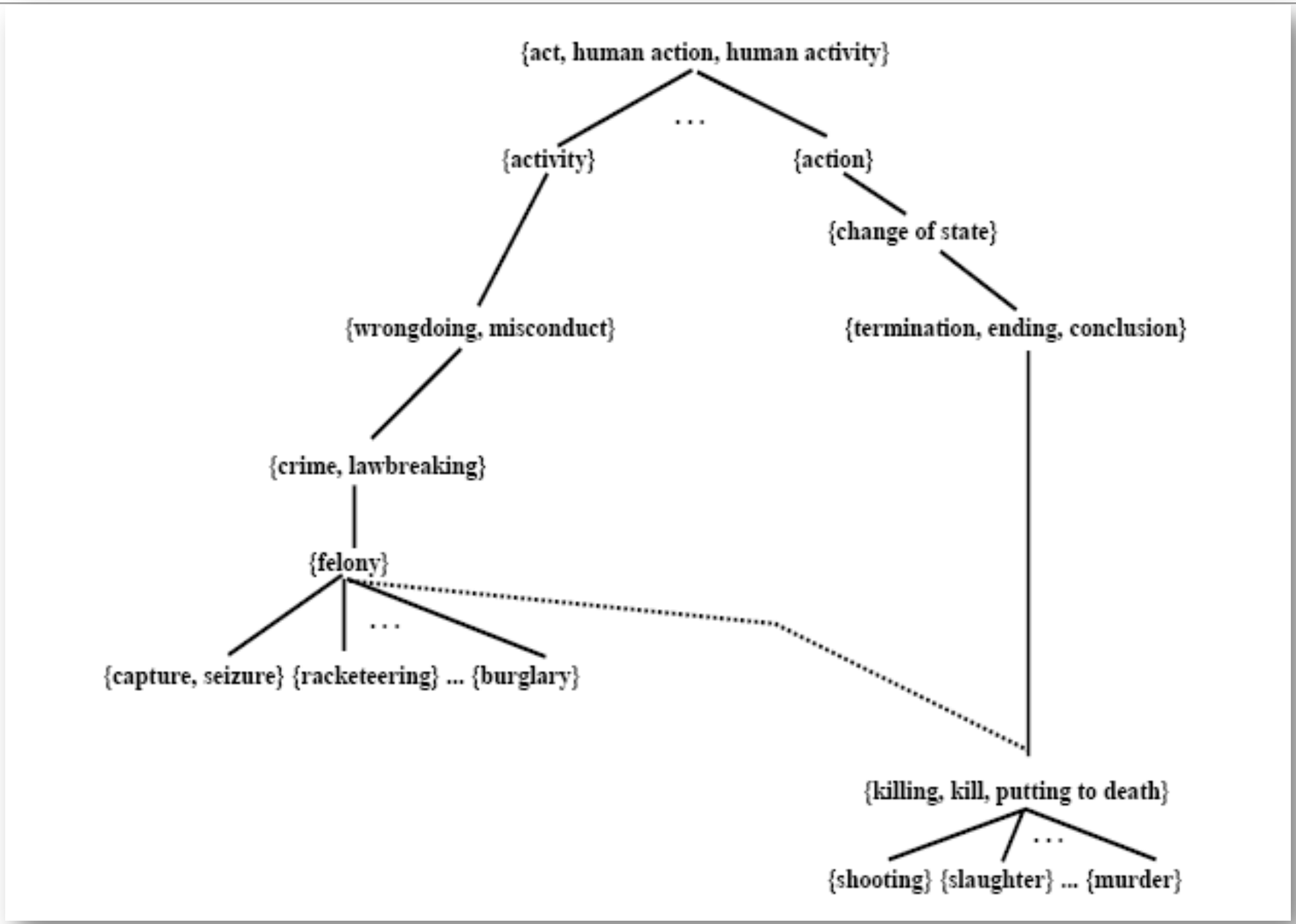
$\Longrightarrow$ *hyponym("author", "Herrick"),*
hyponym("author", "Goldsmith"),
hyponym("author", "Shakespeare")

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

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

$\Longrightarrow$ *hyponym("bruise", "injury"),*
hyponym("broken bone", "injury")

Text Mining:
Adding a New Relation to a Lexicon



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.

Wikipedia

Ngram Viewer on the Chess Metaphor

Google books Ngram Viewer

Graph these case-sensitive comma-separated phrases: chess opening,chess endgame,chess middlegame

between 1900 and 2000 from the corpus English with smoothing of 3 .

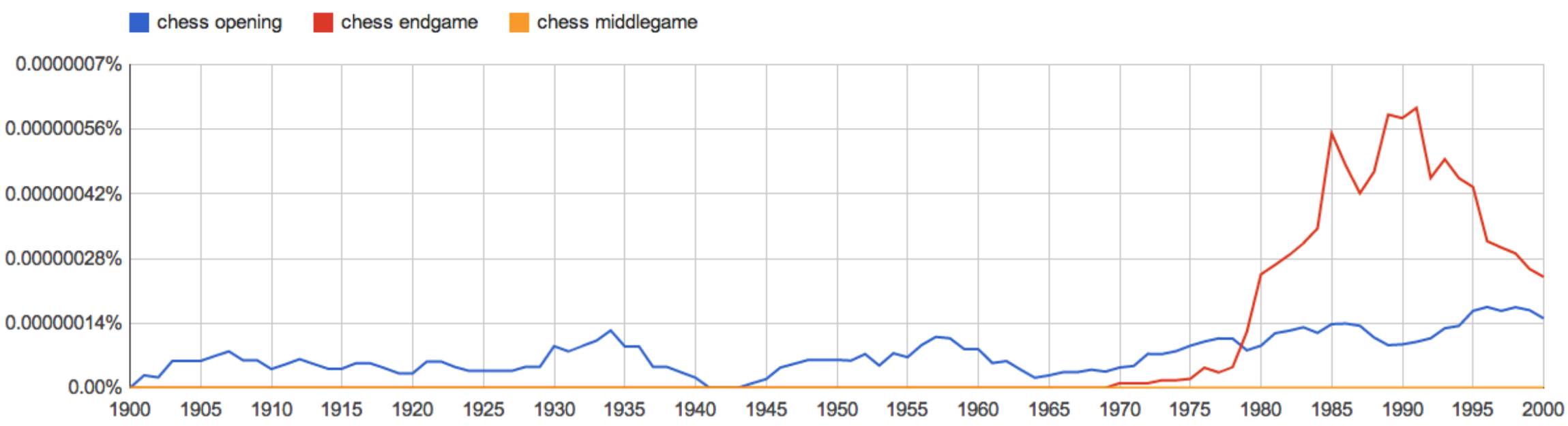
Search lots of books

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!

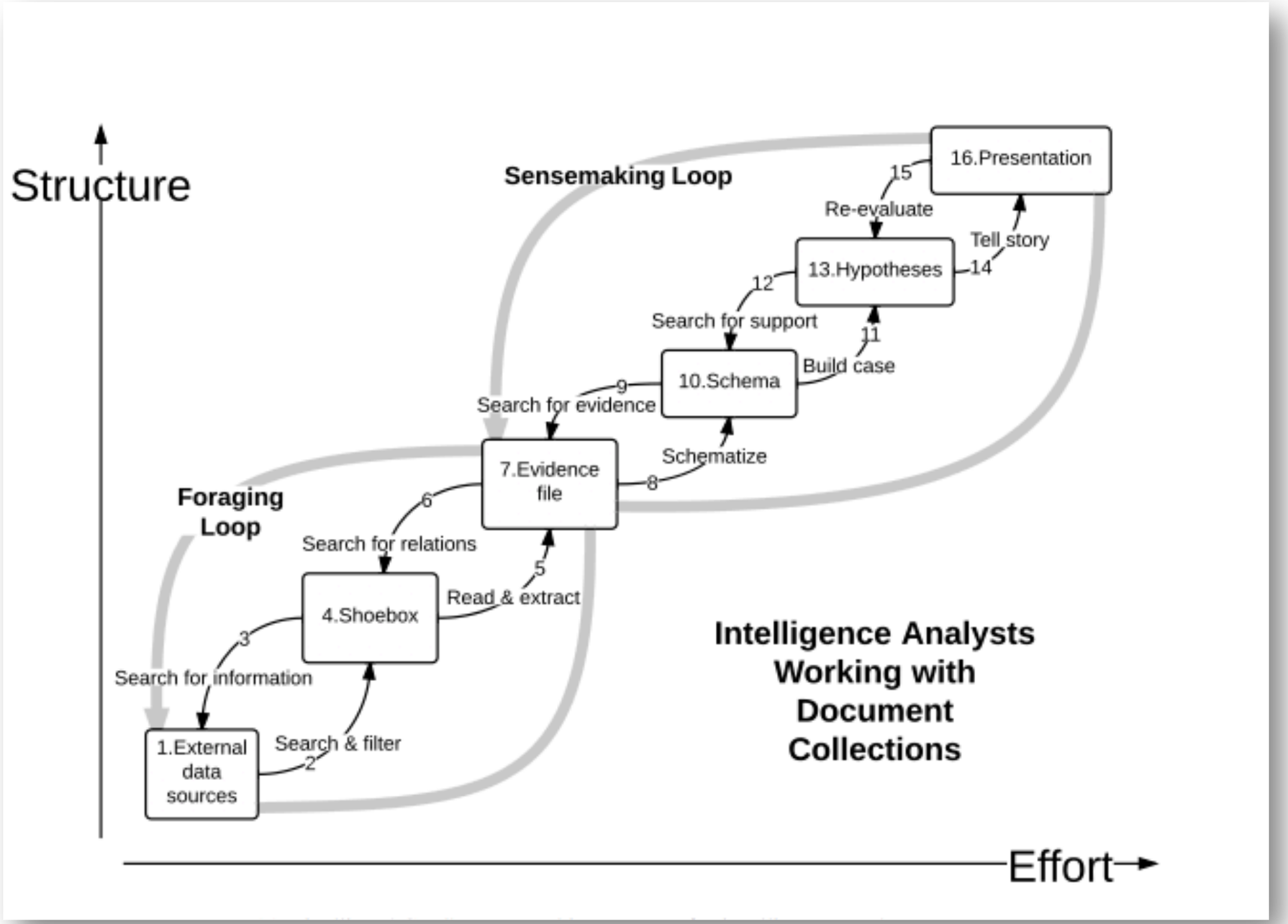
WordSeer Motivation: What do Literary PhDs Want to Do?

- 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)

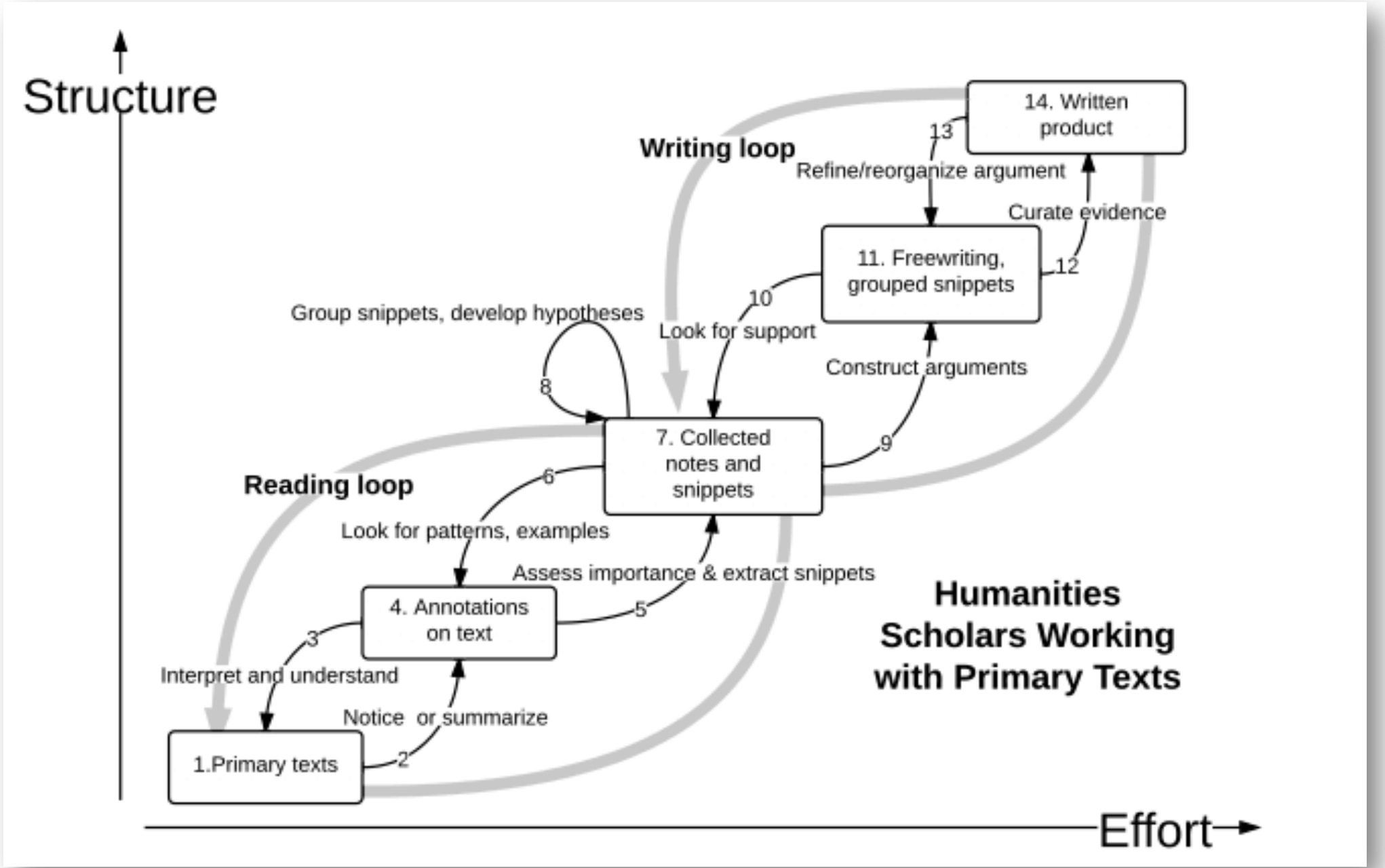
    “I’ve been trying to find out if anyone has ever been positive towards the ‘new woman’, being approving of her. I would love to see how often that phrase shows up during the 1960s” – P12
  - Syntactic patterns and regularities (2/12)

    “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)



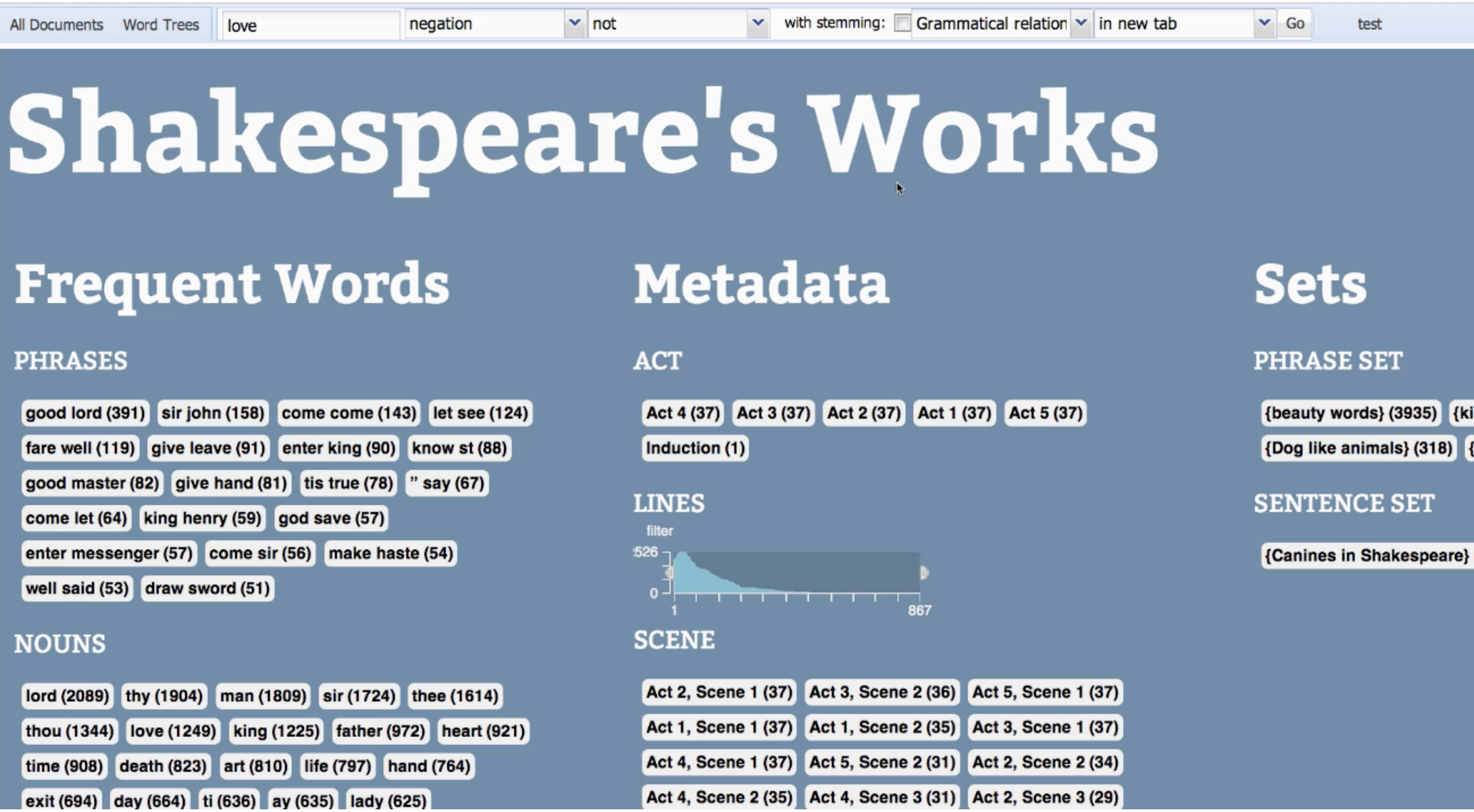
Digital Humanities SenseMaking Loop



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



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

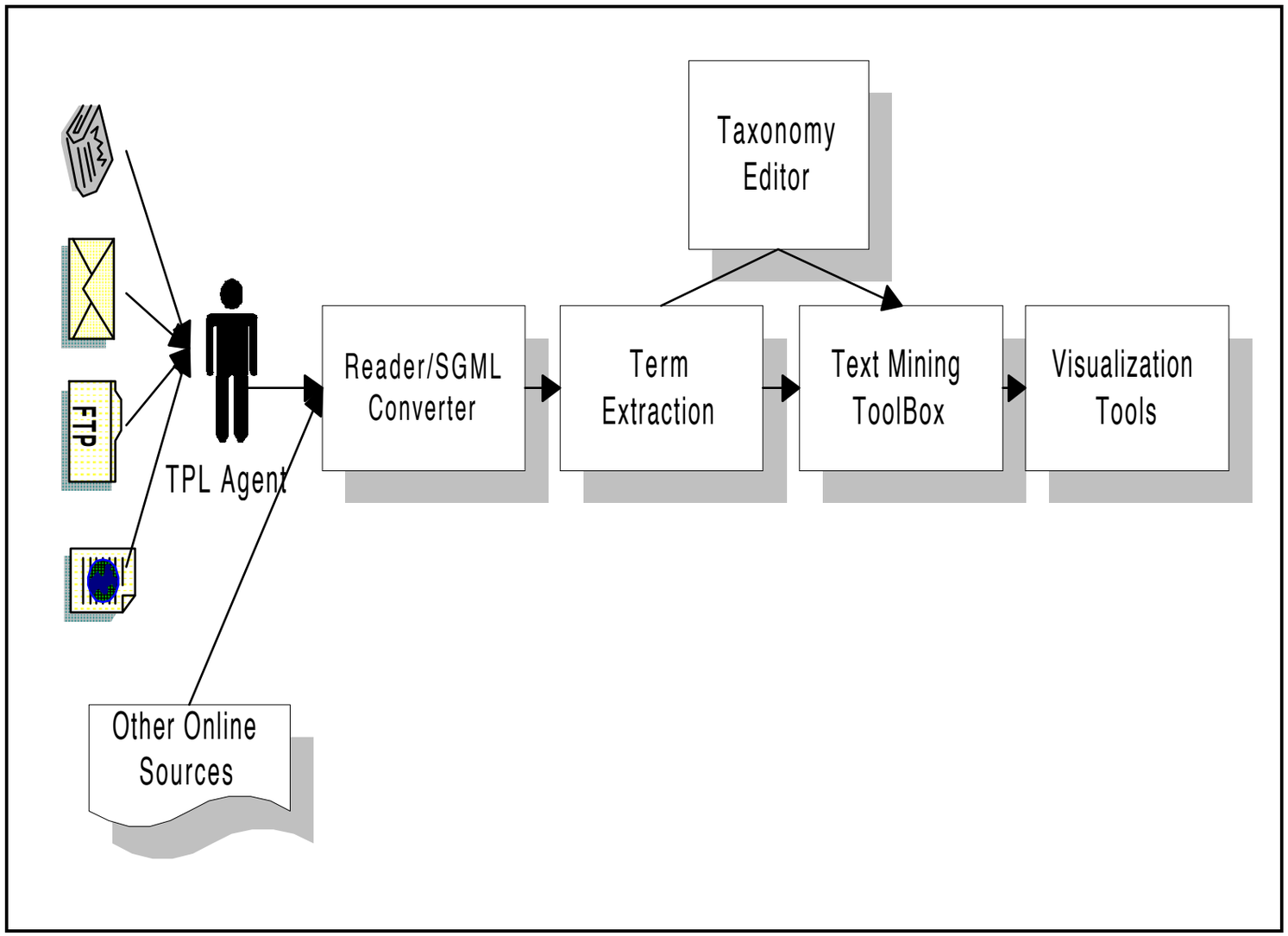
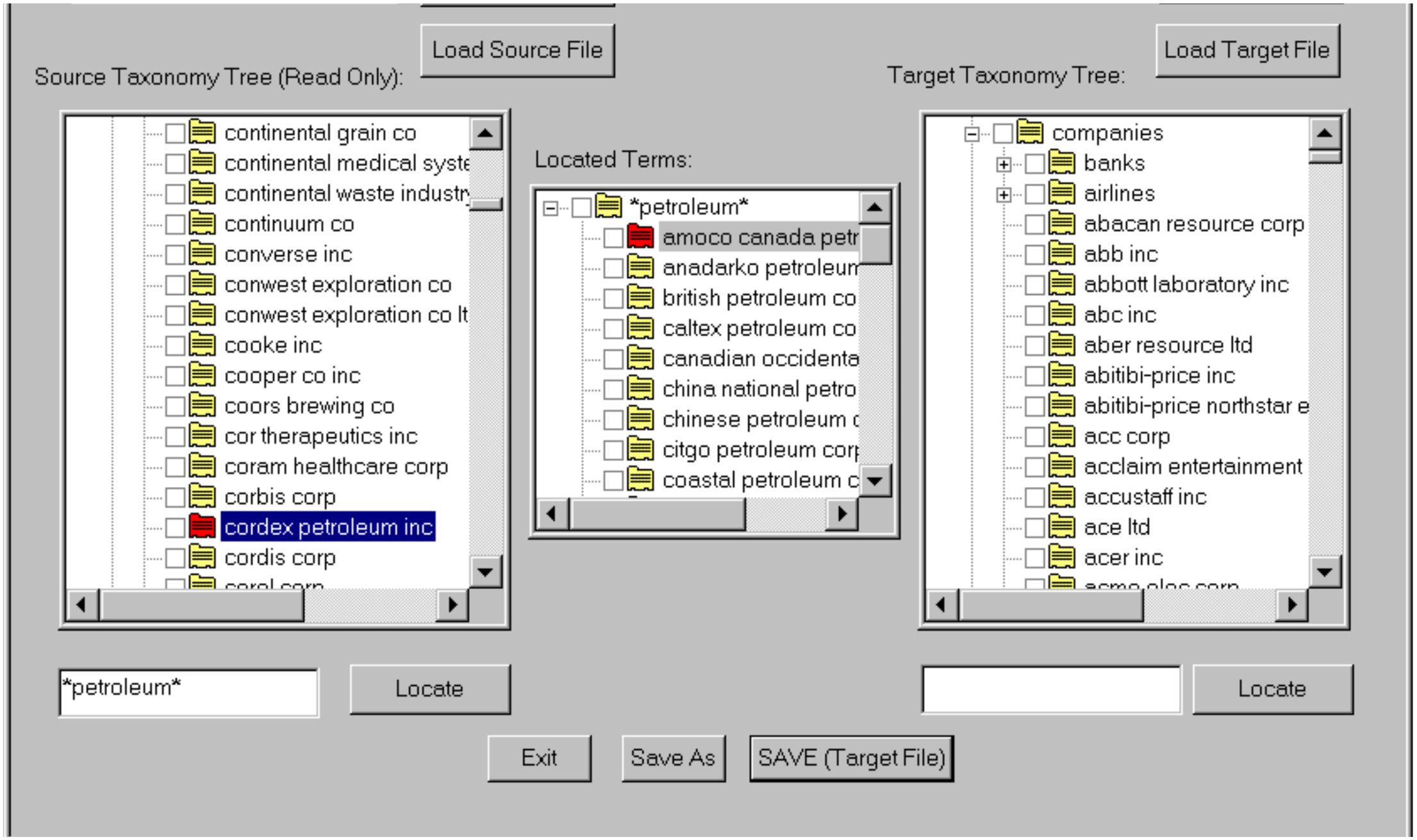
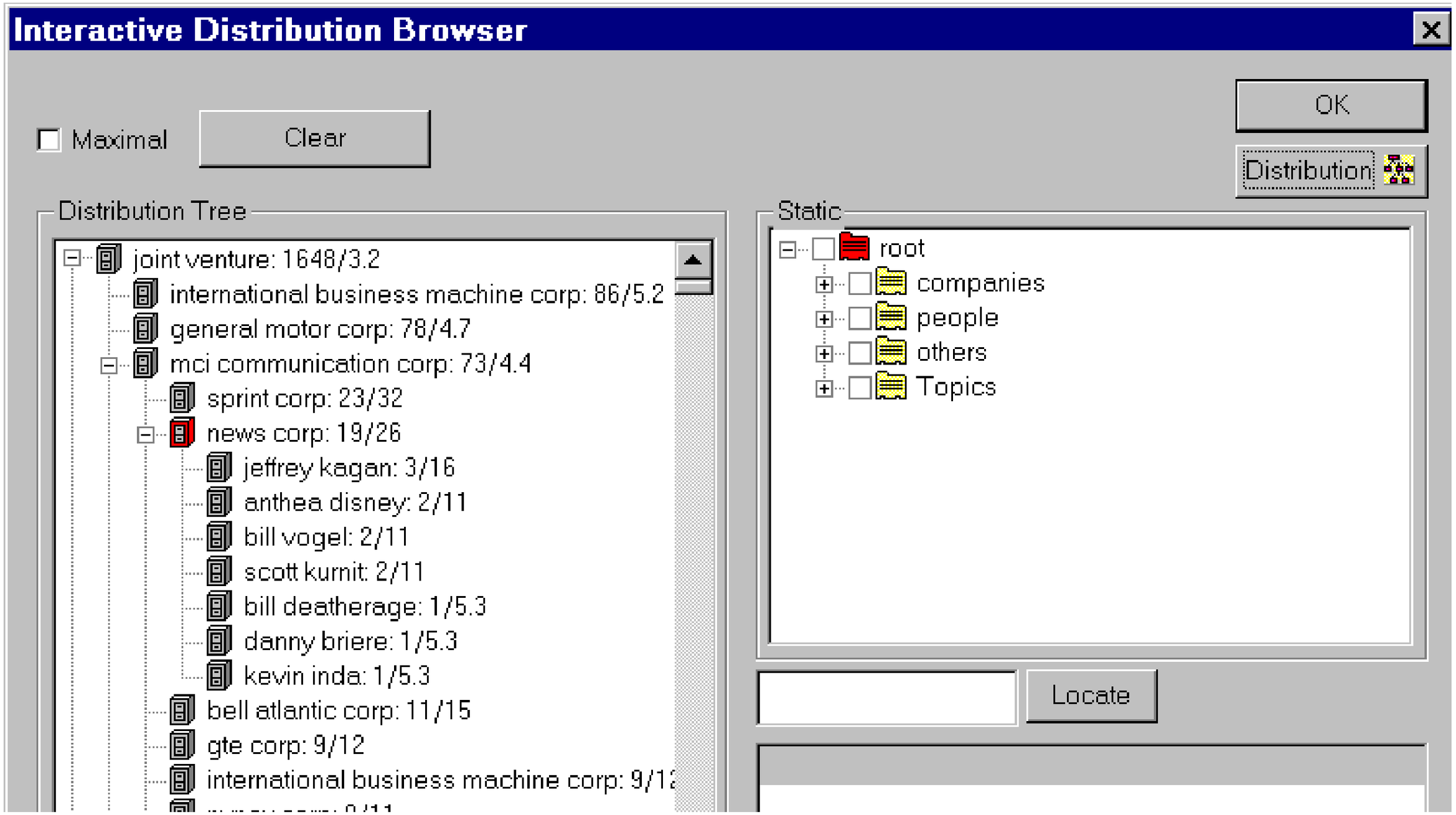


Figure 2. Document Explorer architecture.

Feldman et al. PAKM98



Feldman et al. PAKM98



Feldman et al. PAKM98

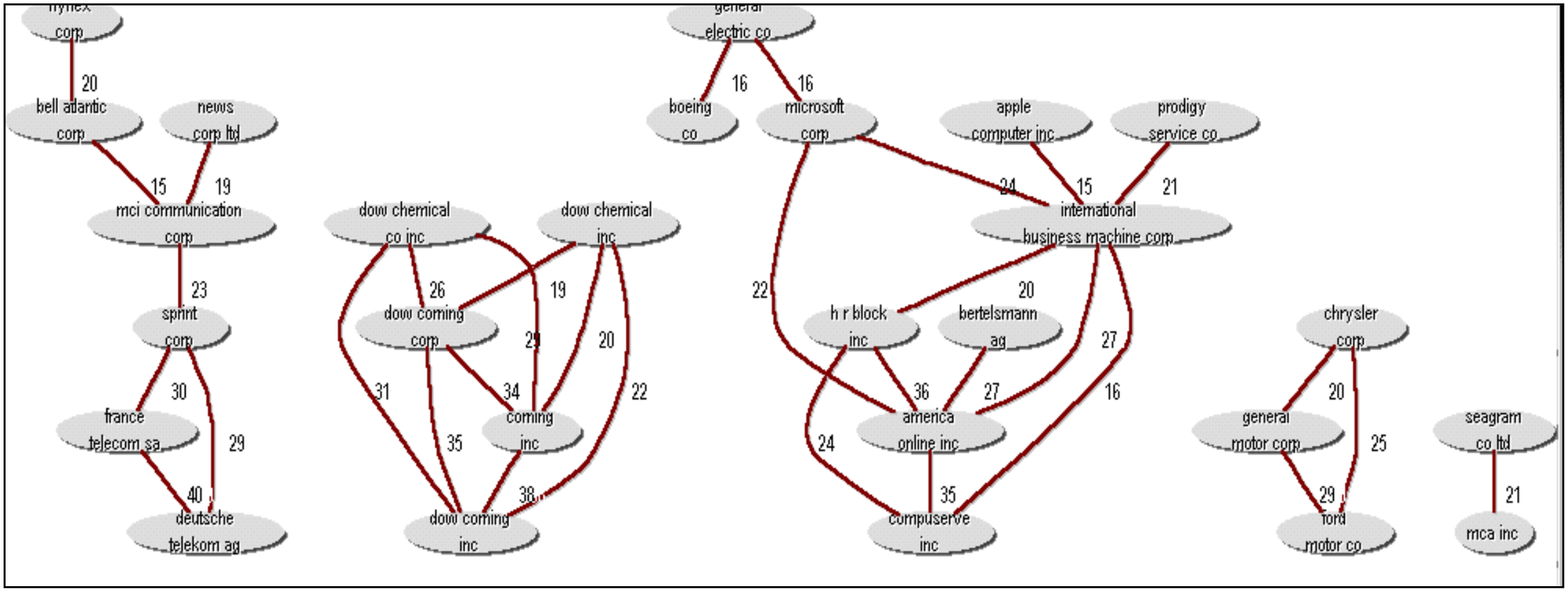


Figure 12 – Context Graph (Companies in Context of "Joint Venture")

Feldman et al. PAKM98

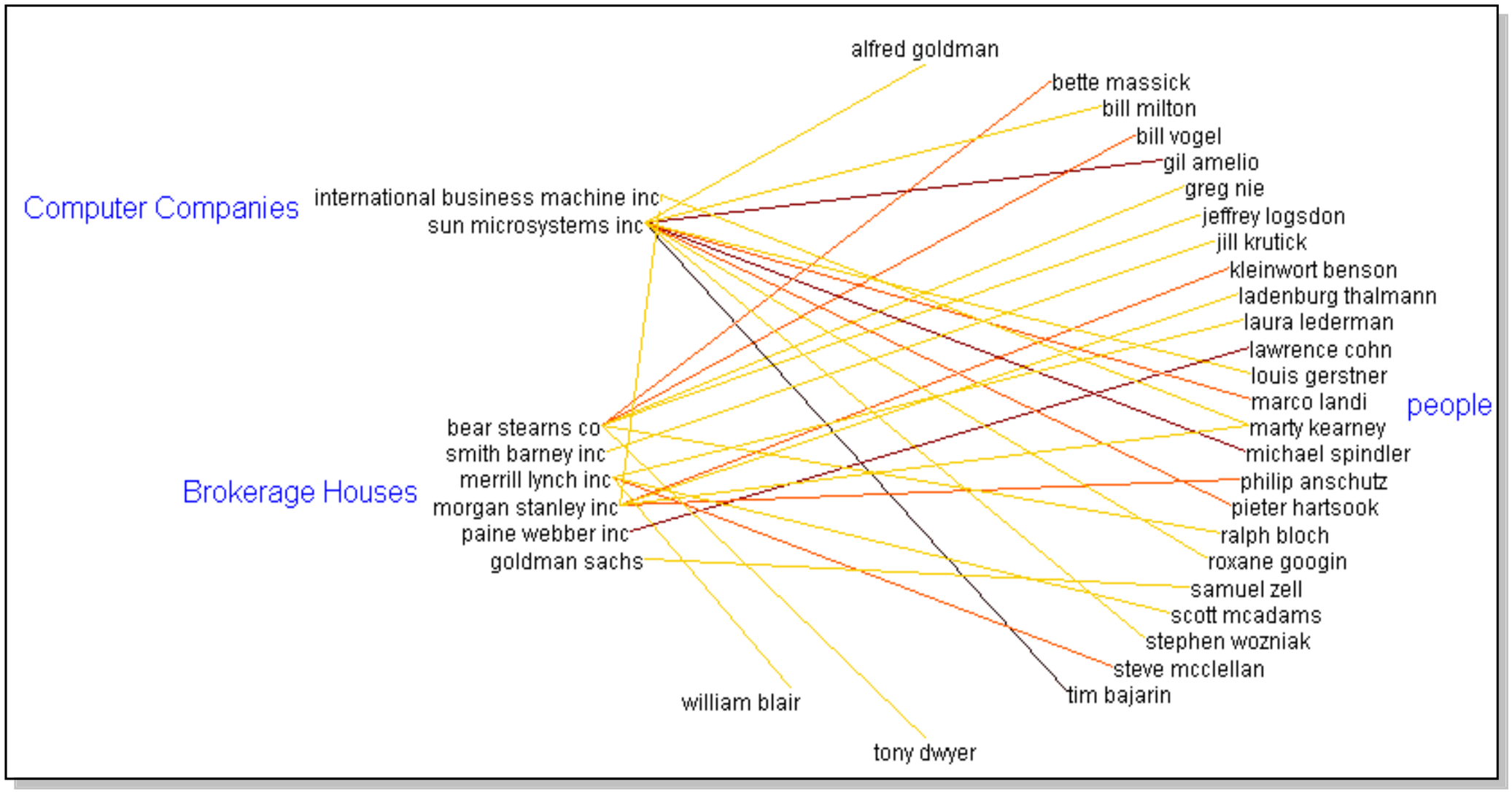
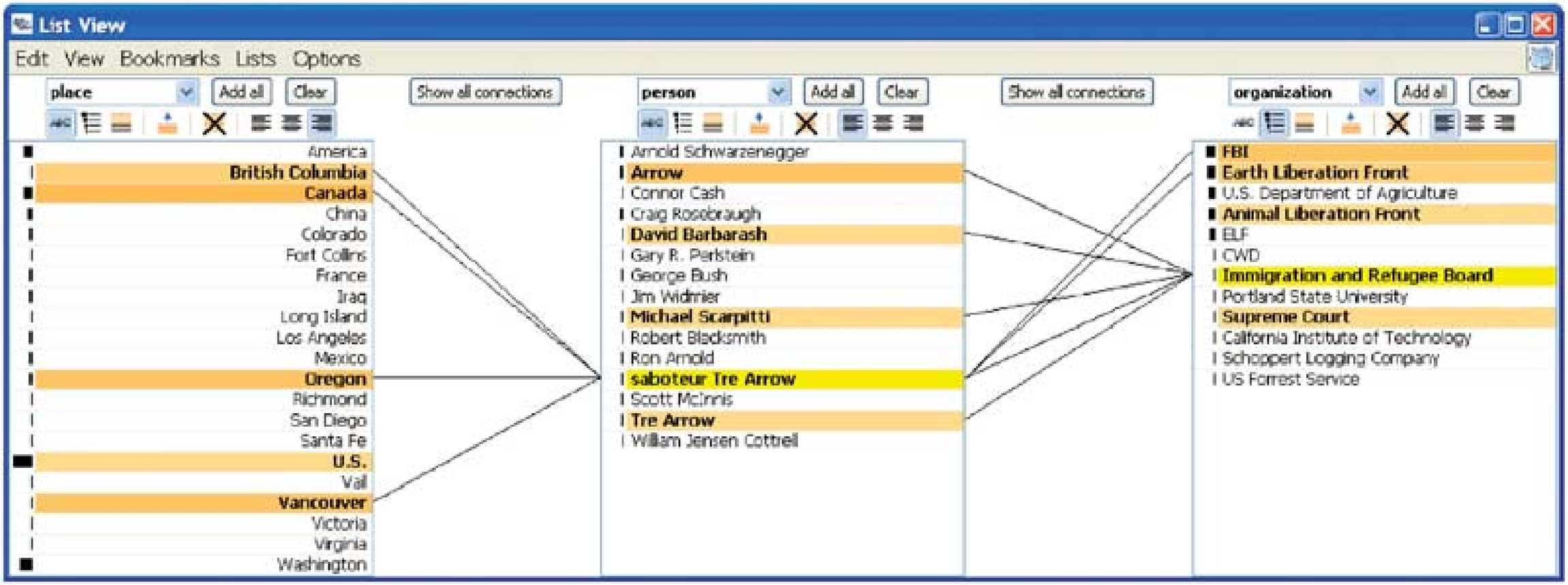


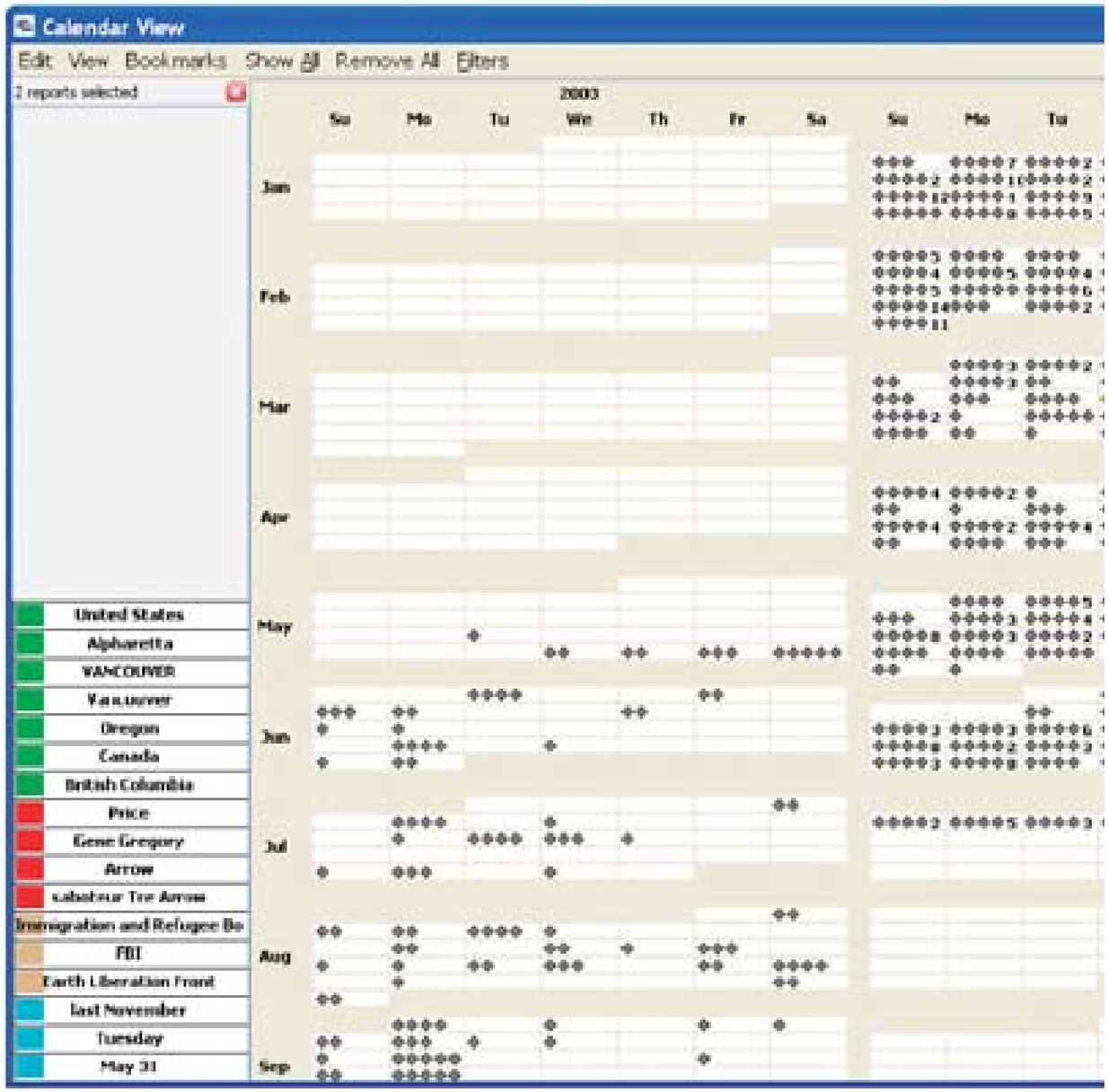
Figure 13 – Category Graph (relationship between People, Brokerage Houses and Computer Companies in Context of "merger")

Stasko et al., Jigsaw, Information Visualization 2008



Stasko et al., Jigsaw, Information Visualization 2008

Stasko et al., Jigsaw, Information Visualization 2008



IBM Content Analytics Version 2.2 (2011)

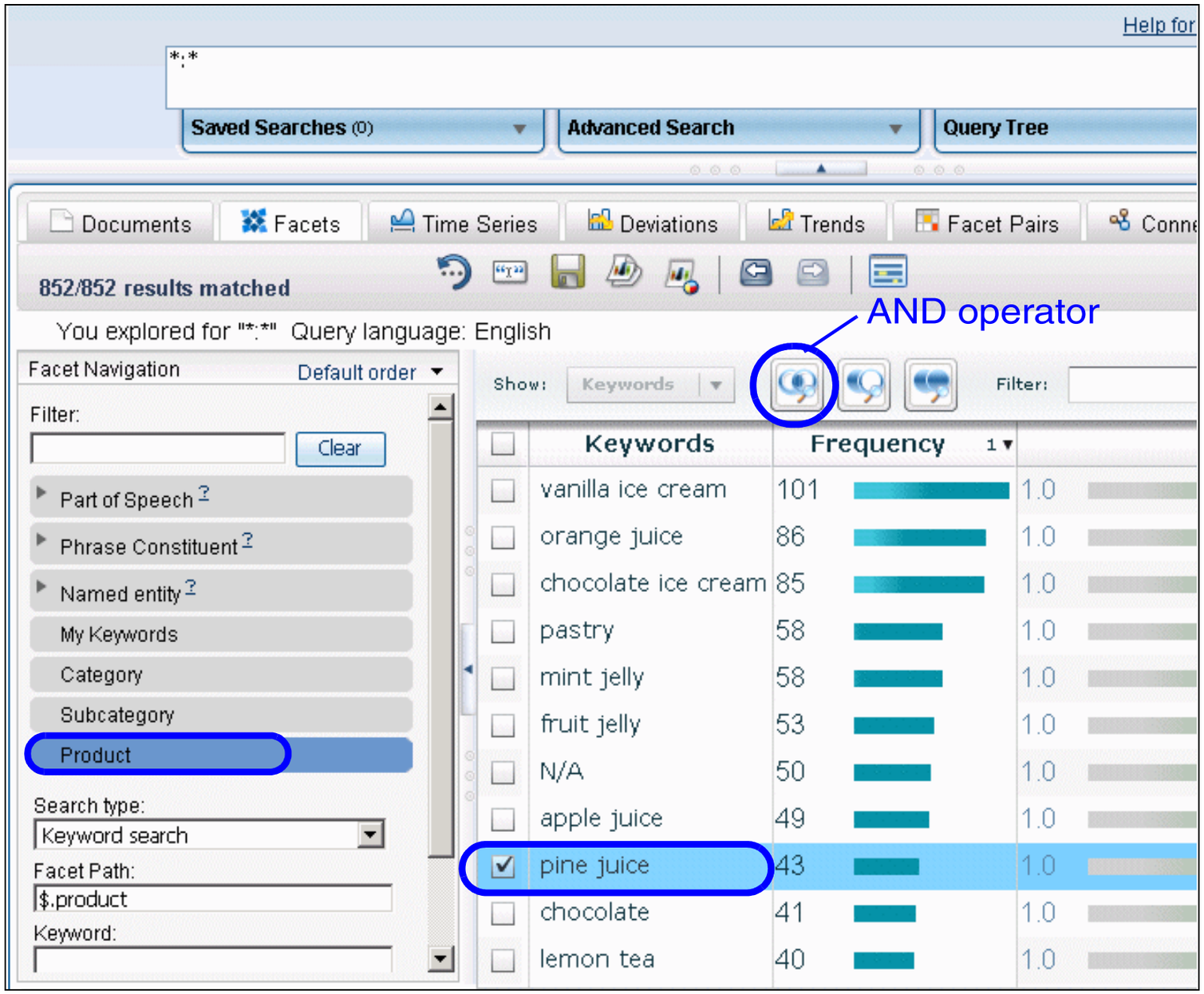


Figure 5-17 Selecting the Product facet and pine juice and clicking the AND operator

IBM Content Analytics Version 2.2 (2011)

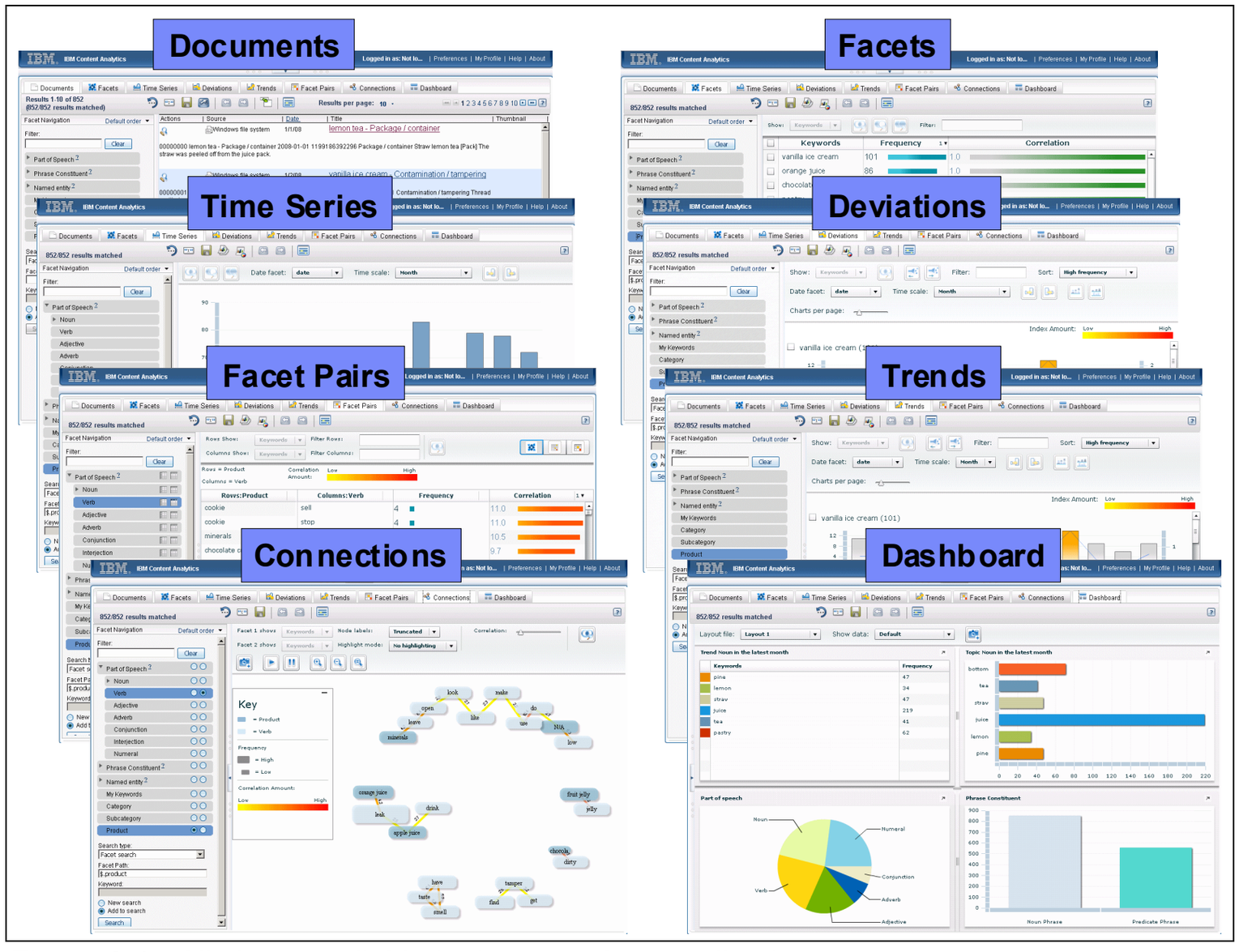


Figure 6-1 All views available in the text miner application

IBM Content Analytics Version 2.2 (2011)

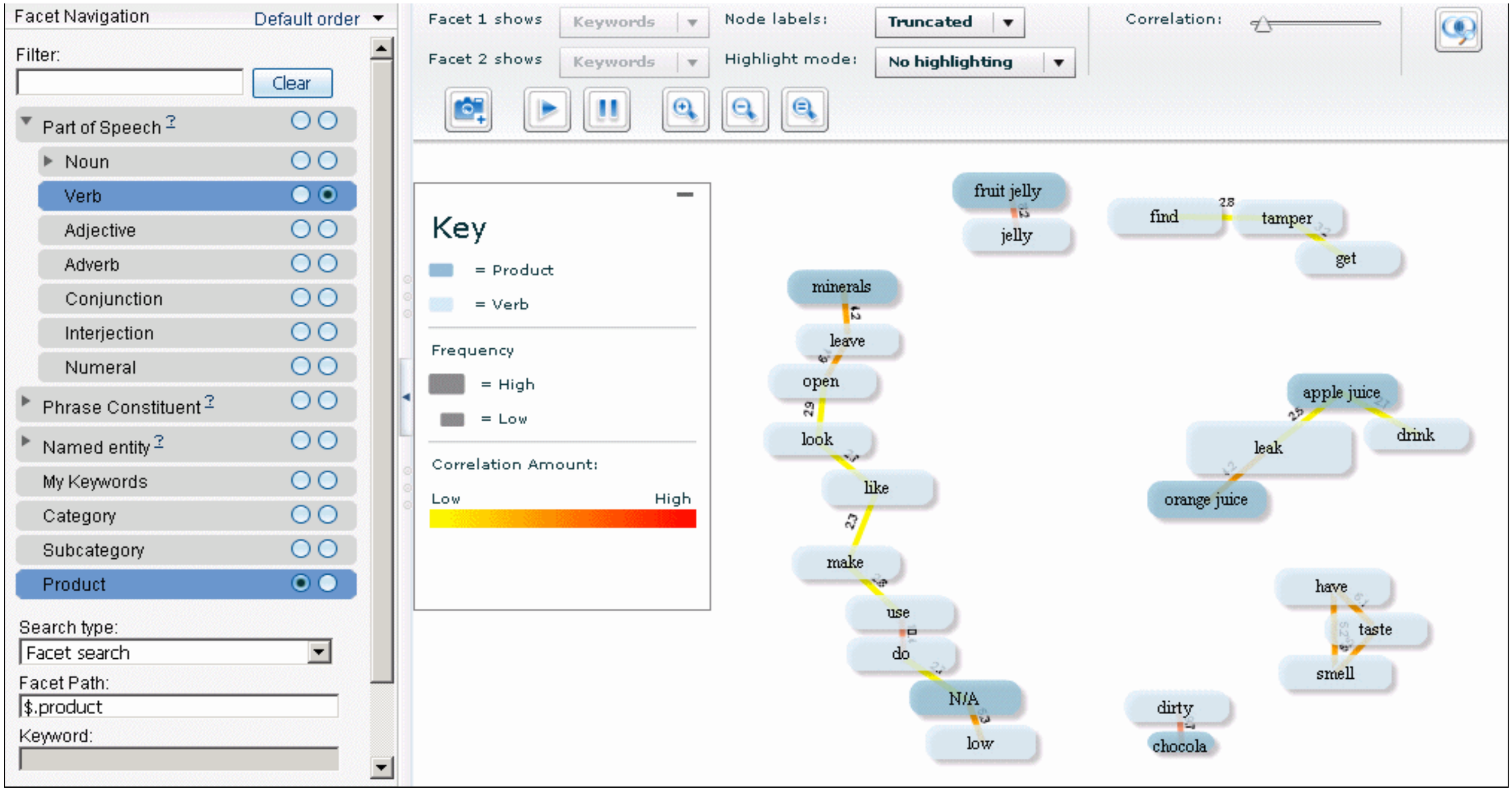
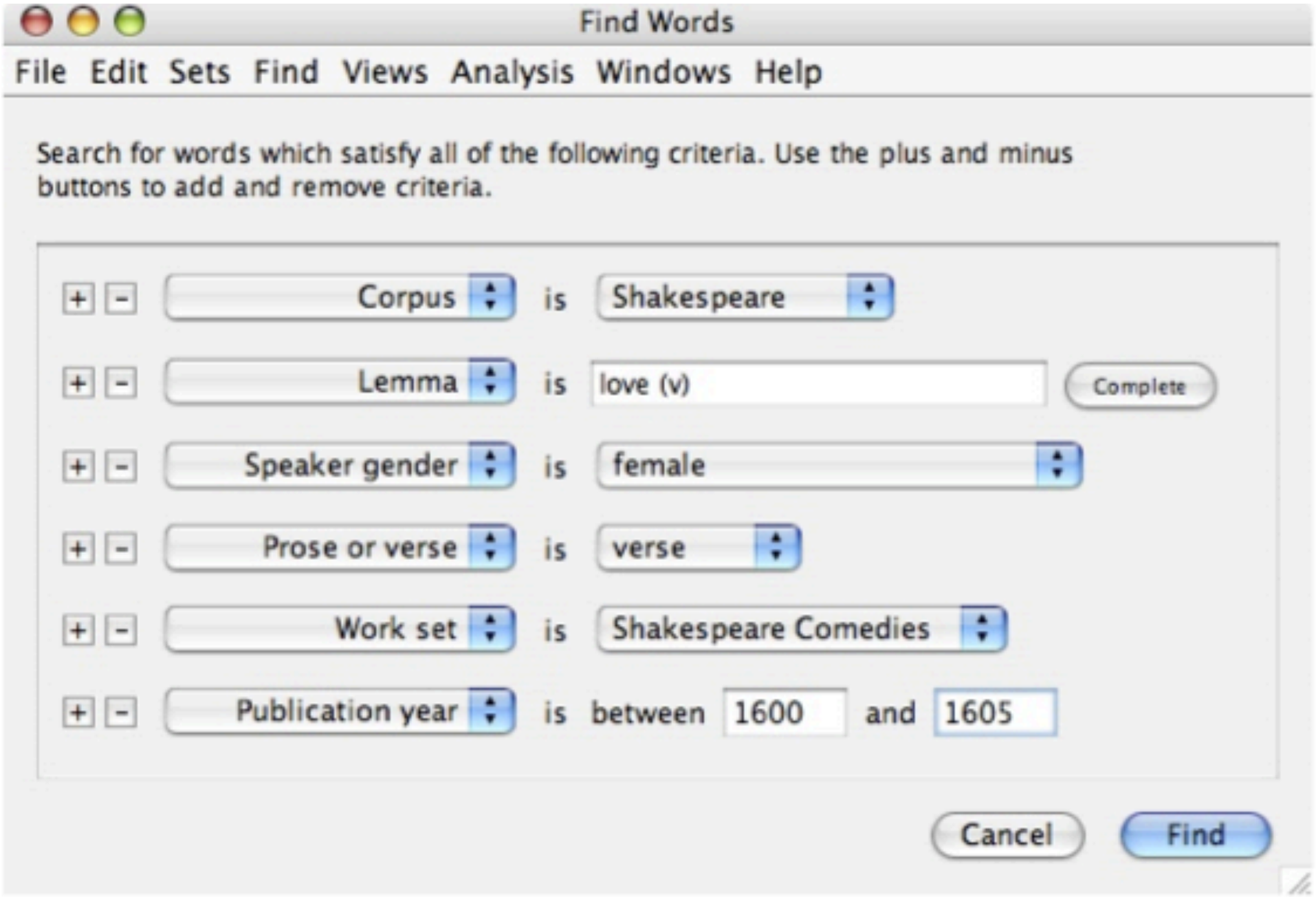
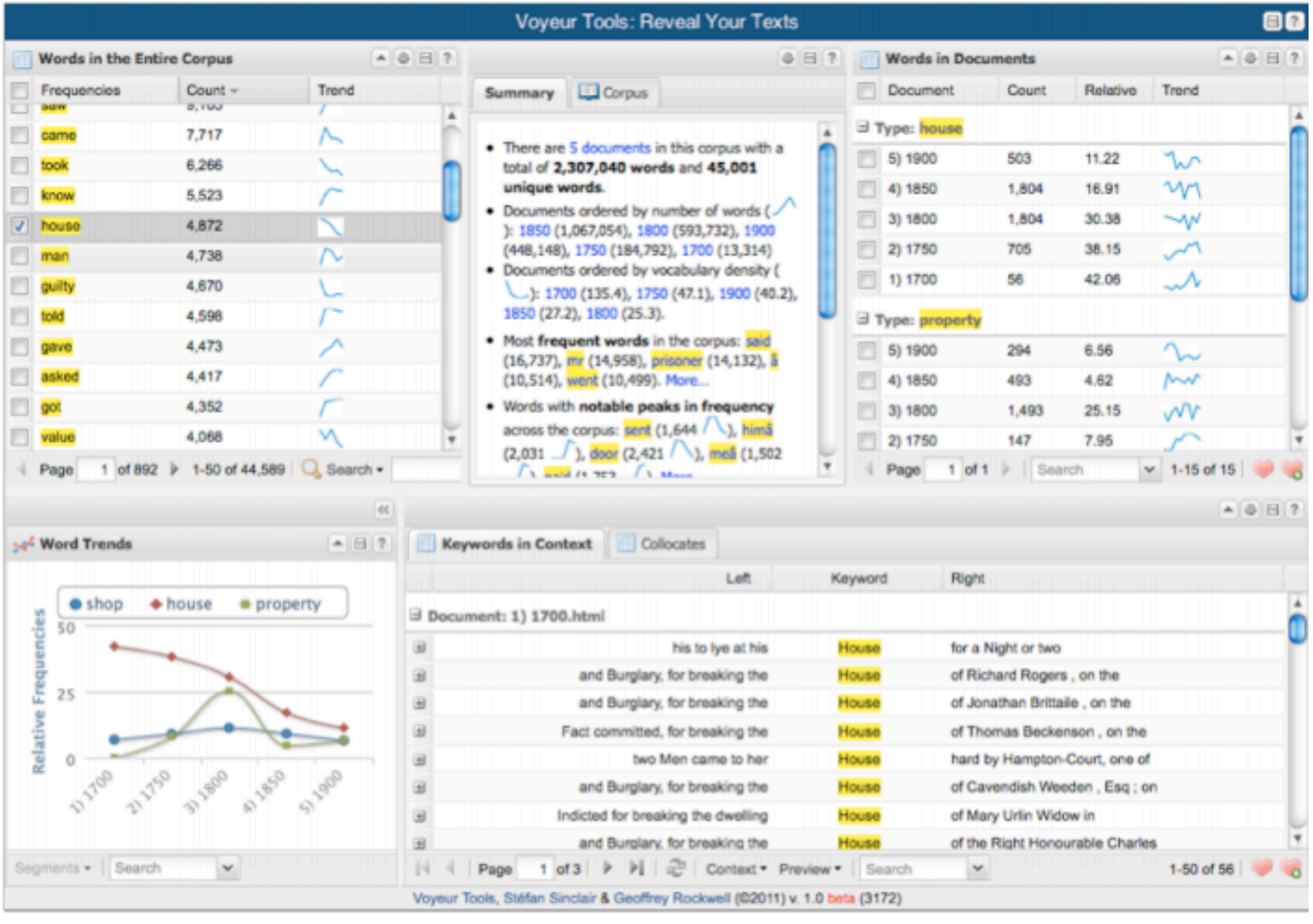


Figure 6-20 Connections view when selecting Product facet and Verb facet

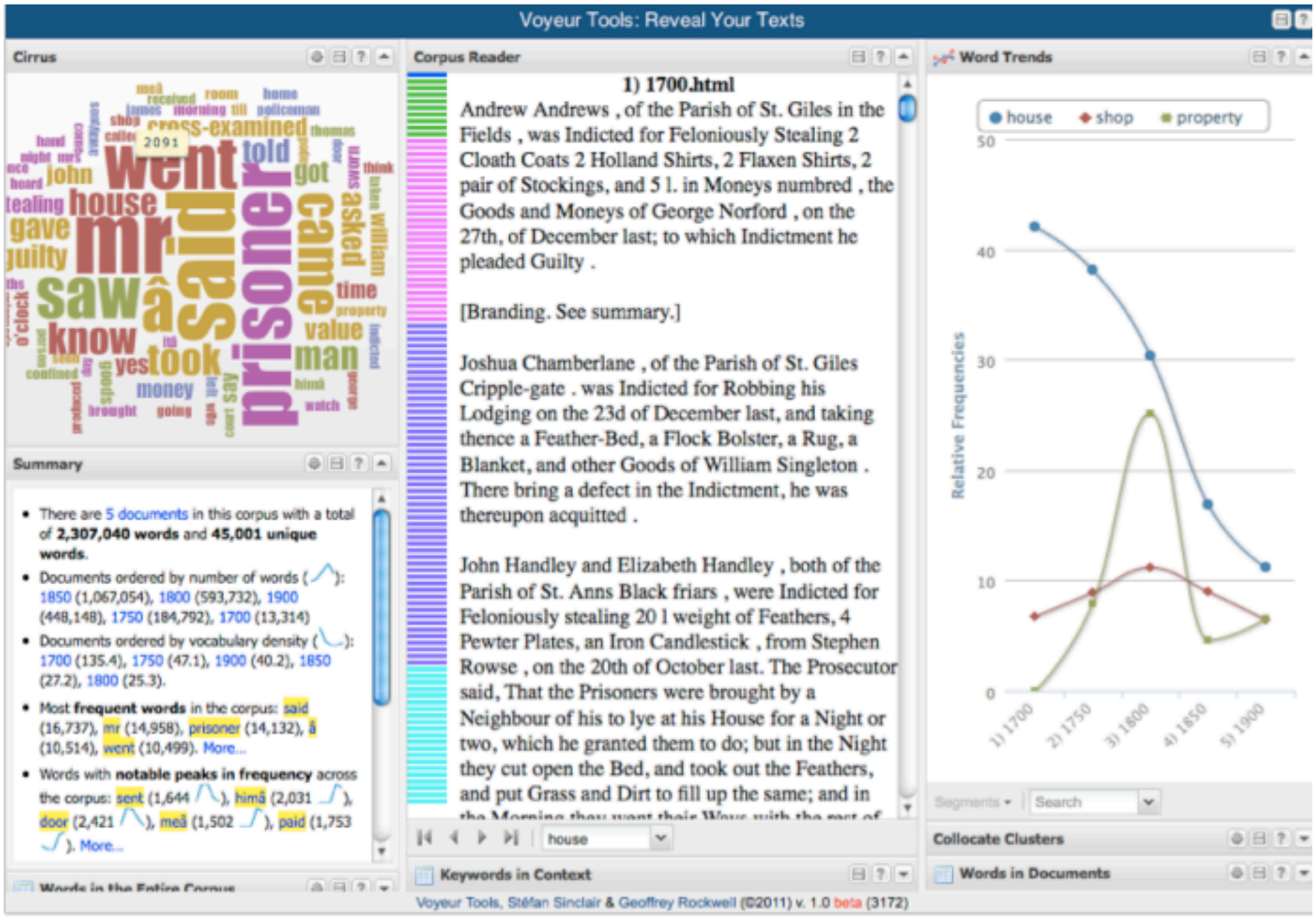
Mueller, WordHoard, Shakespeare, 2008



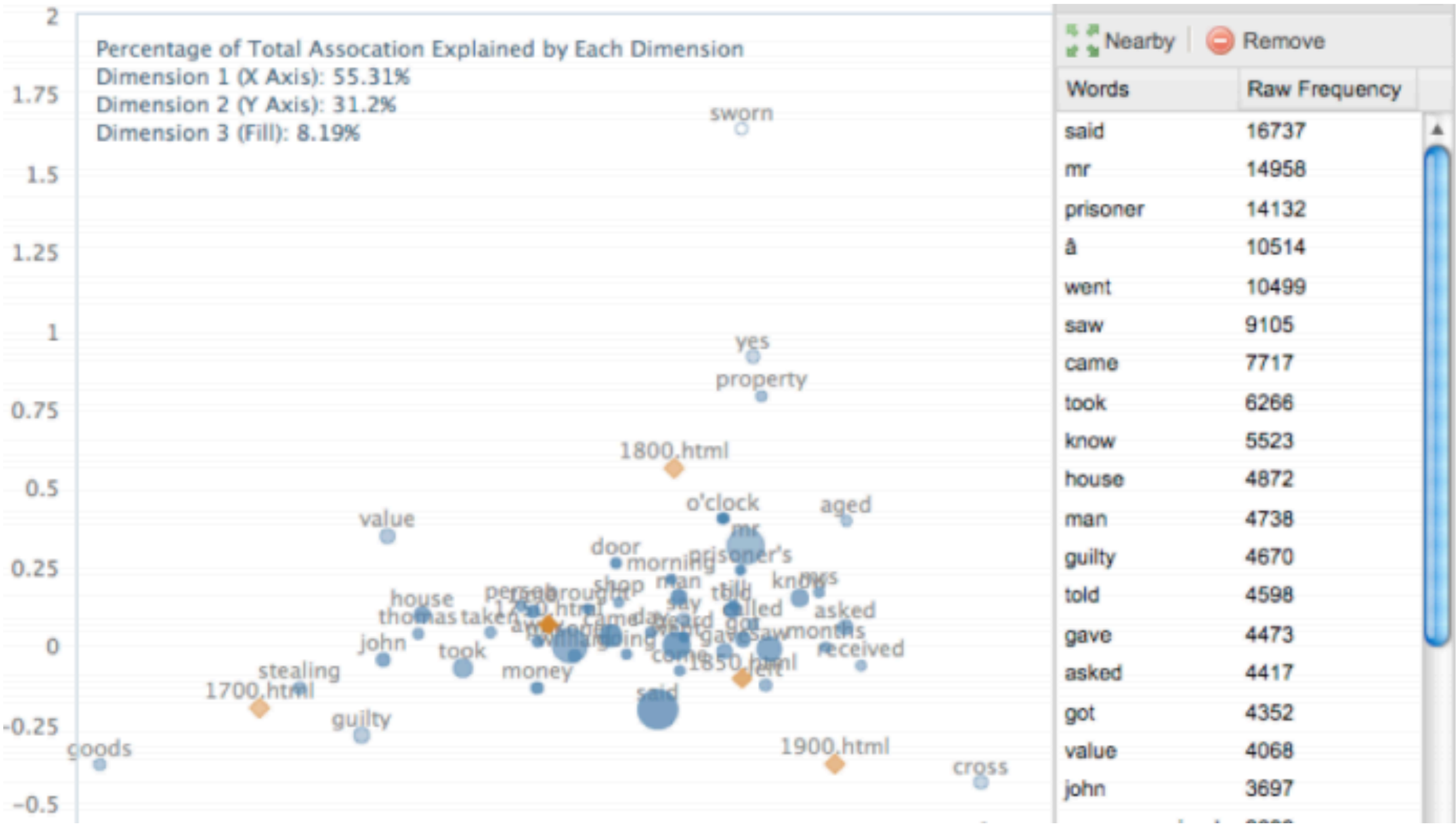
Sinclair & Rockwell, Voyant/Voyeur, TLLO 2009



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

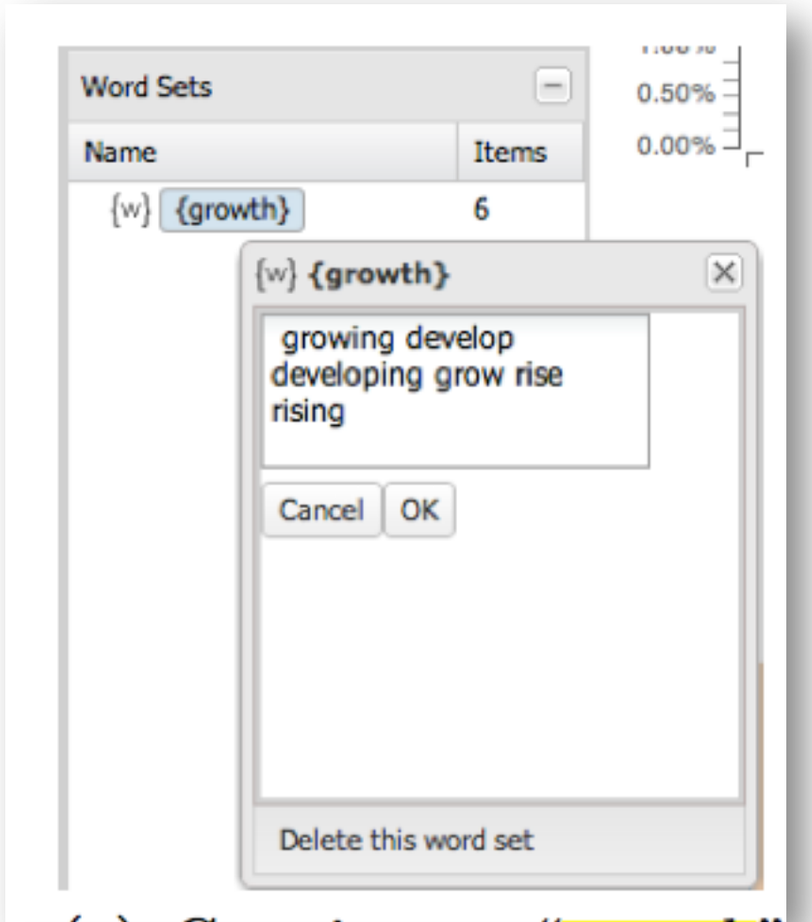
1. Literary scholar studying American literature's reaction to China's rise

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

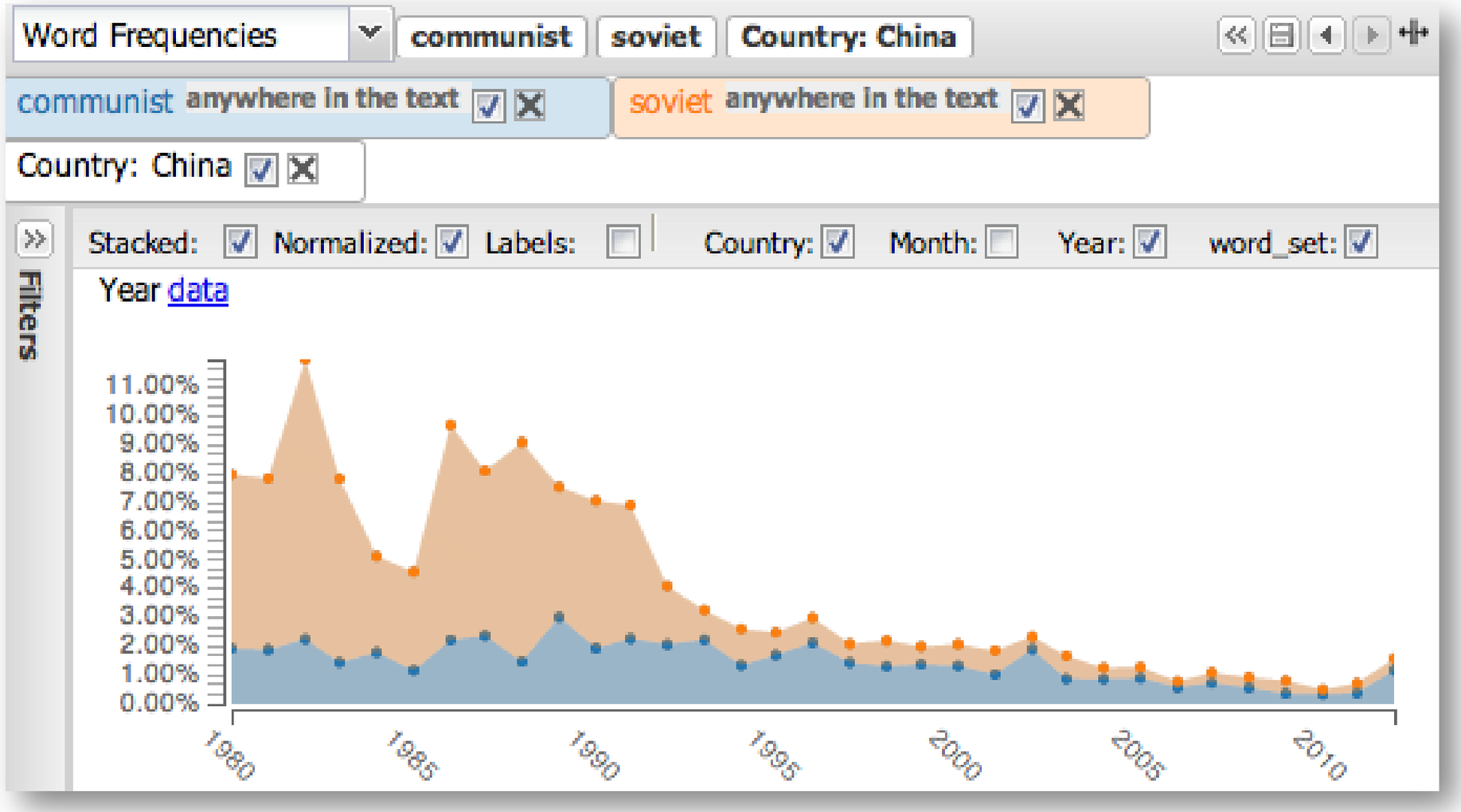


China Scholar: Concept Trend

- 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.)

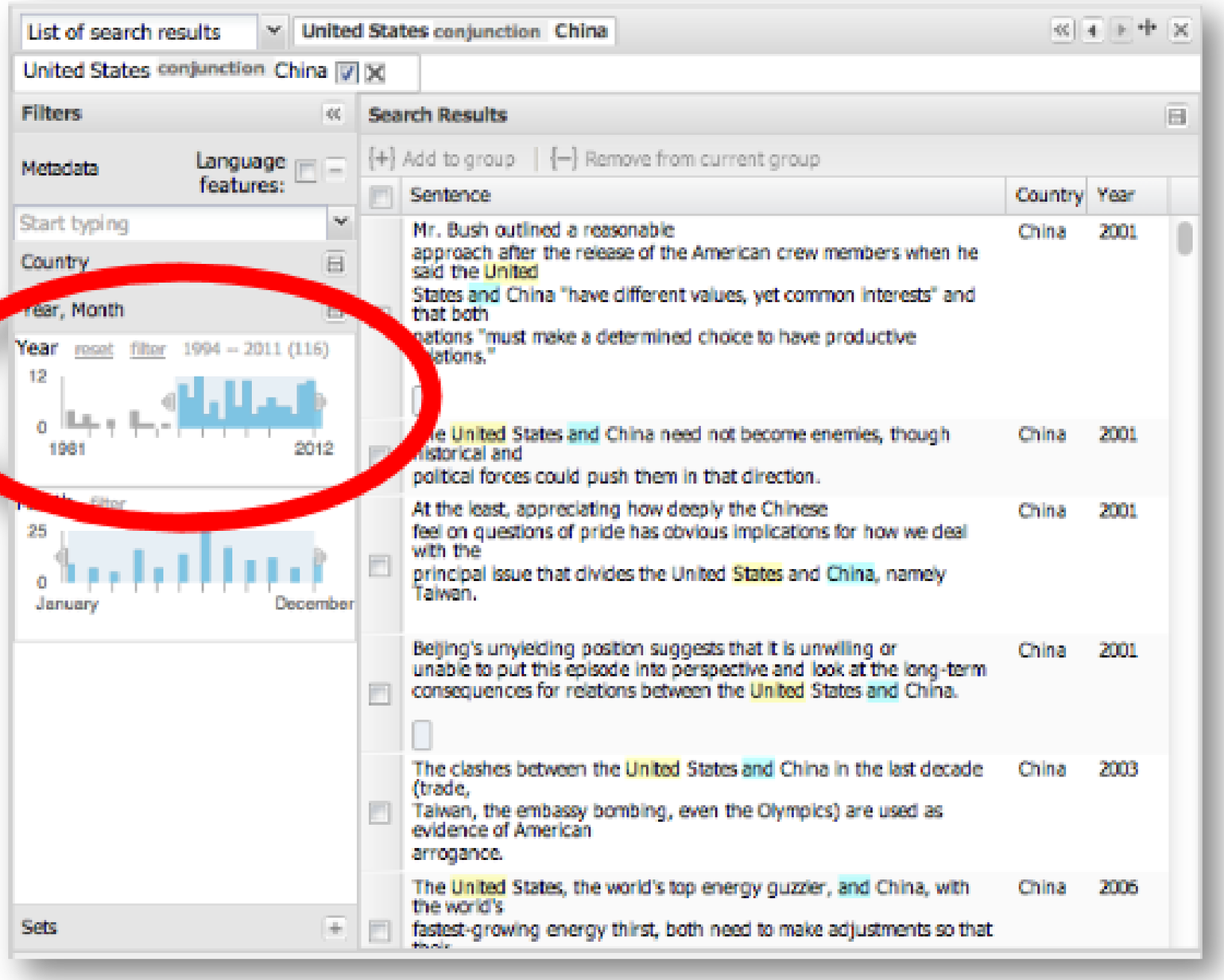
- China joined the WTO in 2001; this is when China-US relations are thought to have become more inter-dependent.

- 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). "

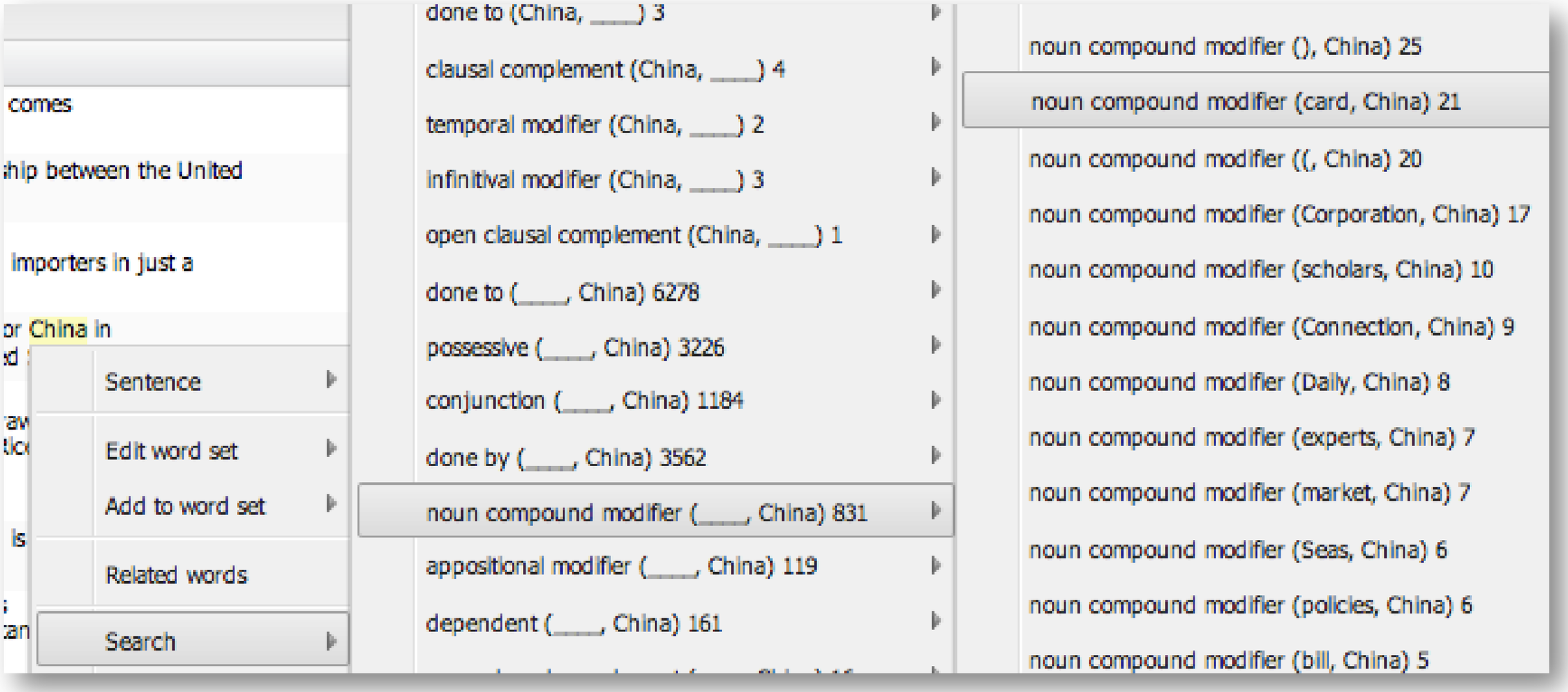
China Scholar: Mid-90s Onward: conjunction(China,U.S.)

- Conjunctions much more frequent after 1994



China Scholar: Explore the “Grammatical Neighborhood’

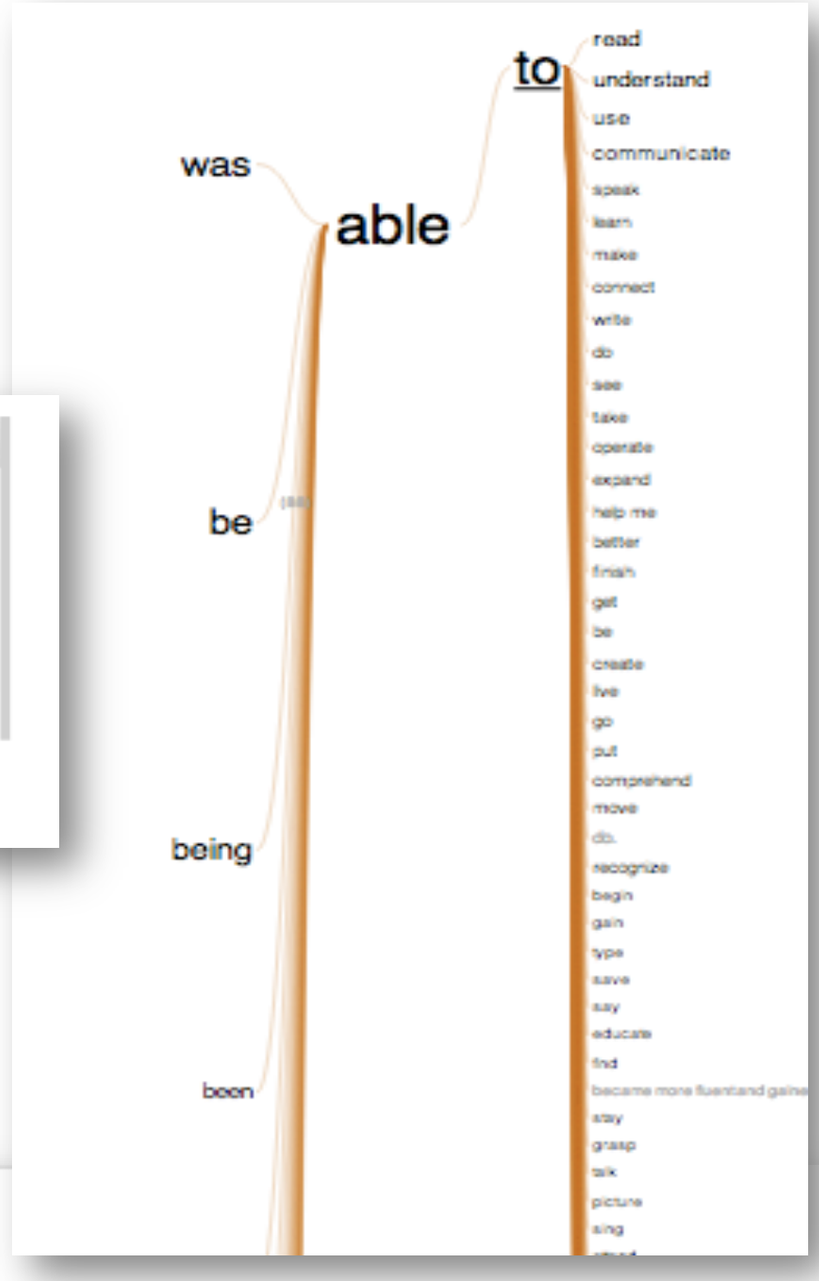
- Interested in: how China is talked about: anything interesting or unexpected?

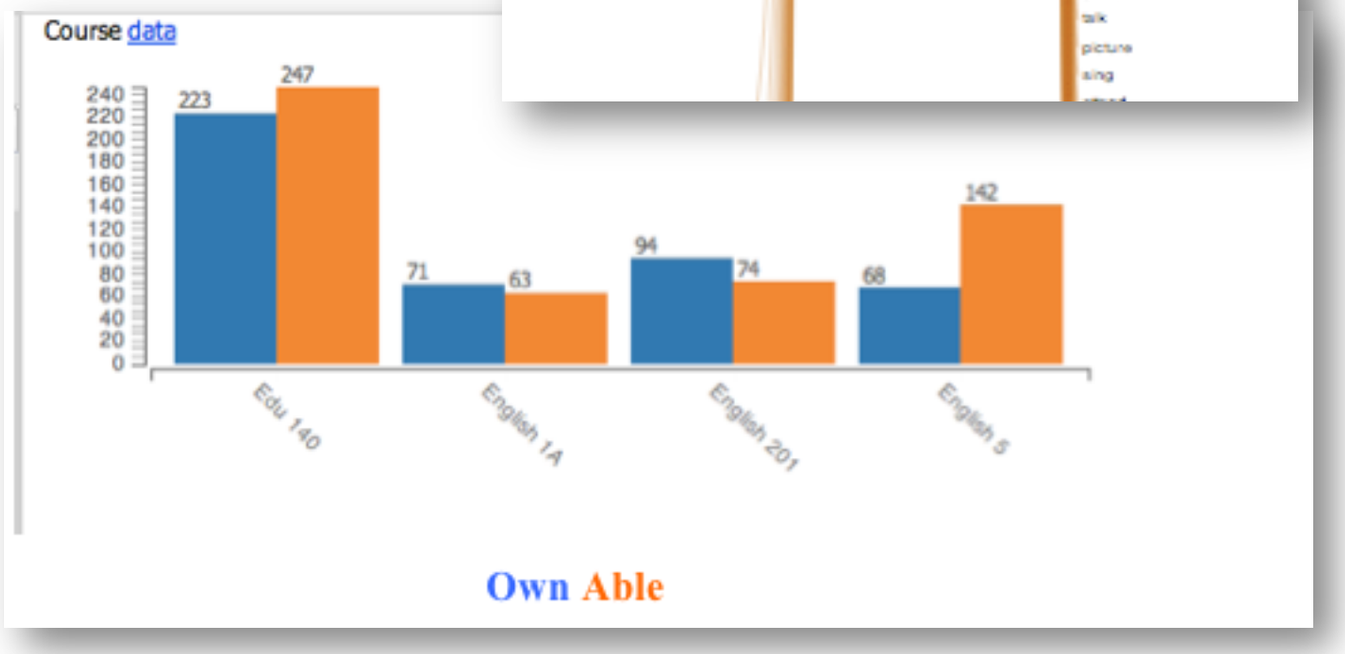


Literacy Essays

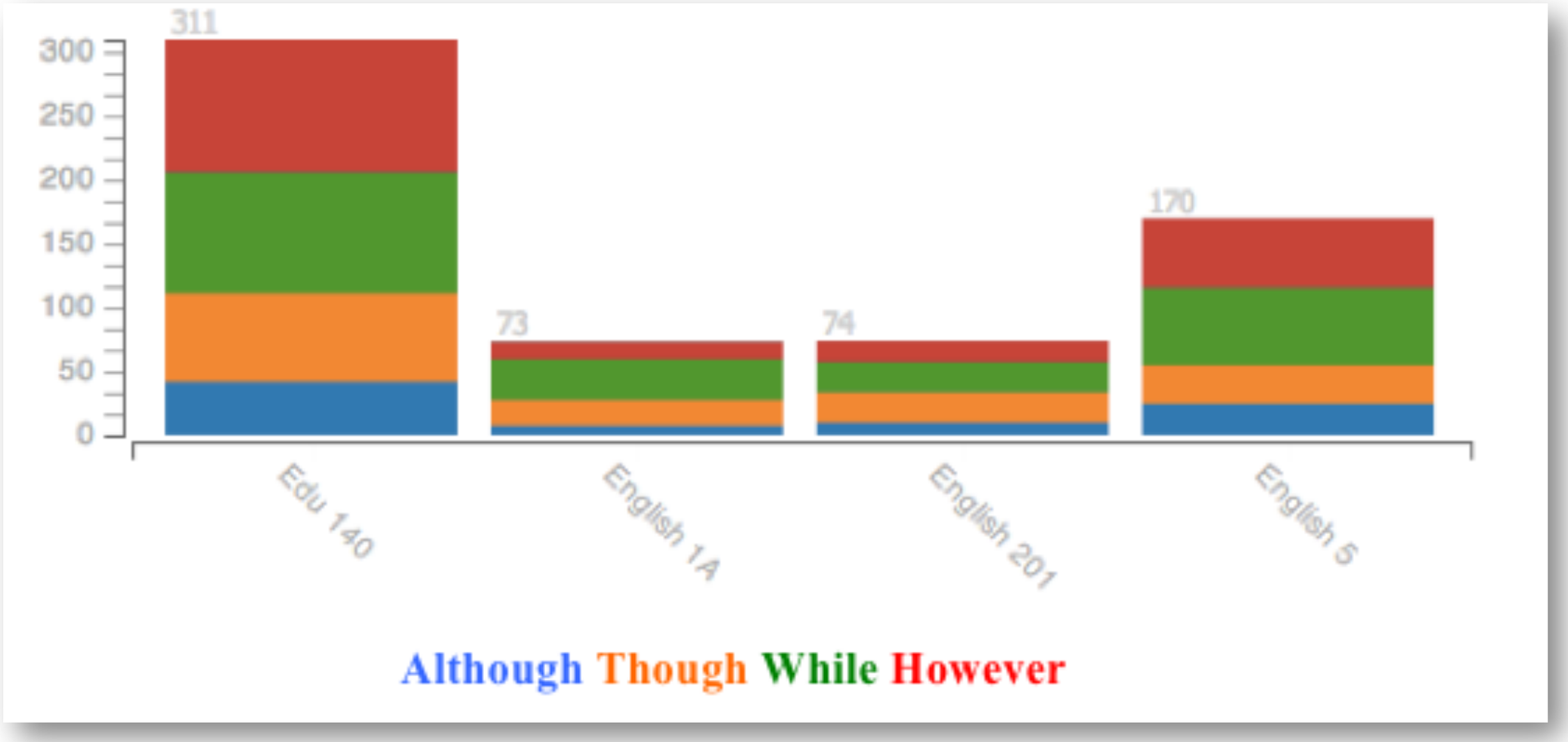
Frequent Adjectives	group by stem:
other	462
first	382
new	371
able	316
english	286
own	273

Figure 2 – Most Frequent Adjectives

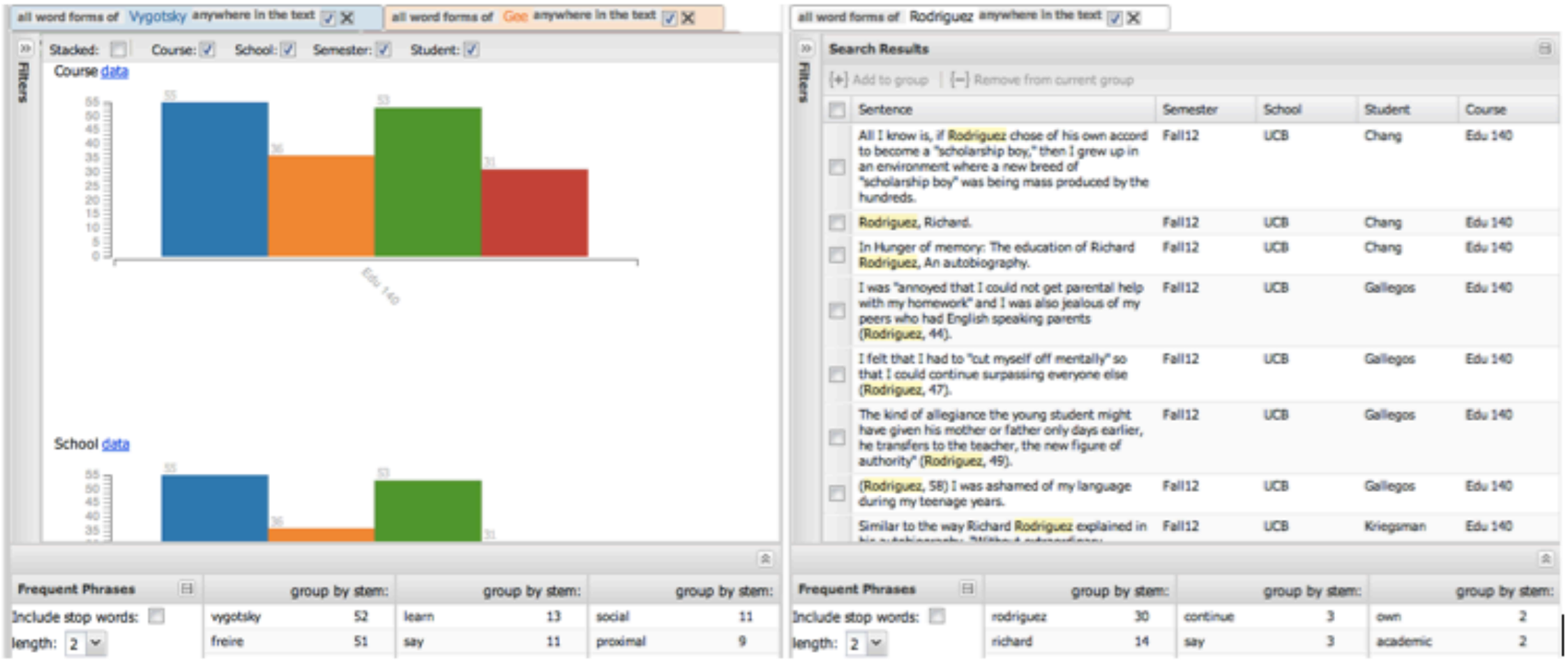




Literacy Essays

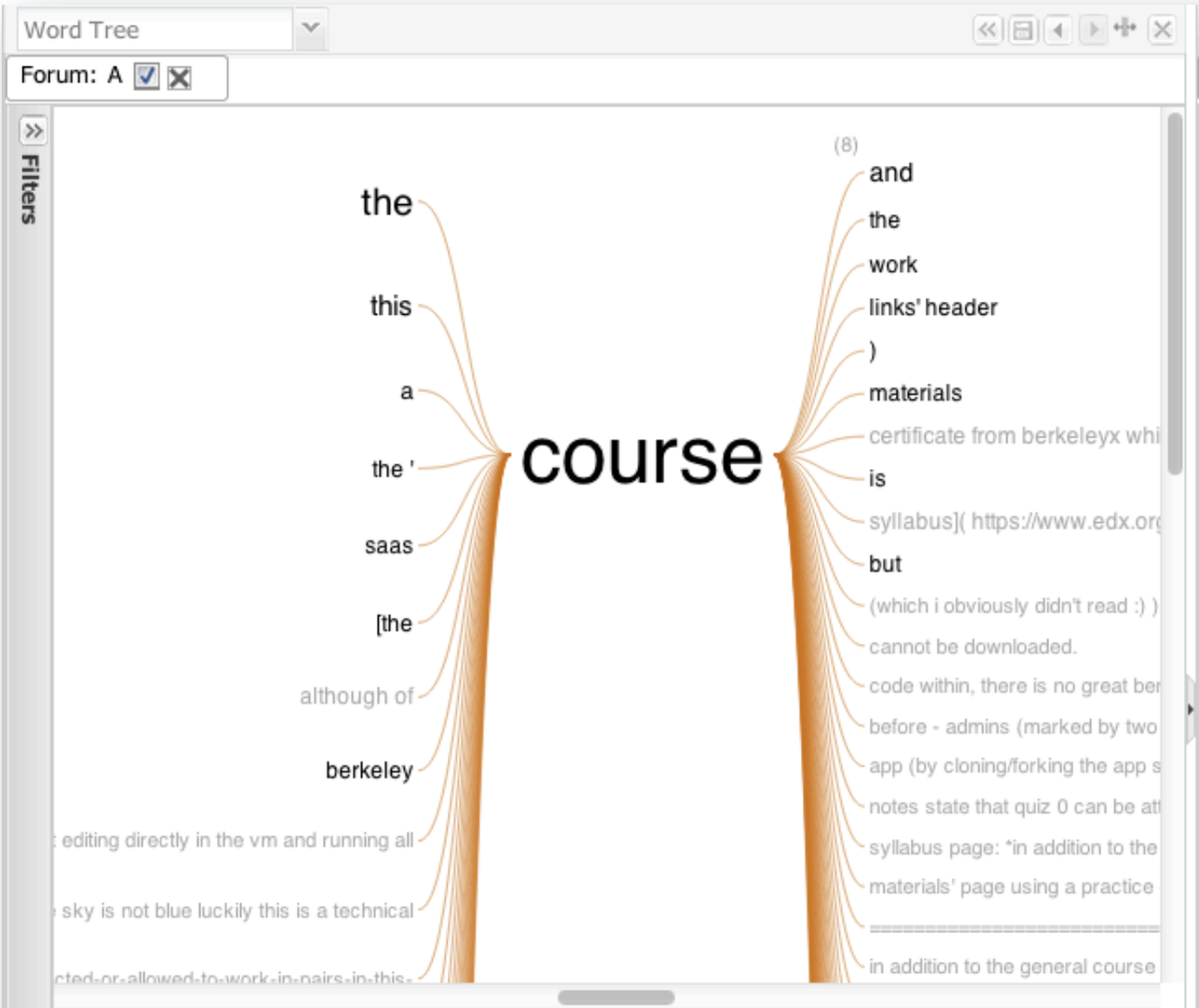


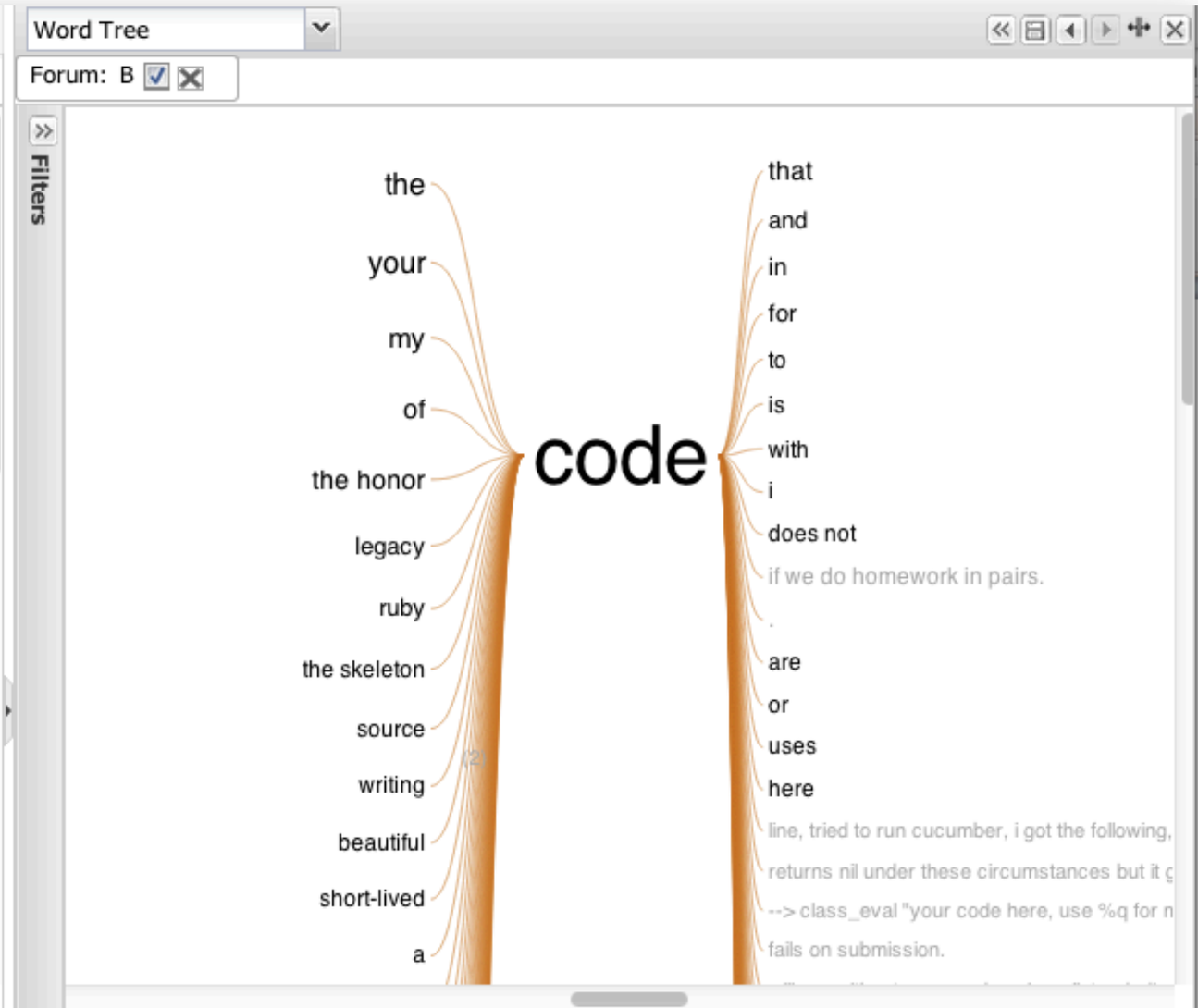
Literacy Essays



Vygotsky **Gee** **Freire** **Rodriguez**

Forum Analysis





Frequent Phrases

Include stop words:

length: 2

usr/local/lib ruby/gems/1	35
ruby/gems/1 .9.1	31
lib/rspec runner	30
rvm/rubies/ru...	24

Fre...	group by stem:
course	147
question	142
answer	121
time	112
error	102
homework	91
thanks	83

Fre...	group by stem:
get	203
use	176
try	131
work	130
think	122
run	110
see	106

Fre...	group by stem:
same	86
other	69
new	58
=	55
sure	47
more	45
first	43

Frequent Phrases

Include stop words:

length: 2

puts rps_game_wi...	16
due date	13
rake db	12
lib/rspec runner	12

Fre...	group by stem:
movie	99
code	86
method	84
thanks	80
time	67
problem	66
error	62

Fre...	group by stem:
get	135
use	122
work	103
try	83
see	70
run	59
find	50

Fre...	group by stem:
=	77
same	68
other	29
new	28
due	27
more	27
wrong	26

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

Refresh New Subset Delete

Search Results

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

Sentence	Author ID	Tags	Type	TotalVotes	Forum	Initiator	Num Respons	Num Answers	Num Comme
o awseome - someone else liked my README - ty Jay	0100	quiz0	answerco...	1	A	0203	27	7	20
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
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
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
IMHO, such information is more appropriately placed in the help FAQ for the edX site.	0203	quiz0	answerco...	1				7	20
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				7	20
@Mr-Jonze Please read post with information from my direct e-mail correspondence with edX.	0203	quiz0	answerco...	1				7	20
as a rule we should try not to reveal the answers to assignments but as this particular quiz has unlimited	0100	quiz0	answer	1				7	20

{w} {appeal for help}

someone
anyone
help
please
anybody
appreciate
somebody

Delete this word set

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

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

group by stem:

	16
	12
best	10
more	9
such	8
own	8

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.

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