University of Konstanz
Data Analysis and Visualization Group

NLP and Visualization for Digital Humanities II

Guest Lecture: Dr. Christopher Collins
University of Ontario Institute of Technology
http://vialab.ca

Outline

- Review
- Tools for Supporting Specific DH Analysis Tasks
- Case Study: Metatation
- Case Study: VisArgue

Difficulties Visualizing Text Data

- *Not* pre-attentive
  - Must foveate to read
- Abstract
  - Difficult to visualize
- Very high dimensionality
  - Tens to hundreds of thousands of features
- Compositional
  - Can be combined together in innumerable ways
- Subtle
  - Small differences matter

Why text is (deceptively) easy

- Text is easier when you have a lot of it
  - Web search is now usually conjunction
- Text has a lot of redundancy
  - An algorithm can:
    - Pull out "important" phrases
    - Find "meaningfully" related words
    - Create a "summary" from document
    - Group "related" documents

Stages of Document Visualization

Feature Extraction → Analysis → Visualization

Feature Extraction

- Statistical Features
- Structural Features
- Syntactic Features
- …

Analysis

- Feature engineering for higher-level features
- Clustering
- Classification
- Pattern detection
- …

Visualization

- Display of features
- Project high dimensional feature vector into 2D space
- Text visualization
- …

Text Visualization Examples

… with a Digital Humanities Bias

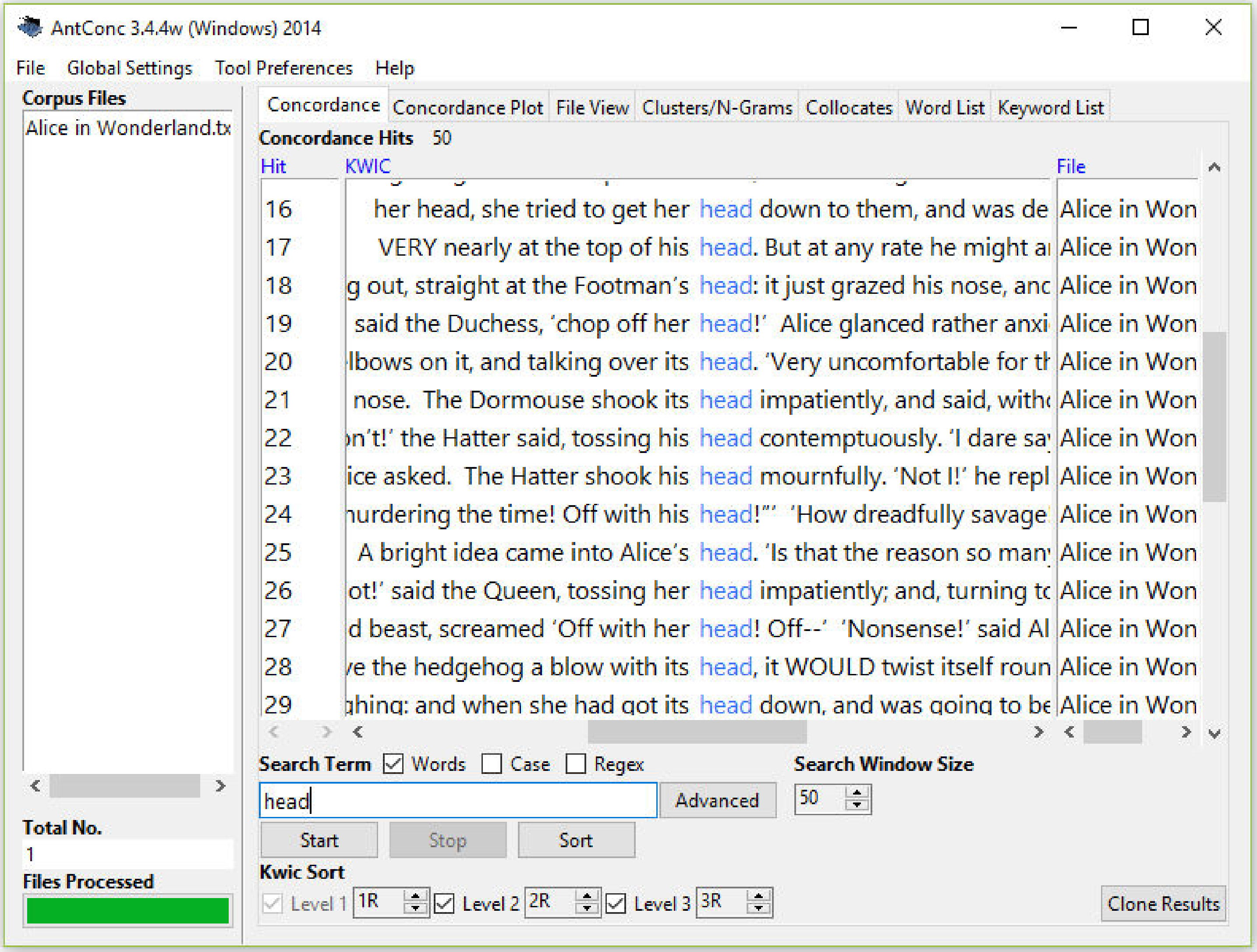
Tools – Lexical Data

- AntConc
- Wordandphrase.info
- Word Wanderer
- Text Arc
- Arc Diagram
- Word Tree
- Phrase Net

Word counts, concordances

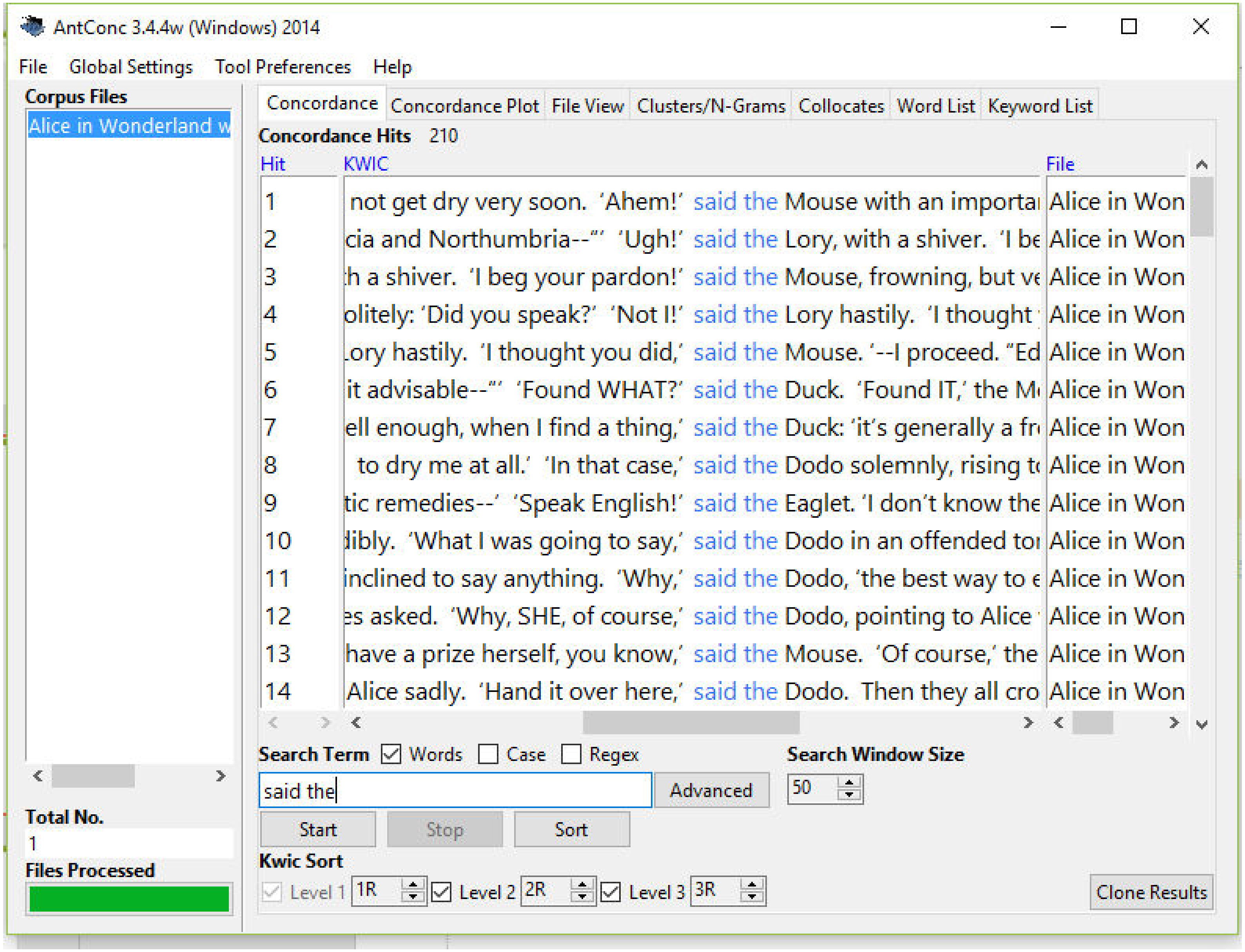
Word relationships, Structure

Concordance Lines (KWIC)

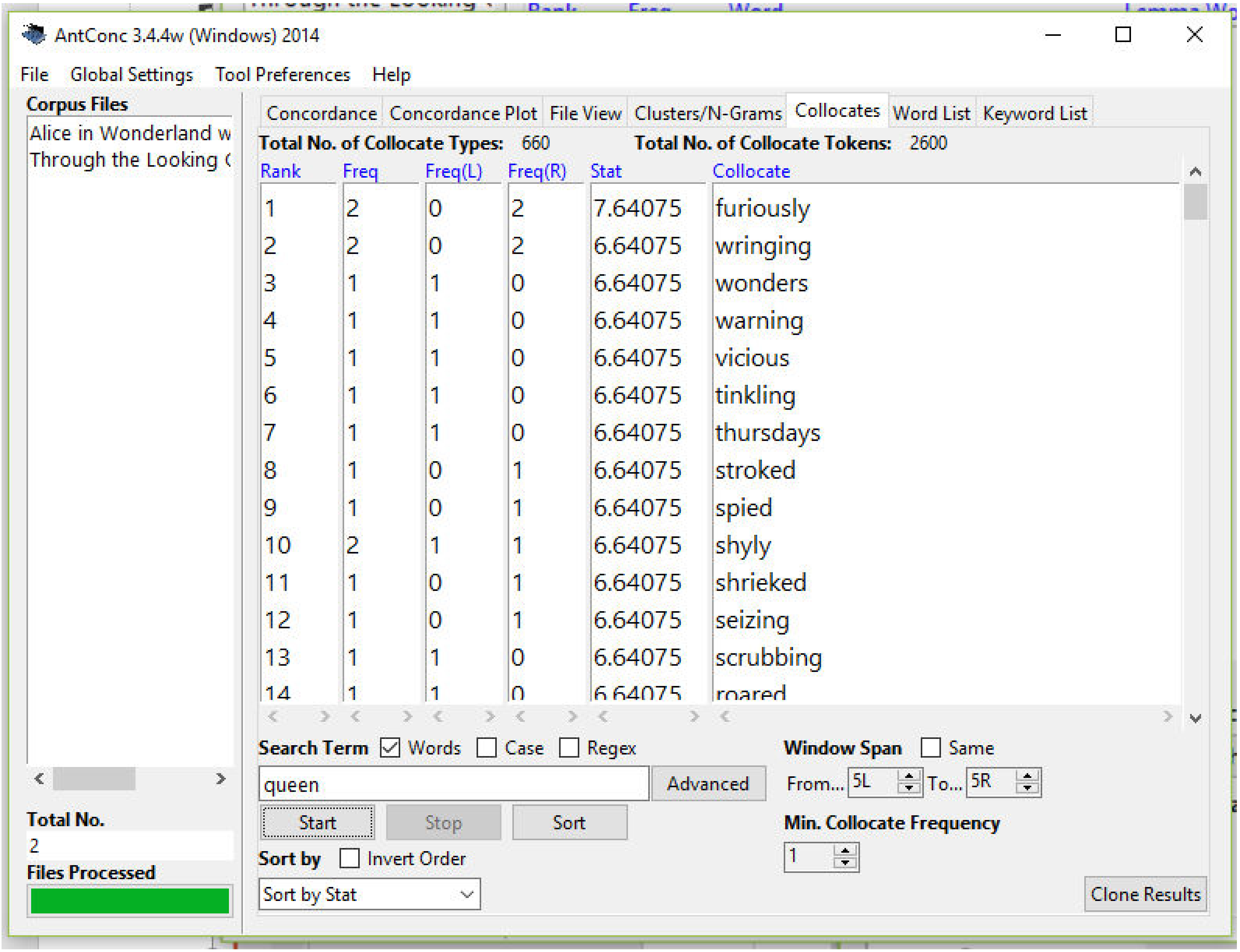


Keyword Plots

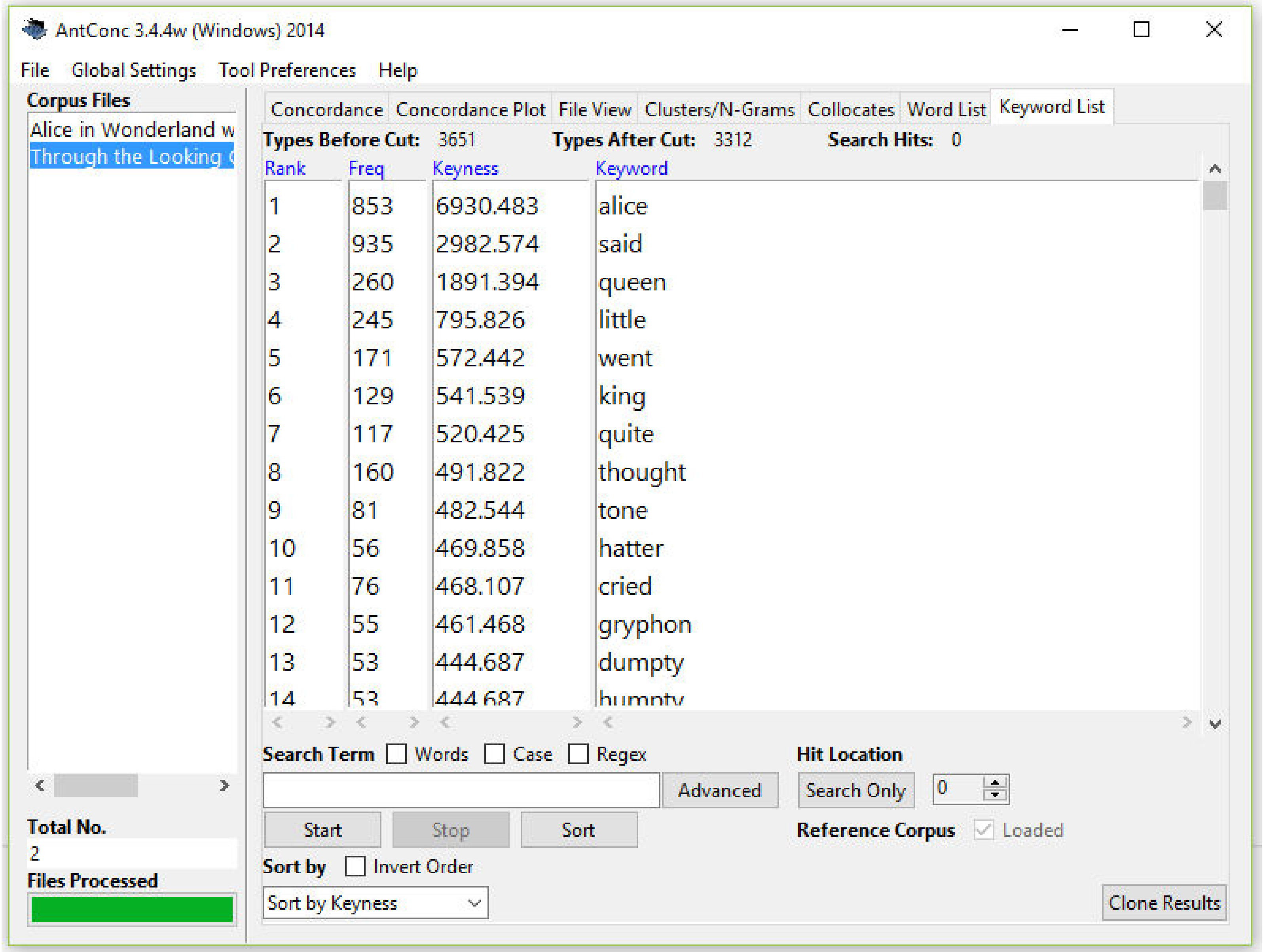
N-grams


AntConc 3.4.4w (Windows) 2014
File Global Settings Tool Preferences Help
Corpus Files
Alice in Wonderland w
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Concordance Hits 210
Hit KWIC File
1 not get dry very soon. 'Ahem!' said the Mouse with an importa Alice in Won
2 cia and Northumbria--"' 'Ugh!' said the Lory, with a shiver. 'I be Alice in Won
3 h a shiver. 'I beg your pardon!' said the Mouse, frowning, but ve Alice in Won
4 olitely: 'Did you speak?' 'Not I!' said the Lory hastily. 'I thought Alice in Won
5 ory hastily. 'I thought you did,' said the Mouse. '--I proceed. "Ed Alice in Won
6 it advisable--"' 'Found WHAT?' said the Duck. 'Found IT,' the M Alice in Won
7 ell enough, when I find a thing,' said the Duck: 'it's generally a fr Alice in Won
8 to dry me at all.' 'In that case,' said the Dodo solemnly, rising t Alice in Won
9 tic remedies--' 'Speak English!' said the Eaglet. 'I don't know the Alice in Won
10 dibly. 'What I was going to say,' said the Dodo in an offended to Alice in Won
11 inclined to say anything. 'Why,' said the Dodo, 'the best way to e Alice in Won
12 es asked. 'Why, SHE, of course,' said the Dodo, pointing to Alice Alice in Won
13 have a prize herself, you know,' said the Mouse. 'Of course,' the Alice in Won
14 Alice sadly. 'Hand it over here,' said the Dodo. Then they all cro Alice in Won
Search Term Words Case Regex
said the
Advanced
Search Window Size
50
Start Stop Sort
Kwic Sort
Level 1 1R Level 2 2R Level 3 3R
Clone Results
Total No.
1
Files Processed

Collocates

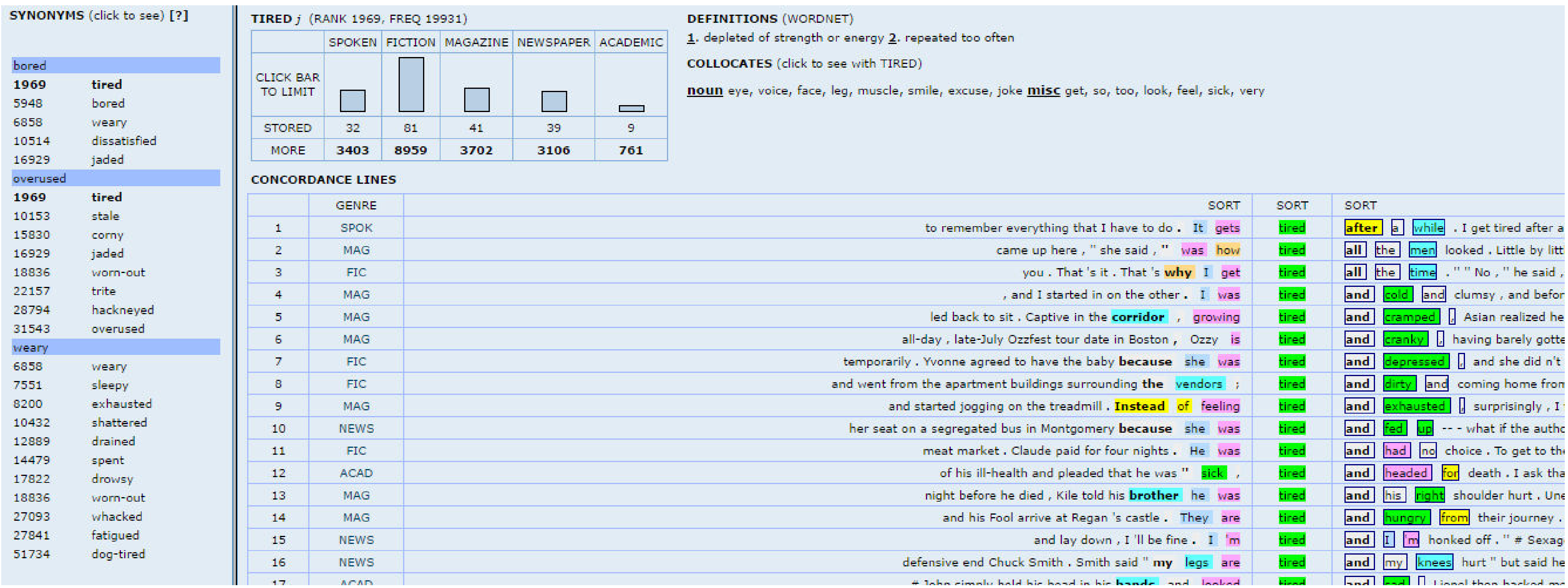

AntConc 3.4.4w (Windows) 2014
File Global Settings Tool Preferences Help
Corpus Files
Alice in Wonderland w
Through the Looking (
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Total No. of Collocate Types: 660
Total No. of Collocate Tokens: 2600
Rank Freq Freq(L) Freq(R) Stat Collocate
1 2 0 2 7.64075 furiously
2 2 0 2 6.64075 wringing
3 1 1 0 6.64075 wonders
4 1 1 0 6.64075 warning
5 1 1 0 6.64075 vicious
6 1 1 0 6.64075 tinkling
7 1 1 0 6.64075 thursdays
8 1 0 1 6.64075 stroked
9 1 0 1 6.64075 spied
10 2 1 1 6.64075 shyly
11 1 0 1 6.64075 shrieked
12 1 0 1 6.64075 seizing
13 1 1 0 6.64075 scrubbing
14 1 1 0 6.64075 roared
Search Term Words Case Regex
queen
Advanced
Start Stop Sort
Sort by Invert Order
Sort by Stat
Window Span Same
From... 5L To... 5R
Min. Collocate Frequency
1
Clone Results
Total No.
2
Files Processed

Keywords

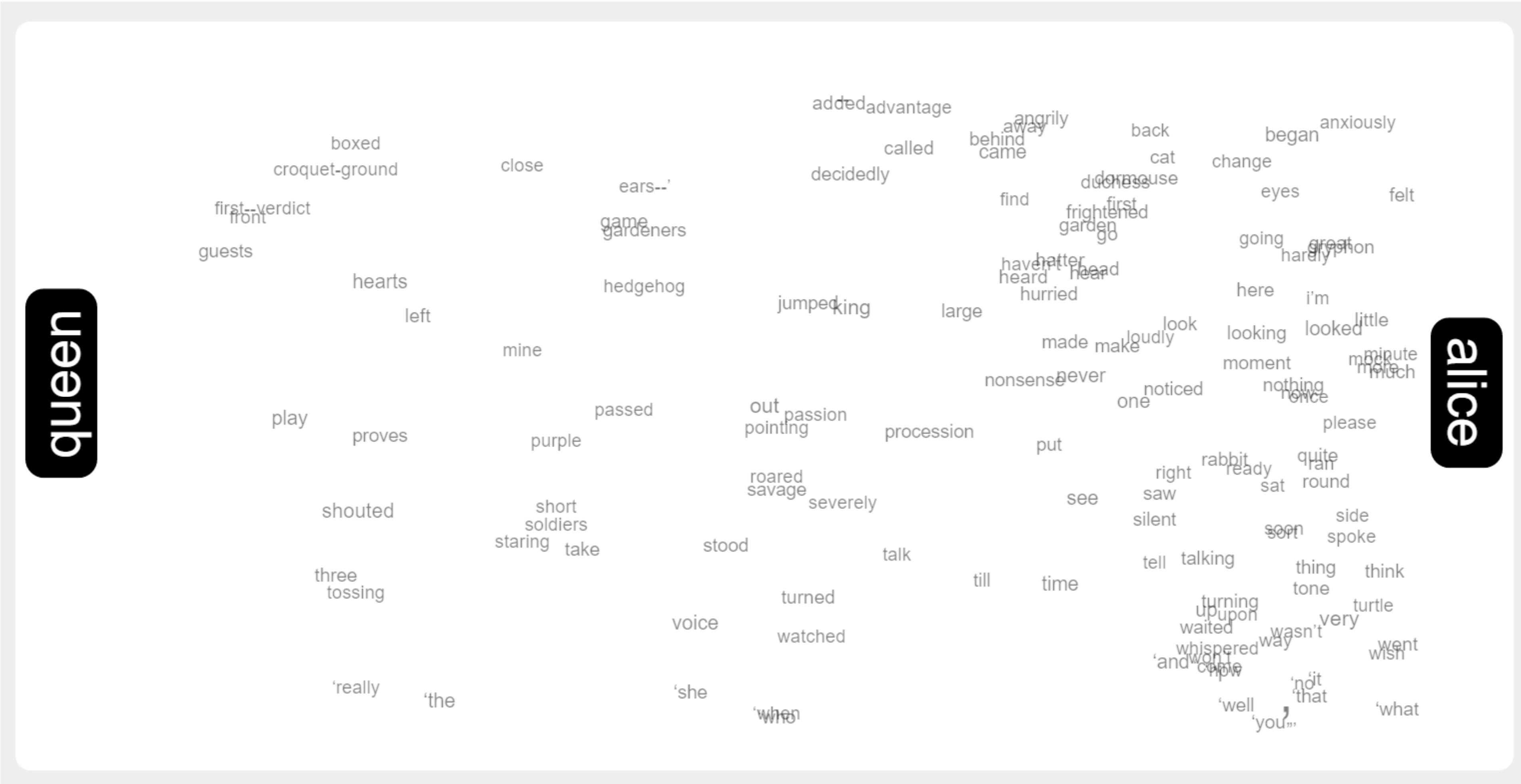

AntConc 3.4.4w (Windows) 2014
File
Global Settings
Tool Preferences
Help
Corpus Files
Alice in Wonderland w
Through the Looking G
Total No.
2
Files Processed
Concordance
Concordance Plot
File View
Clusters/N-Grams
Collocates
Word List
Keyword List
Types Before Cut: 3651
Types After Cut: 3312
Search Hits: 0
Rank Freq Keyness Keyword
1 853 6930.483 alice
2 935 2982.574 said
3 260 1891.394 queen
4 245 795.826 little
5 171 572.442 went
6 129 541.539 king
7 117 520.425 quite
8 160 491.822 thought
9 81 482.544 tone
10 56 469.858 hatter
11 76 468.107 cried
12 55 461.468 gryphon
13 53 444.687 dumpty
Search Term
Words
Case
Regex
Advanced
Start
Stop
Sort
Sort by
Invert Order
Sort by Keyness
Hit Location
Search Only
0
Reference Corpus
Loaded
Clone Results

Investigate Words in Context

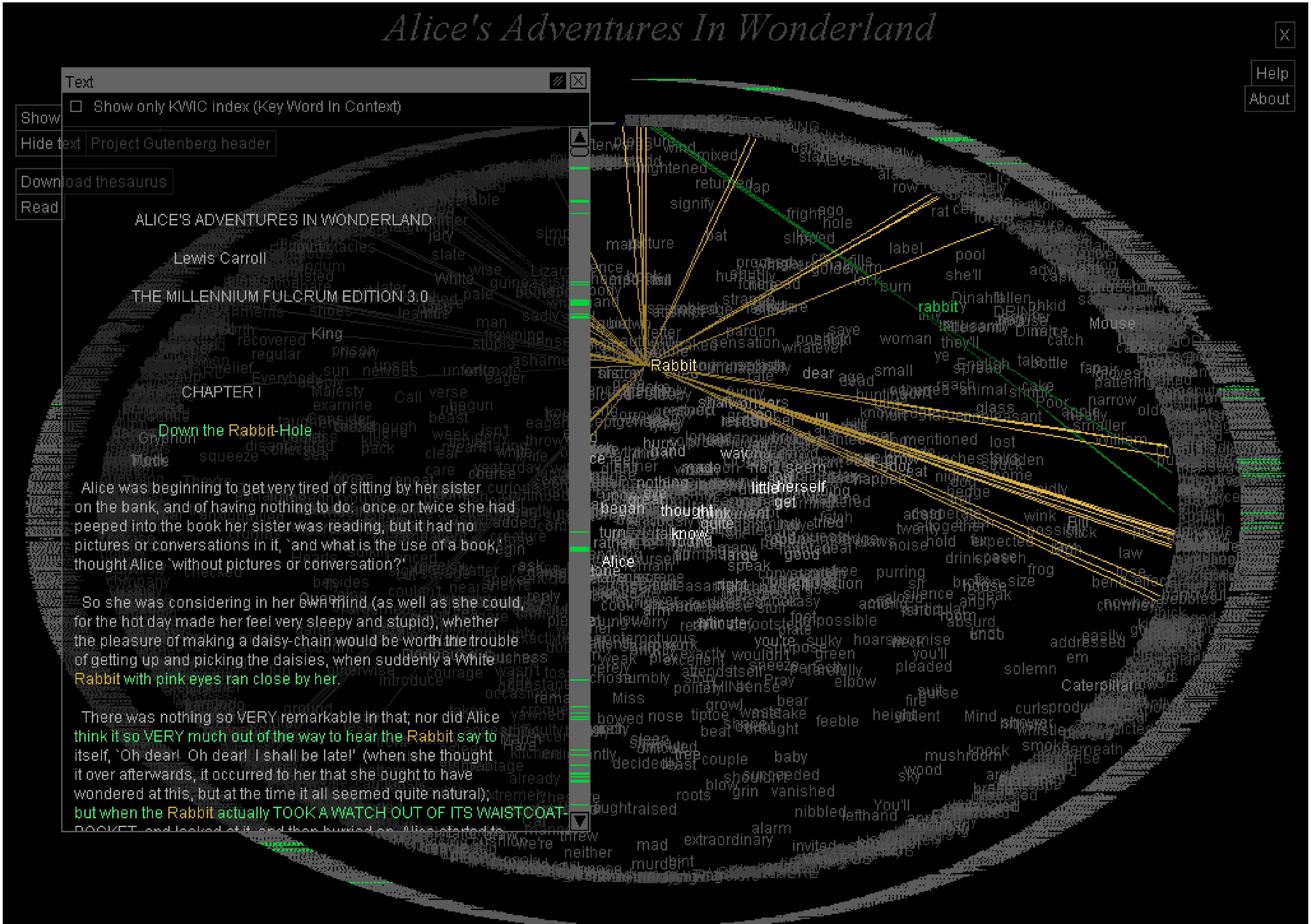
- External context: From Corpus of Contemporary American English



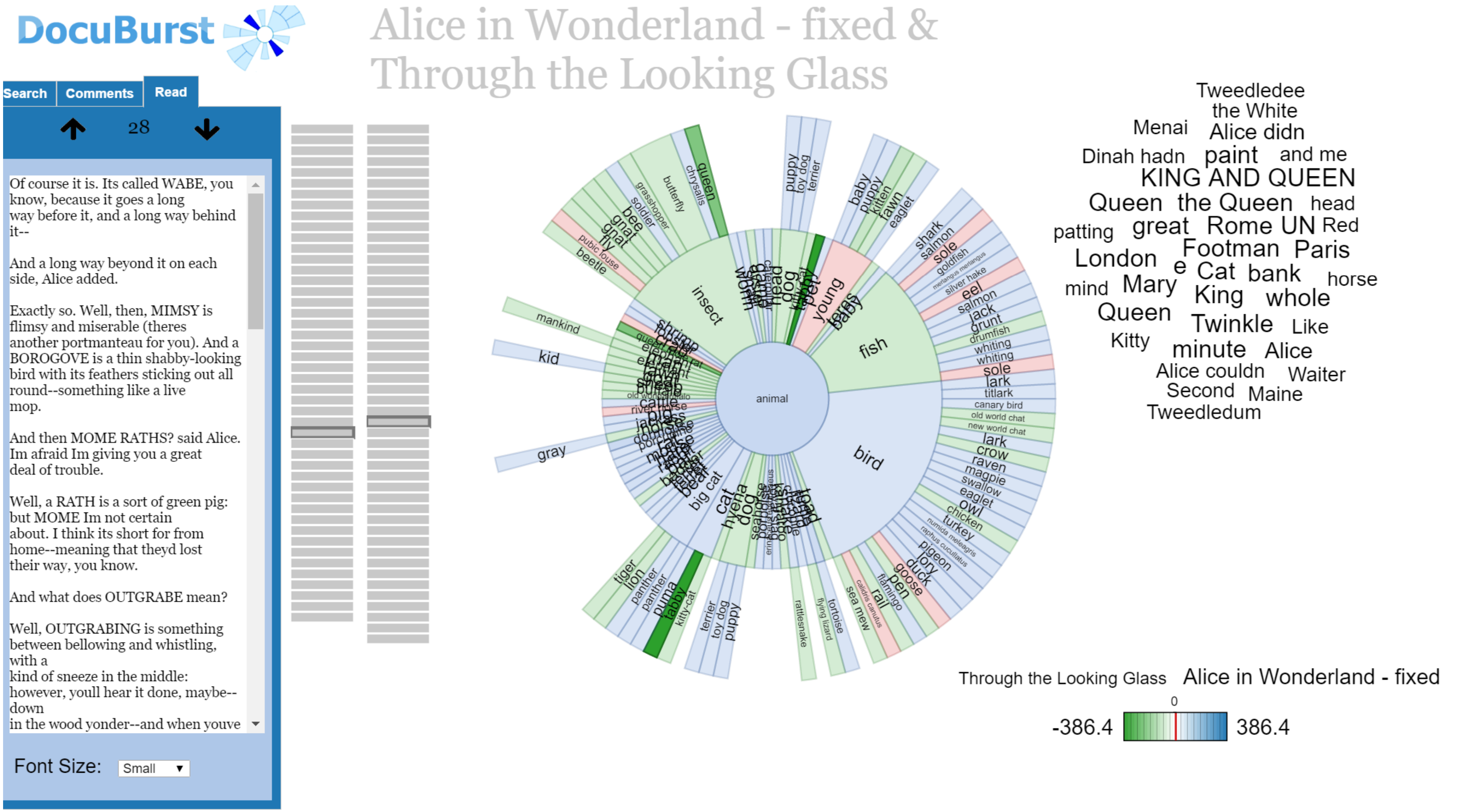
Word Wanderer



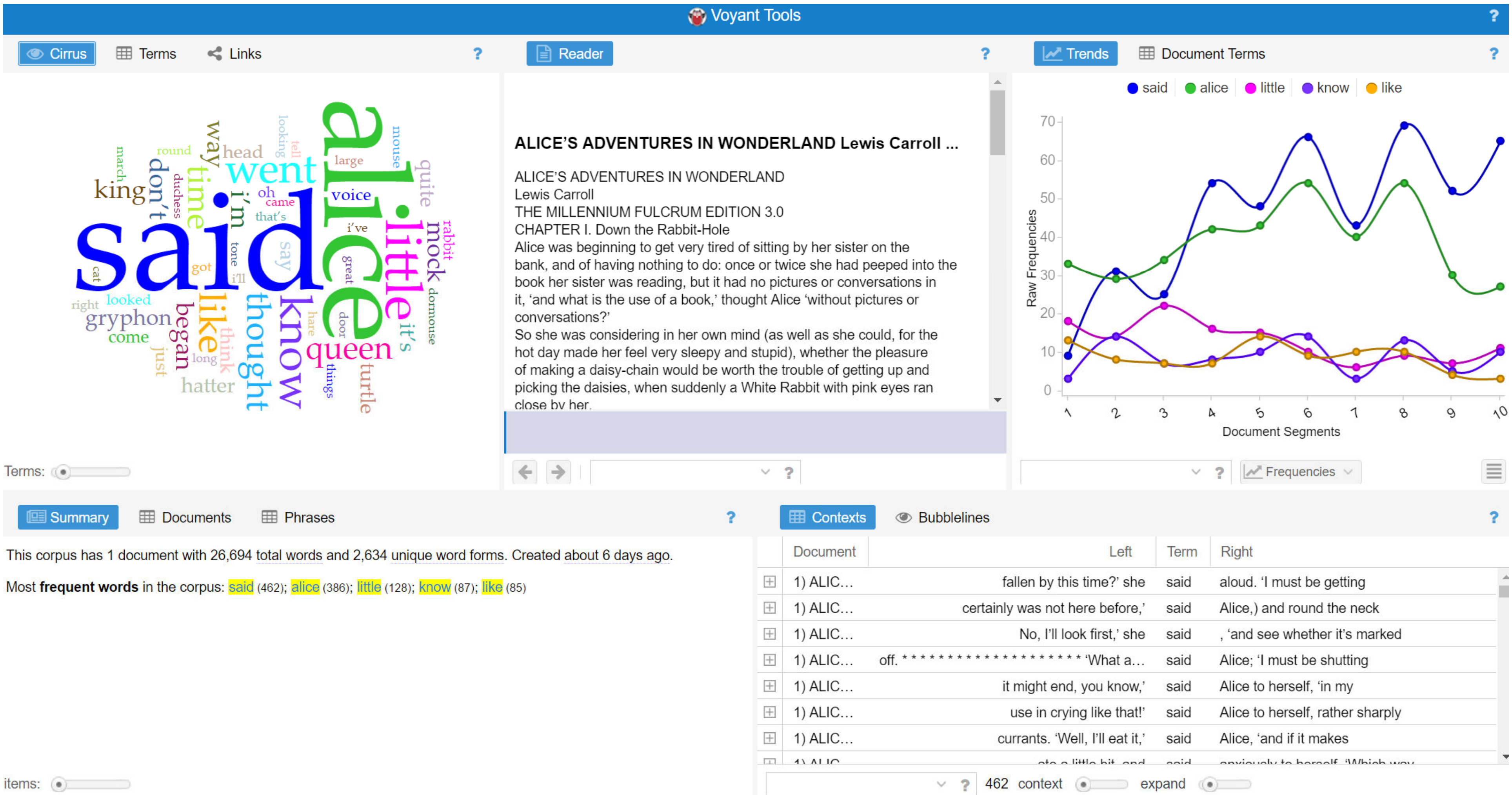
By Marian Dörk and Dawn Knight; http://wordwanderer.org/



W. Bradford Paley, 2002 , http://www.textarc.org/


DocuBurst
Search
Comments
Read
28
Alice in Wonderland - fixed & Through the Looking Glass
Of course it is. Its called WABE, you know, because it goes a long way before it, and a long way behind it--
And a long way beyond it on each side, Alice added.
Exactly so. Well, then, MIMSY is flimsy and miserable (theres another portmanteau for you). And a BOROGOVE is a thin shabby-looking bird with its feathers sticking out all round--something like a live mop.
And then MOME RATHS? said Alice. Im afraid Im giving you a great deal of trouble.
Well, a RATH is a sort of green pig: but MOME Im not certain about. I think its short for from home--meaning that theyd lost their way, you know.
And what does OUTGRABE mean?
Well, OUTGRABING is something between bellowing and whistling, with a kind of sneeze in the middle: however, youll hear it done, maybe--down in the wood yonder--and when youve
Font Size:
Small
animal
insect
fish
bird
mankind
kid
gray
Tweedledee
the White
Menai
Alice didn
Dinah hadn
paint
and me
KING AND QUEEN
Queen
the Queen
head
patting
great
Rome
UN
Red
London
e
Footman
Paris
Cat
bank
mind
Mary
horse
King
whole
Queen
Twinkle
Like
Kitty
minute
Alice
Alice couldn
Waiter
Second
Maine
Tweedledum
Through the Looking Glass
Alice in Wonderland - fixed
0
-386.4
386.4

Voyant Tools

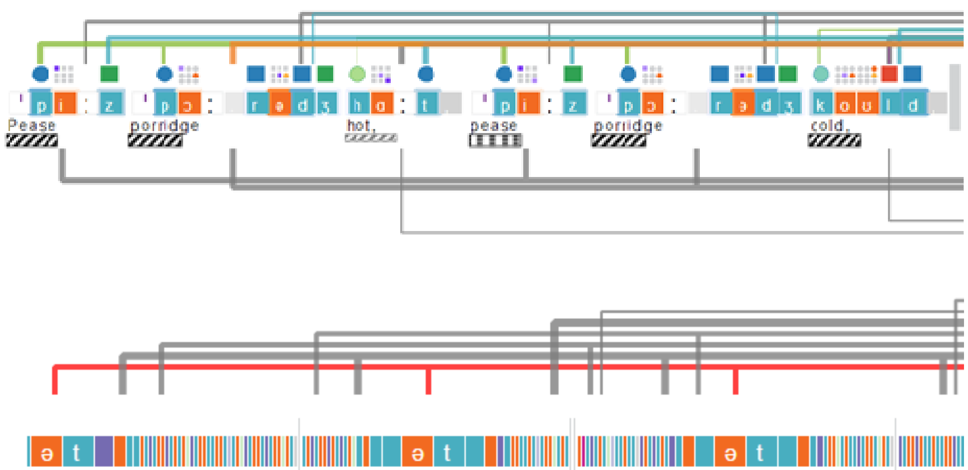


Visualizing Other Document Features

- Sonic Patterns
- Sentence Complexity / Readability
- Sentiment
- Edits and Variations

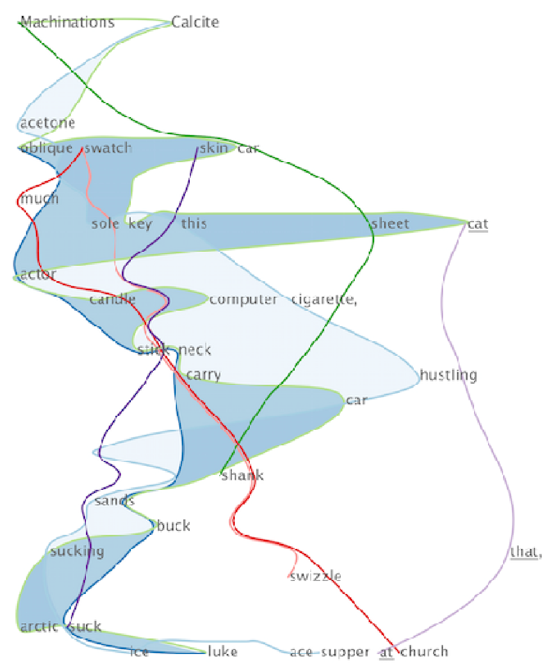
Visualizing Sonic Patterns in Text

Poemviewer



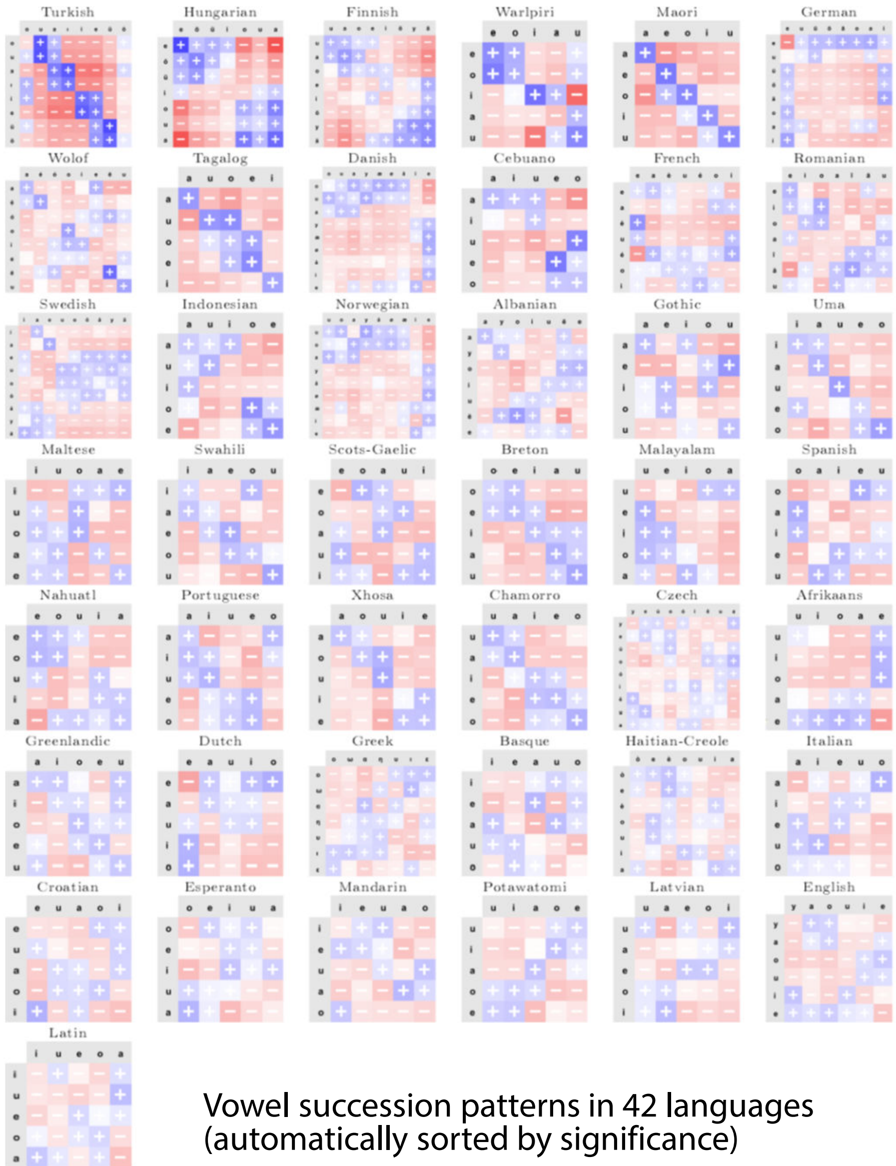
Rule-based Visual Mappings - with a Case Study on Poetry Visualization. A. Abdul-Rahman, J. Lein, K. Coles, E. Maguire, M. Meyer, M. Wynne, C. R. Johnson, A. Trefethen, and M. Chen. In *Computer Graphics Forum*, 32(3):381-390, 2013.

Poemage

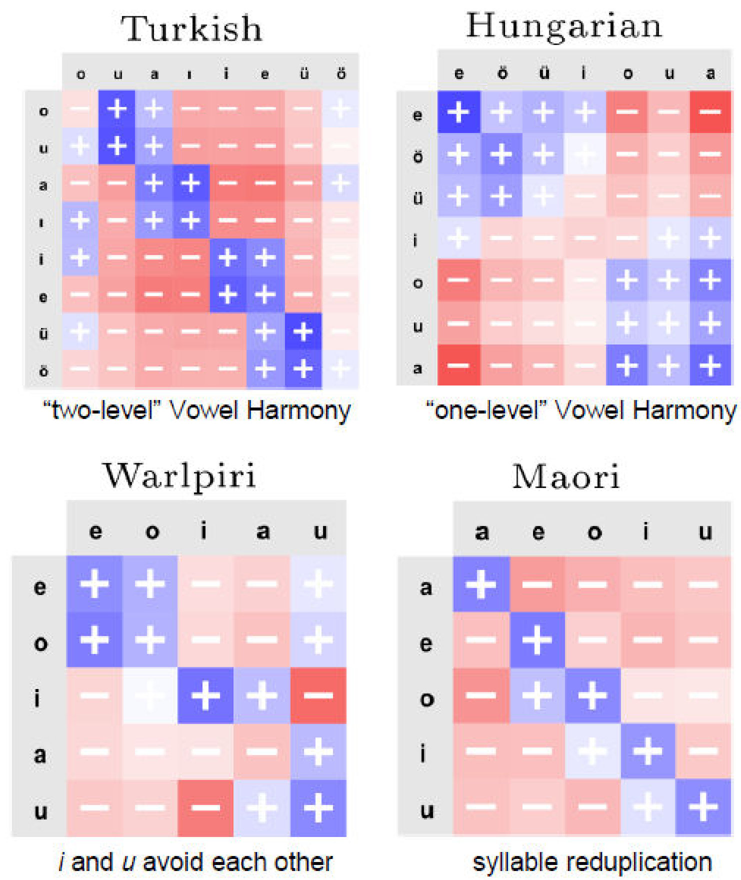


N. McCurdy, J. Lein, K. Coles, M. Meyer. Poemage: Visualizing the Sonic Topology of a Poem. *IEEE TVCG*, pages 439-448, January 2016.

Vowel Harmony: Cross-linguistic Comparison of Complex Language Features

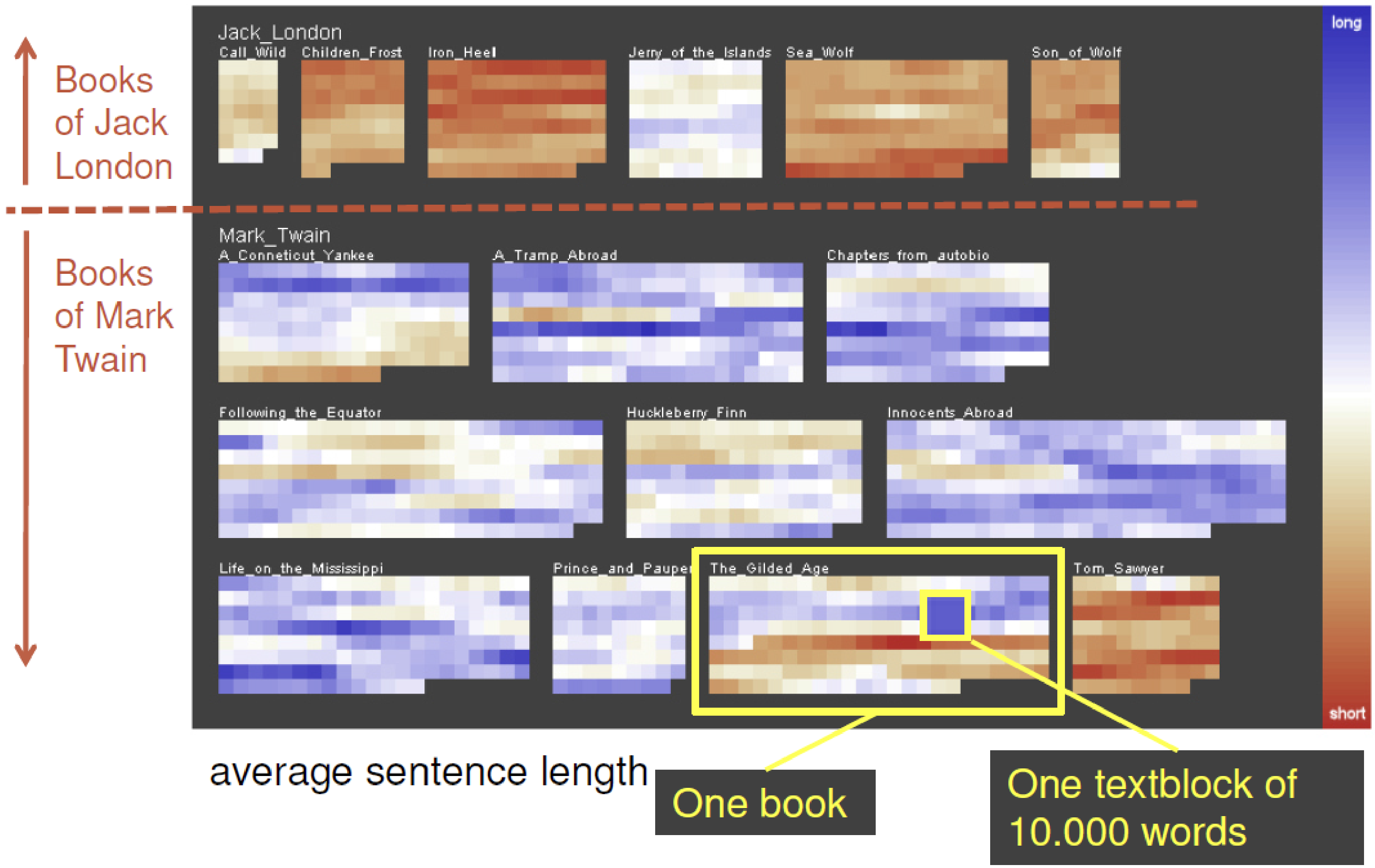


Vowel succession patterns in 42 languages
(automatically sorted by significance)



Comparative visual analysis of cross-linguistic features. Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010).

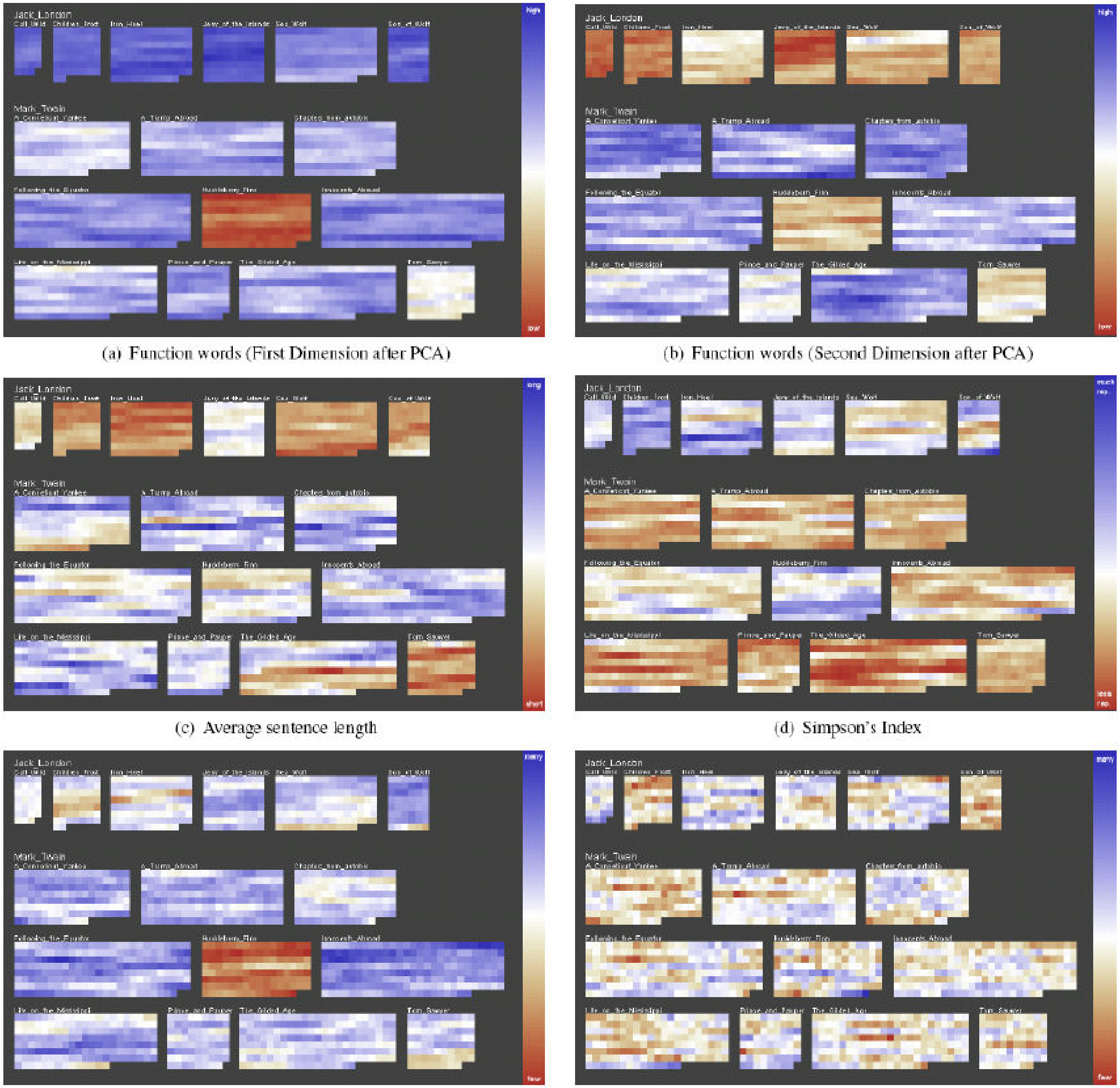
Literature Fingerprinting



Daniel A. Keim and Daniela Oelke. 2007. Literature Fingerprinting: A New Method for Visual Literary Analysis. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology* (VAST '07). IEEE Computer Society, Washington, DC, USA, 115-122.

Features

- Hapax legomena: Word’s that appear exactly once
- Hapax dislegomena: appear twice
- Function words
- Simpson’s index: measure of diversity/repetition
- Sentence length
- Word length
- …



(a) Function words (First Dimension after PCA)

(b) Function words (Second Dimension after PCA)

(c) Average sentence length

(d) Simpson's Index

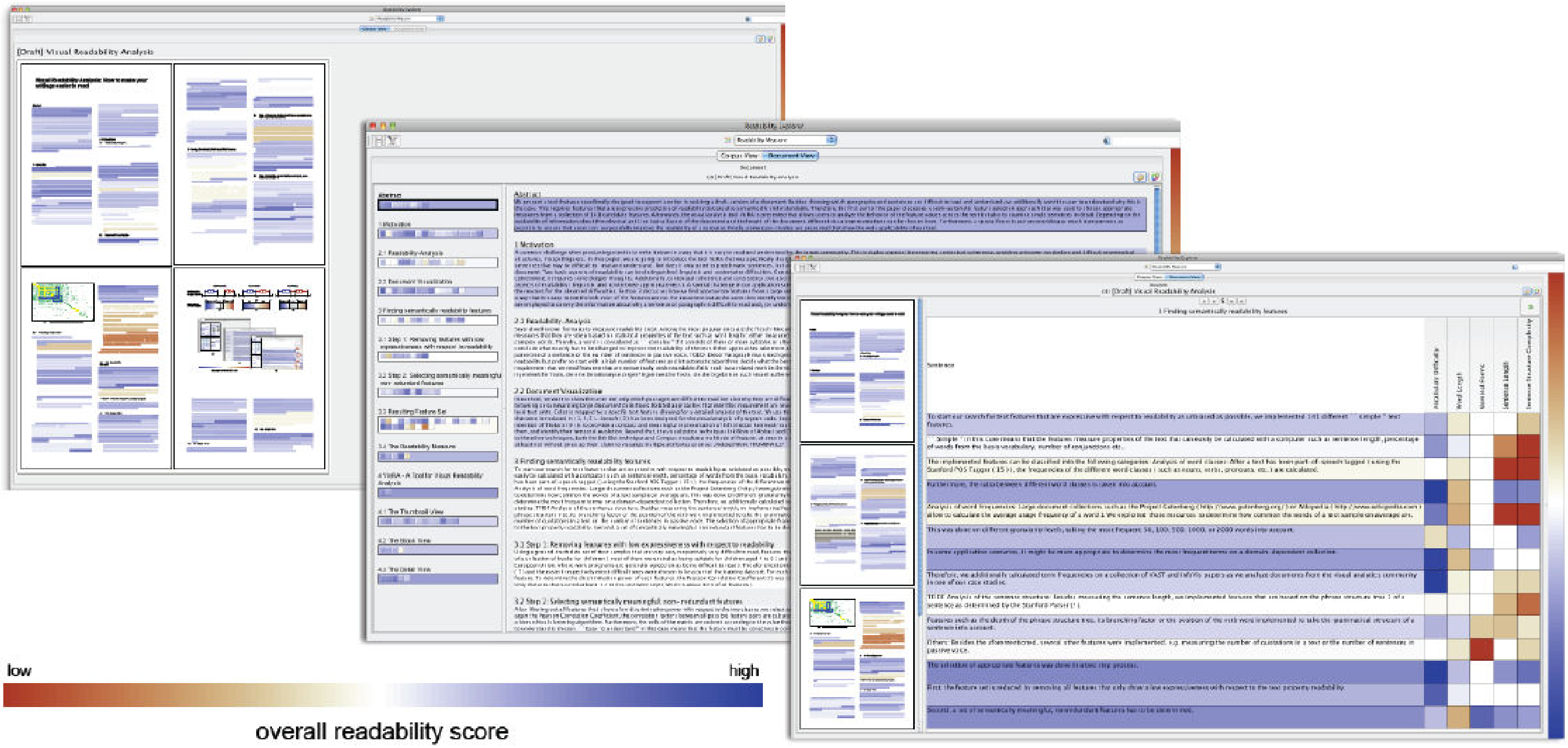
(e) Hapax Legomena

(f) Hapax Dislegomena

Different measures for authorship attribution are tested on books of Mark Twain (last three rows) and Jack London (first row).

If a feature is able to discriminate between the two authors the books in the first row visually set apart from the rest of the books.

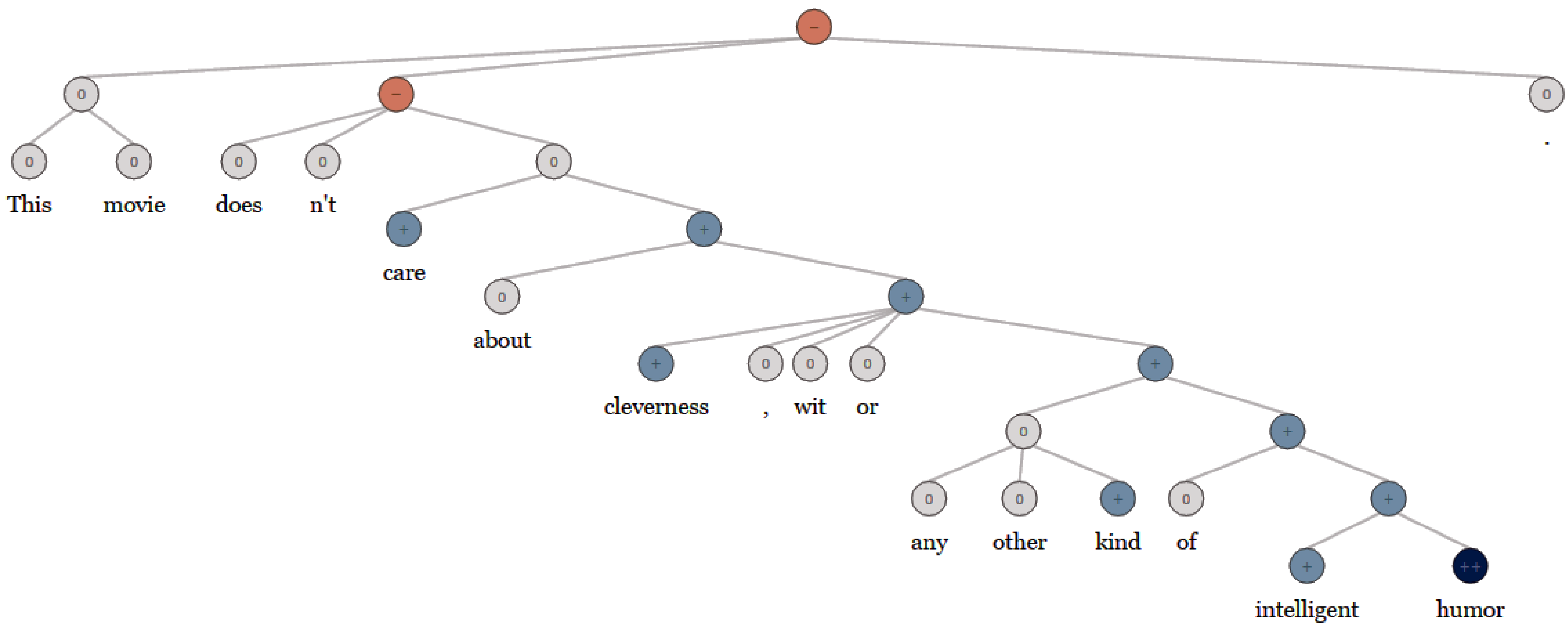
Readability Analysis



D. Oelke, D. Spretke, A. Stoffel and D. A. Keim.
Visual Readability Analysis: How to Make Your Writings Easier to Read.
IEEE Transactions on Visualization and Computer Graphics, 18(5):662-674, 2012.

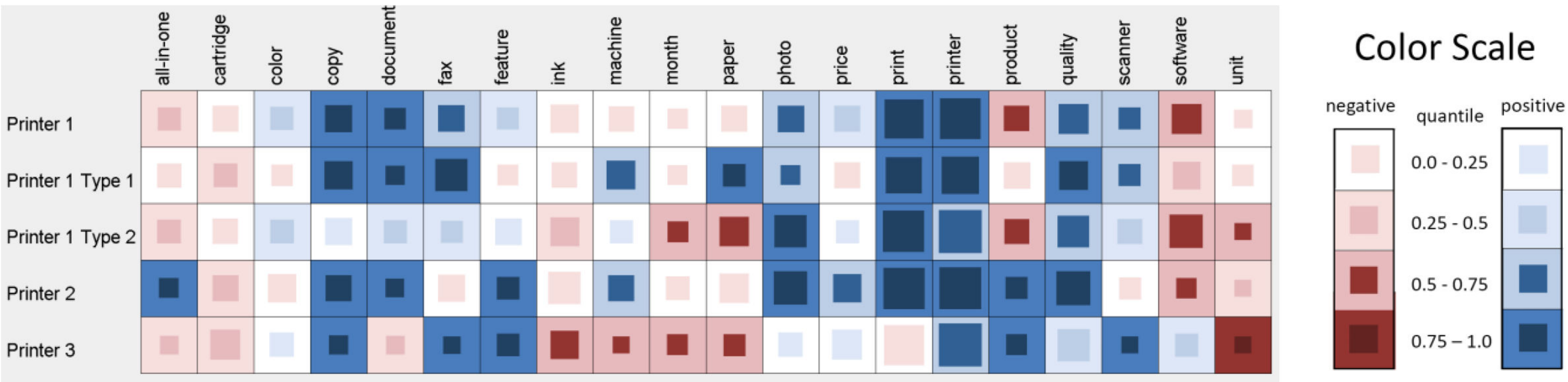
Sentiment Analysis

- Dictionary approaches (emotion word lists)
- Parsing approaches (addresses negation)



http://nlp.stanford.edu/sentiment/

Sentiment Analysis

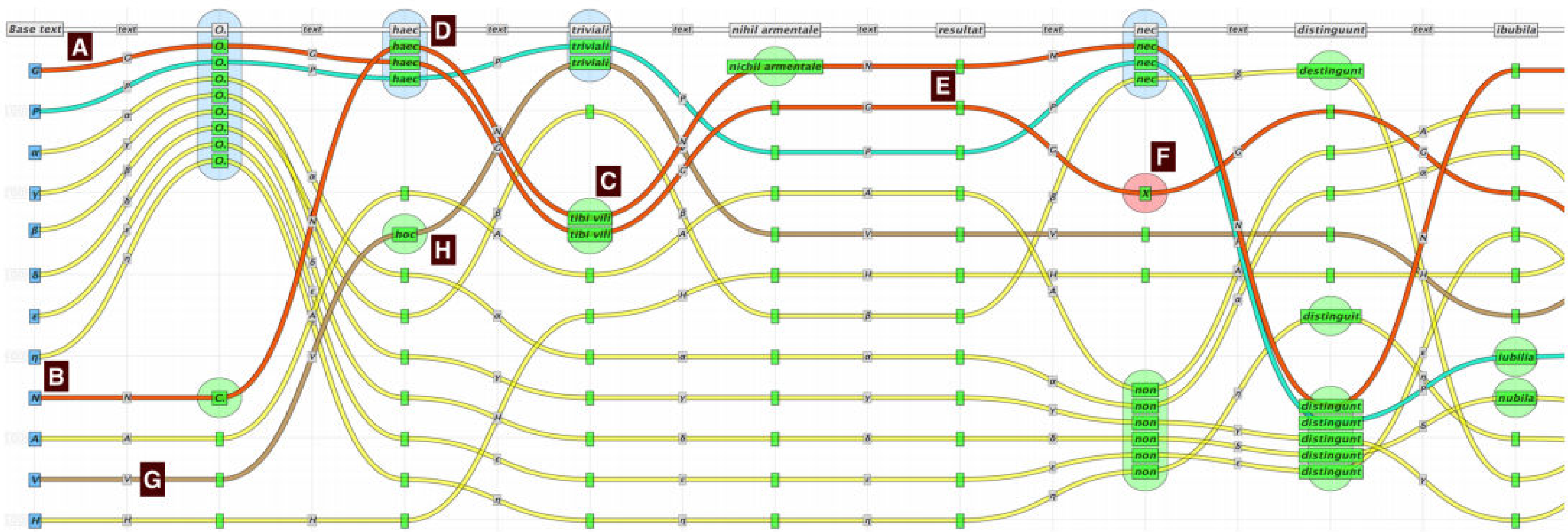


Larger inside square = more comments

D. Oelke, M. C. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L.-E. Haug and H. Janetzko.
Visual Opinion Analysis of Customer Feedback Data. *Proc. IEEE VAST*, pages 187-194, 2009.

Document Variations

- Variations in *editions* or edits of documents
- Used in historical analysis



Shejuti Silvia, Ronak Etemadpour, June Abbas, Sam Huskey, and Chris Weaver. "Visualizing Variation in Classical Text with Force Directed Storylines". *Proceedings of the Workshop on Visualization for the Digital Humanities*, Baltimore, MD, October 2016.

Tools - Document Collections

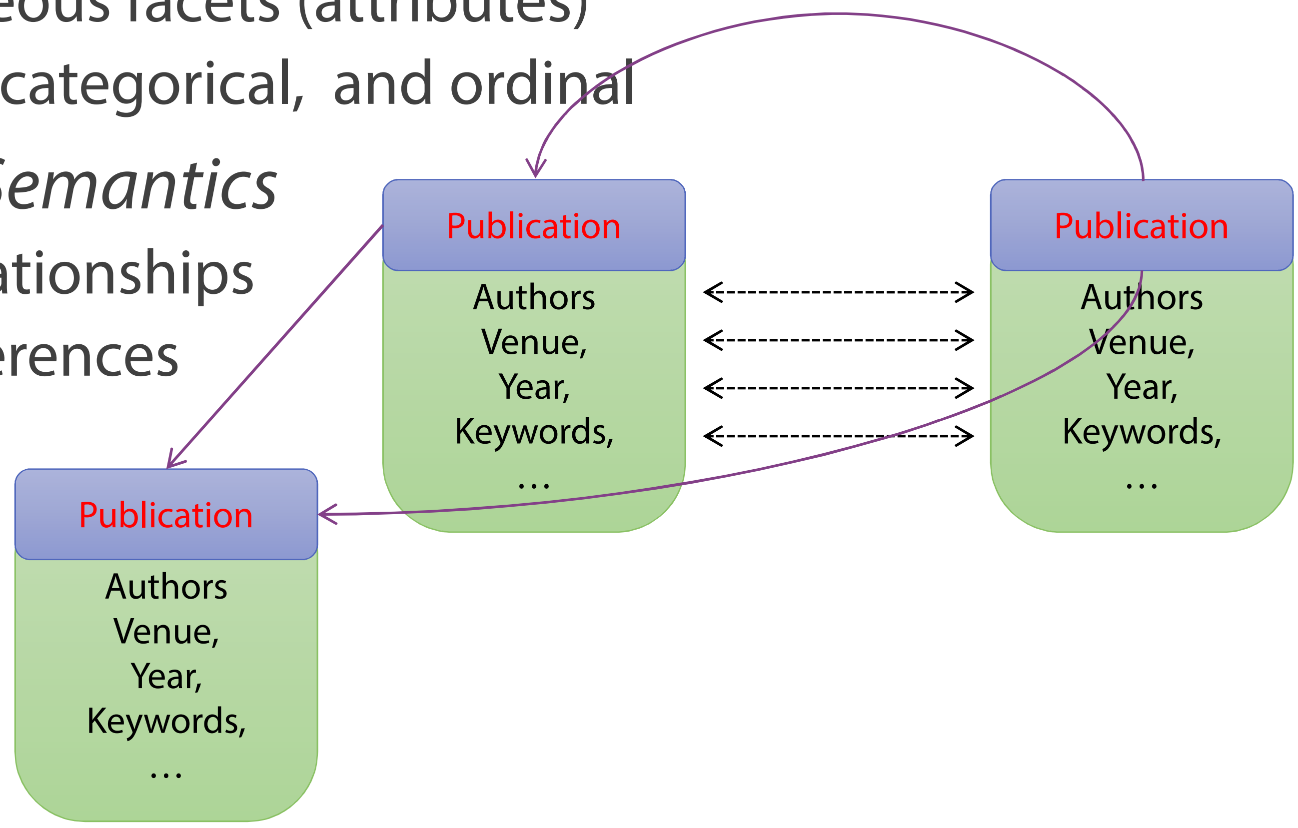
- Visualizing Document Collections
  - Citations and Document Connections
  - Clustering
    - Galaxies
    - Themescapes
  - Annotating and Curating
    - Overview Project
    - Timeline Curator
- Visualizing Corpus Trends
  - Culturomics

Bibliometric Data

- Large in *Size*
- Rich in *Structure*
  - Heterogeneous facets (attributes)
  - Numerical, categorical, and ordinal
- Complex in *Semantics*

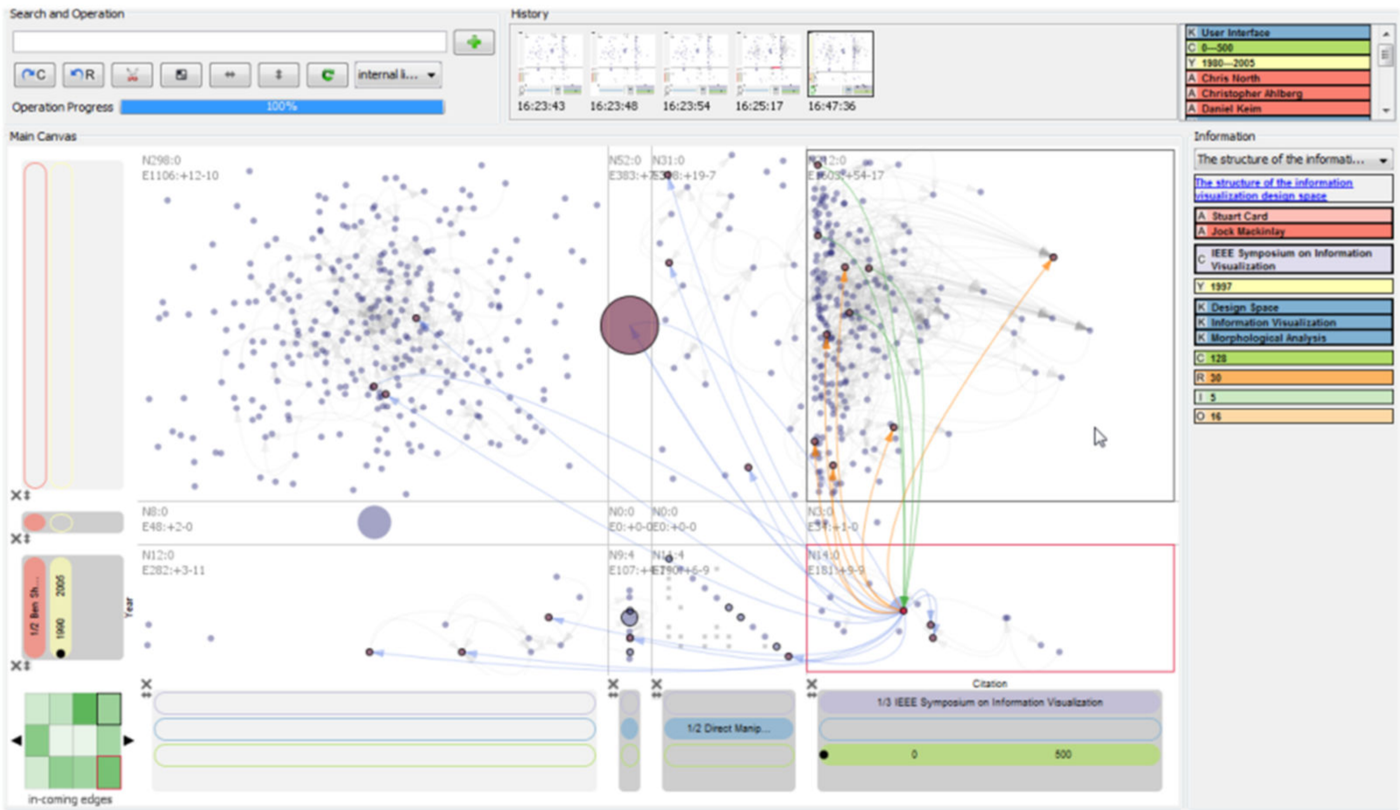
Bibliometric Data

- Large in *Size*
- Rich in *Structure*
  - Heterogeneous facets (attributes)
  - Numerical, categorical, and ordinal
- Complex in *Semantics*
  - Implicit relationships
  - Explicit references



Citation Visualizations

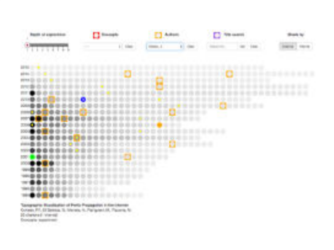
- Citations provide *explicit relations*
- Keywords, authors, etc. provide *implicit relations*



J. Zhao, C. Collins, F. Chevalier, and R. Balakrishnan, "Interactive Exploration of Implicit and Explicit Relations in Faceted Datasets," *IEEE Trans. on Visualization and Computer Graphics (Proc. of the IEEE Conf. on Visual Analytics Science and Technology (VAST))*, vol. 19, iss. 12, pp. 2080-2089, 2013.

Citation Visualizations

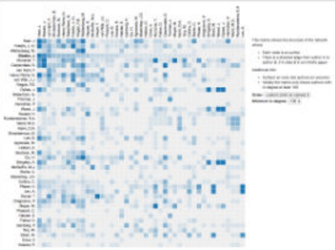
- Citation research by John Stasko's group



CiteVis

This visualization presents all the conference papers, their total numbers of citations, and citations between papers. It draws from the original CiteVis technique, but adds the notion of cascading influence through paper generations.

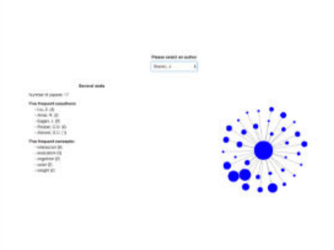
Go!



Author's matrix

This visualization presents citations from specific authors to other authors. It contains only the most referenced authors, and it supports reordering of authors on axes to help explore the data.

Go!



Author's statistics

This webpage shows statistics for specific authors including their most frequent co-authors, the most frequent concepts of their papers, and sets of other authors that cite all of authors in each set (cliques).

Go!

Co-citations network

This visualization presents a node-link network diagram, with nodes representing authors and edges between authors if they cite each other at least once.

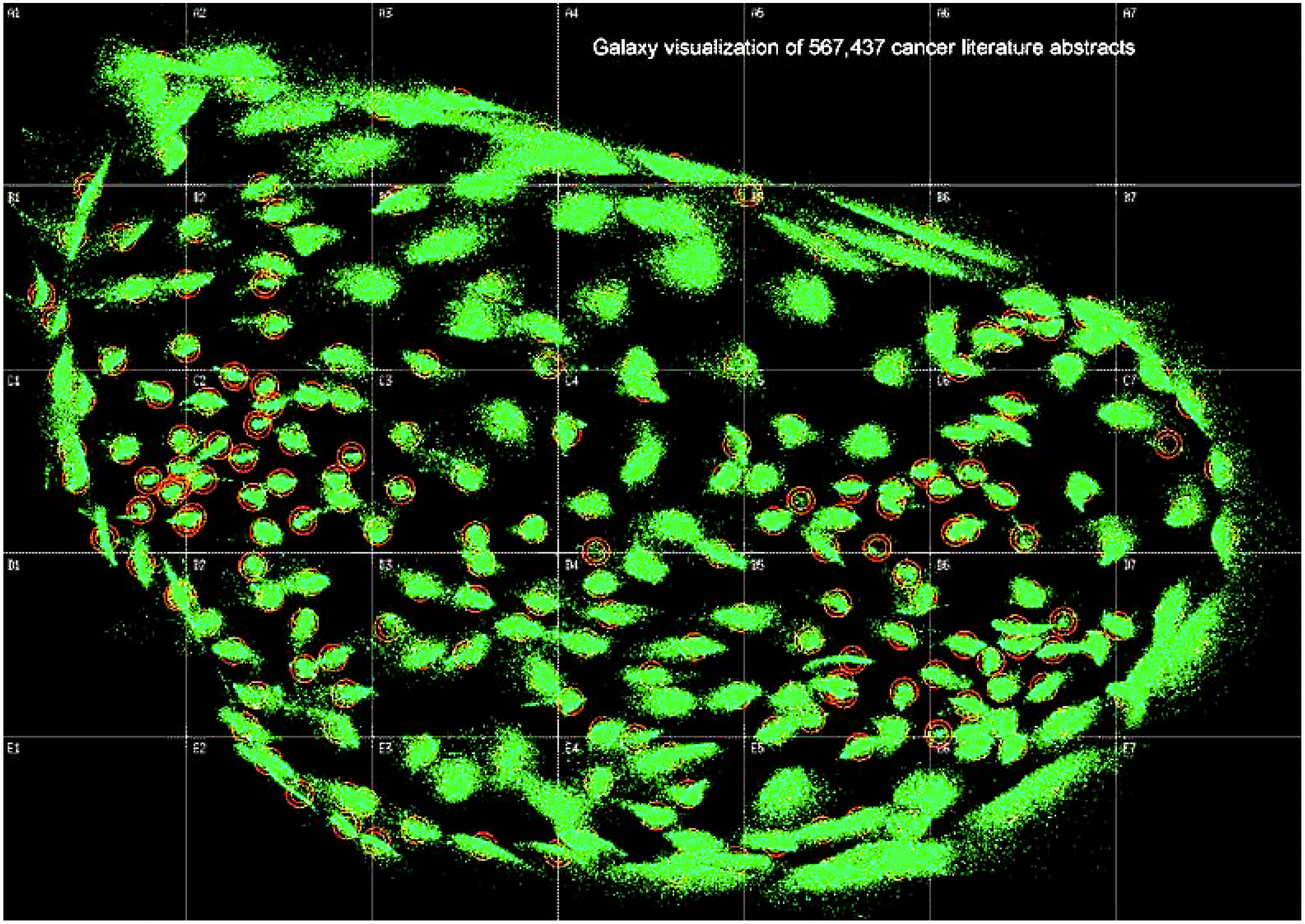
Go!

Concepts

This visualization presents common concepts (keywords) across all the titles and abstracts of papers, and how frequently each concept co-occurs with other concepts.

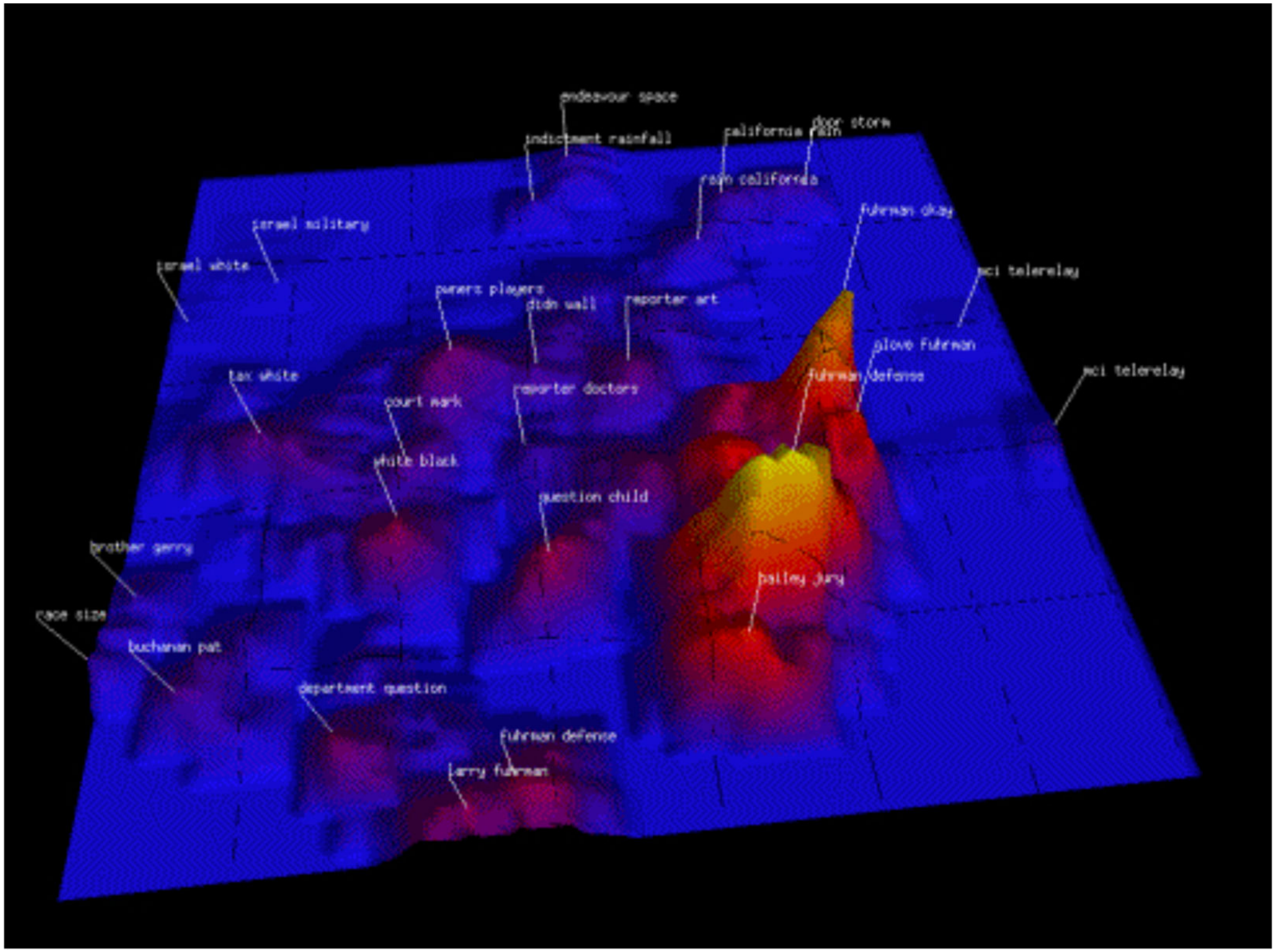
Go!

Clustering - Galaxies



James A. Wise et al. 1995. Visualizing the non-visual: spatial analysis and interaction with information for text documents. In *Proc. of the IEEE Symp. on Information Visualization*, pages 51–58.

Clustering - Themescape



James A. Wise et al. 1995. Visualizing the non-visual: spatial analysis and interaction with information for text documents. In *Proc. of the IEEE Symp. on Information Visualization*, pages 51–58.

Overview Project - Entities

Search for only these:

- [ ] Companies by Suffix
- [ ] Geonames: Cities
- [x] Geonames: Countries
  Any country or disputed country, plus variations and translations, from geonames.org
- [x] Geonames: Political Boundaries
  About 300,000 countries and administrative regions worldwide, plus variations and translations, from geonames.org
- [ ] Numbers

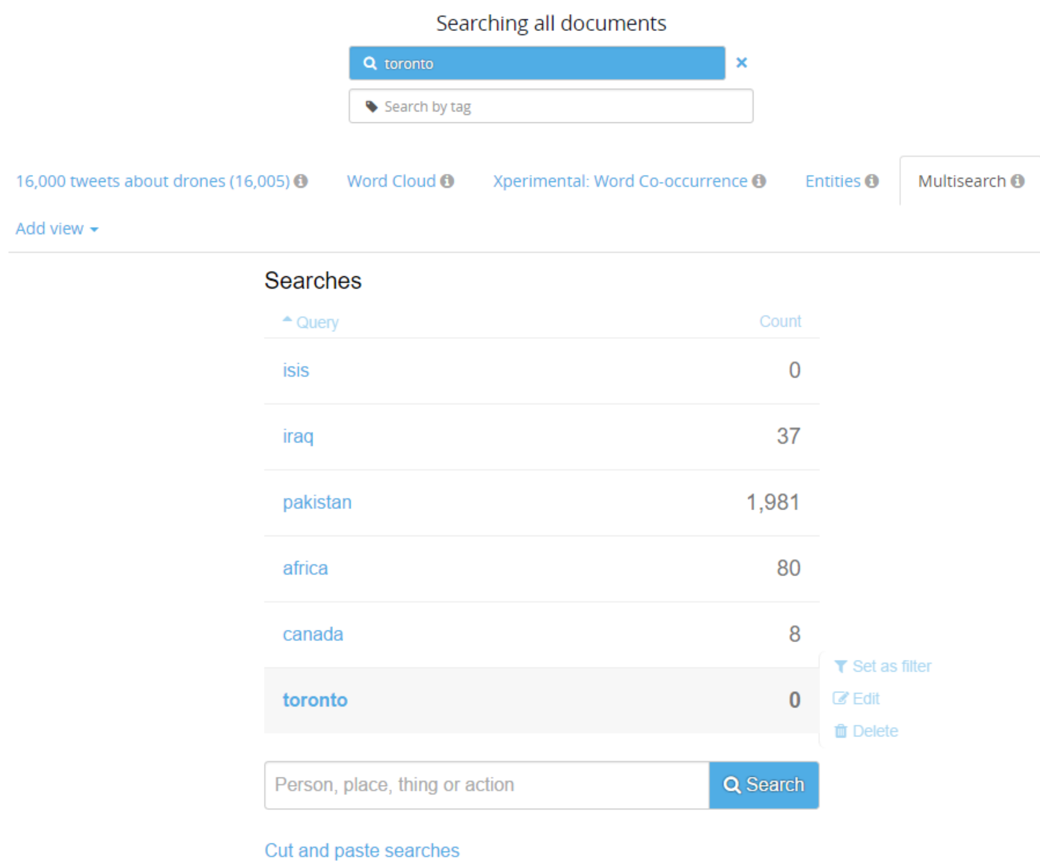
... then remove any of these:

- [x] English: Google Books words
  The most common 50,000 uncapitalized words in English books, according to Google Books (CC BY 3.0)
- [x] English: stop words
  Extremely common English words
- [x] Numbers

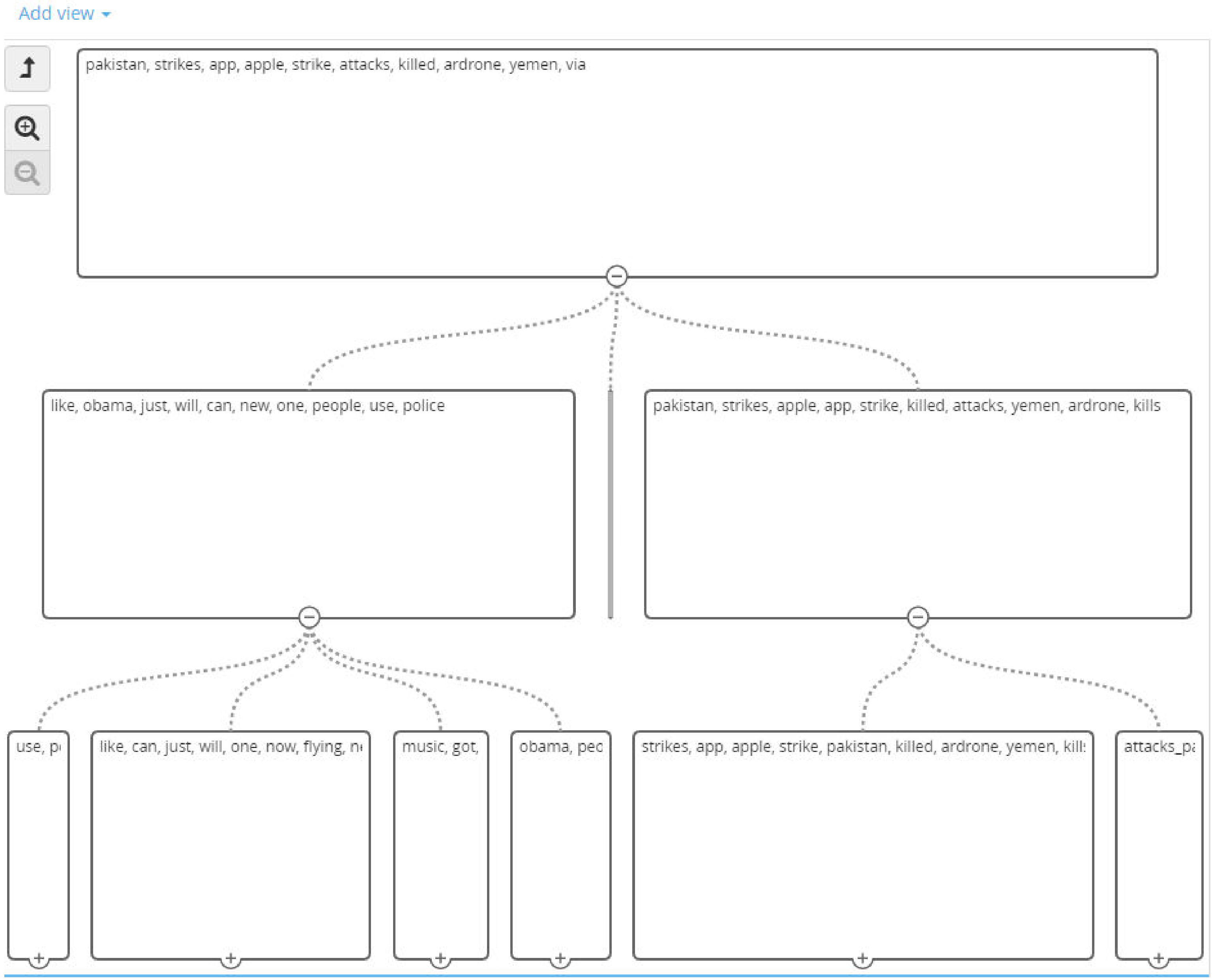
Entity	count	docs
huma Huma Xian	487	352
israel State of Israel	246	64
libya Libya	245	93
syria Syrian Arab Republic	163	39
tripoli Tripoli	145	47
israeli State of Israel	144	53
united states United States	119	75
obama Obama-machi	104	47
washington Washington County	100	58

Overview Project

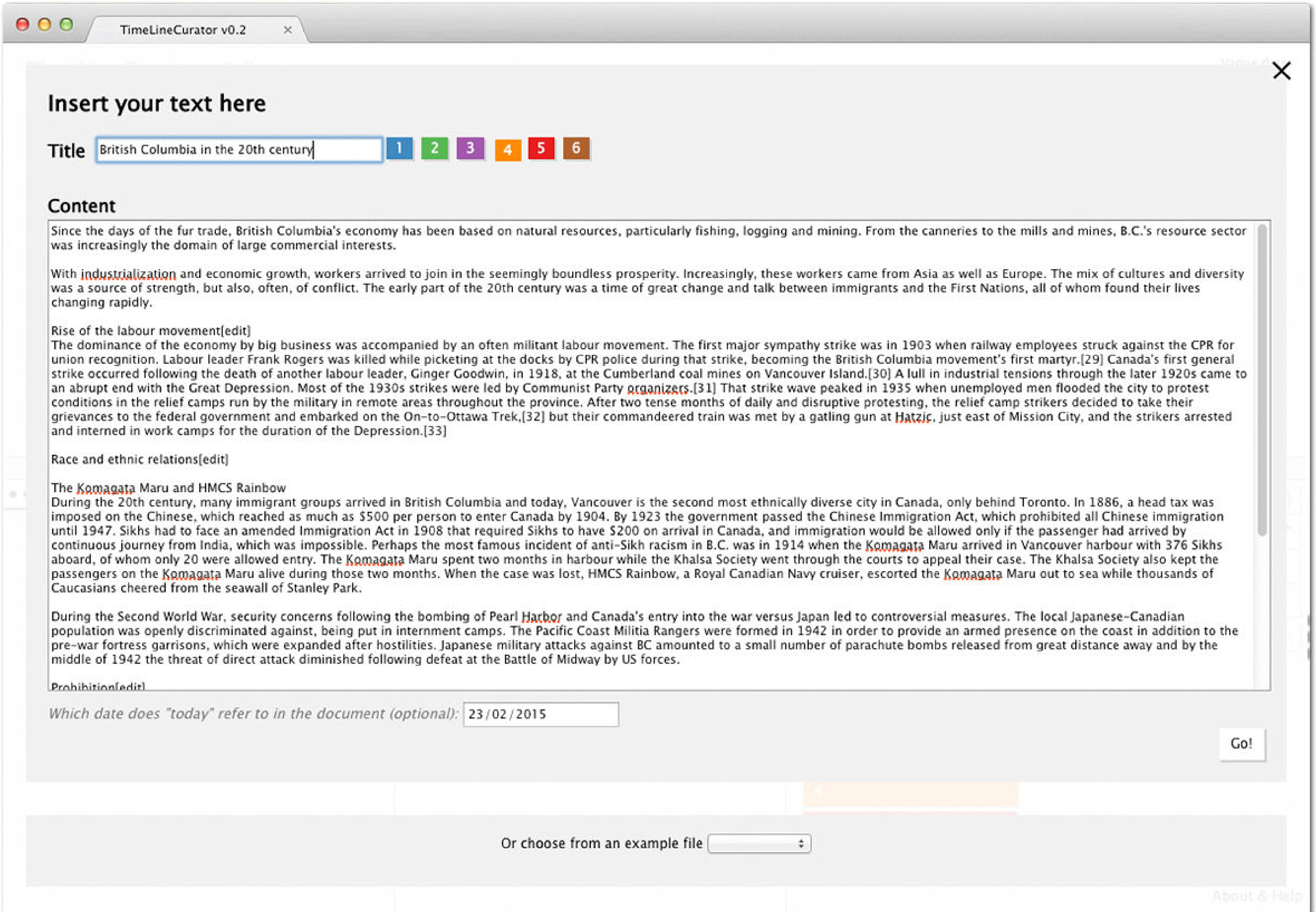
- "Multisearch" is customized to the workflow of journalists – brainstorming.

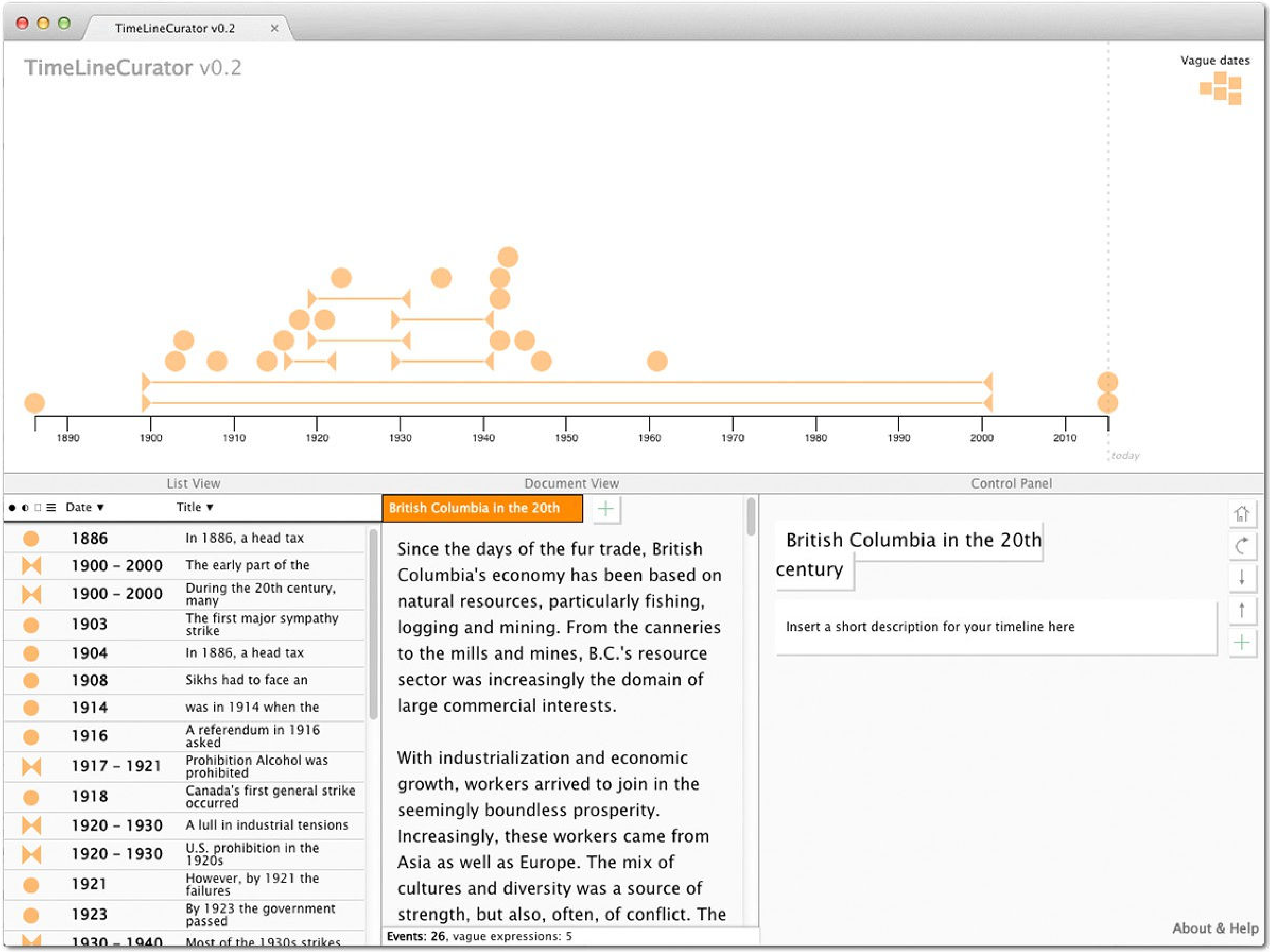


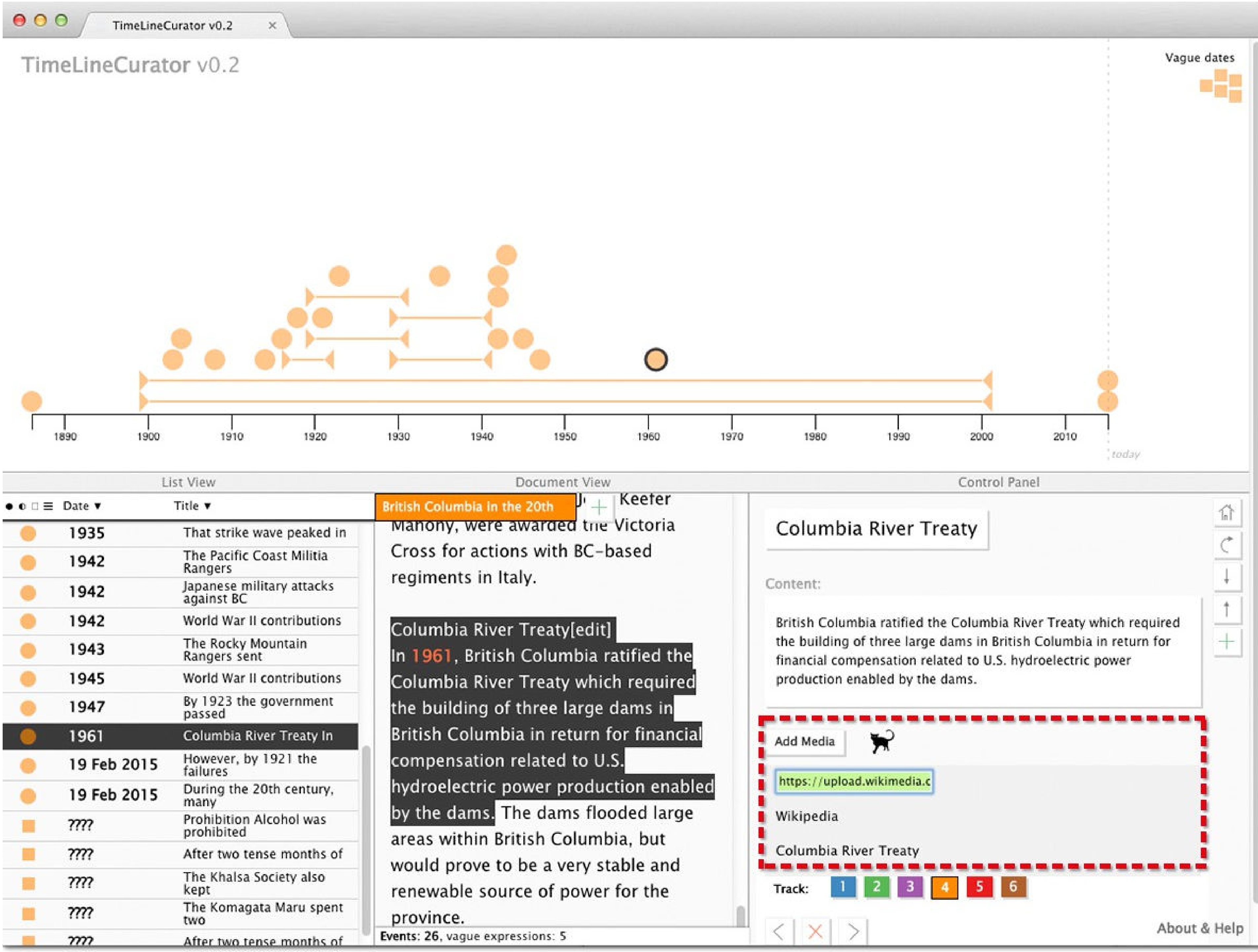
Overview – Document Clustering

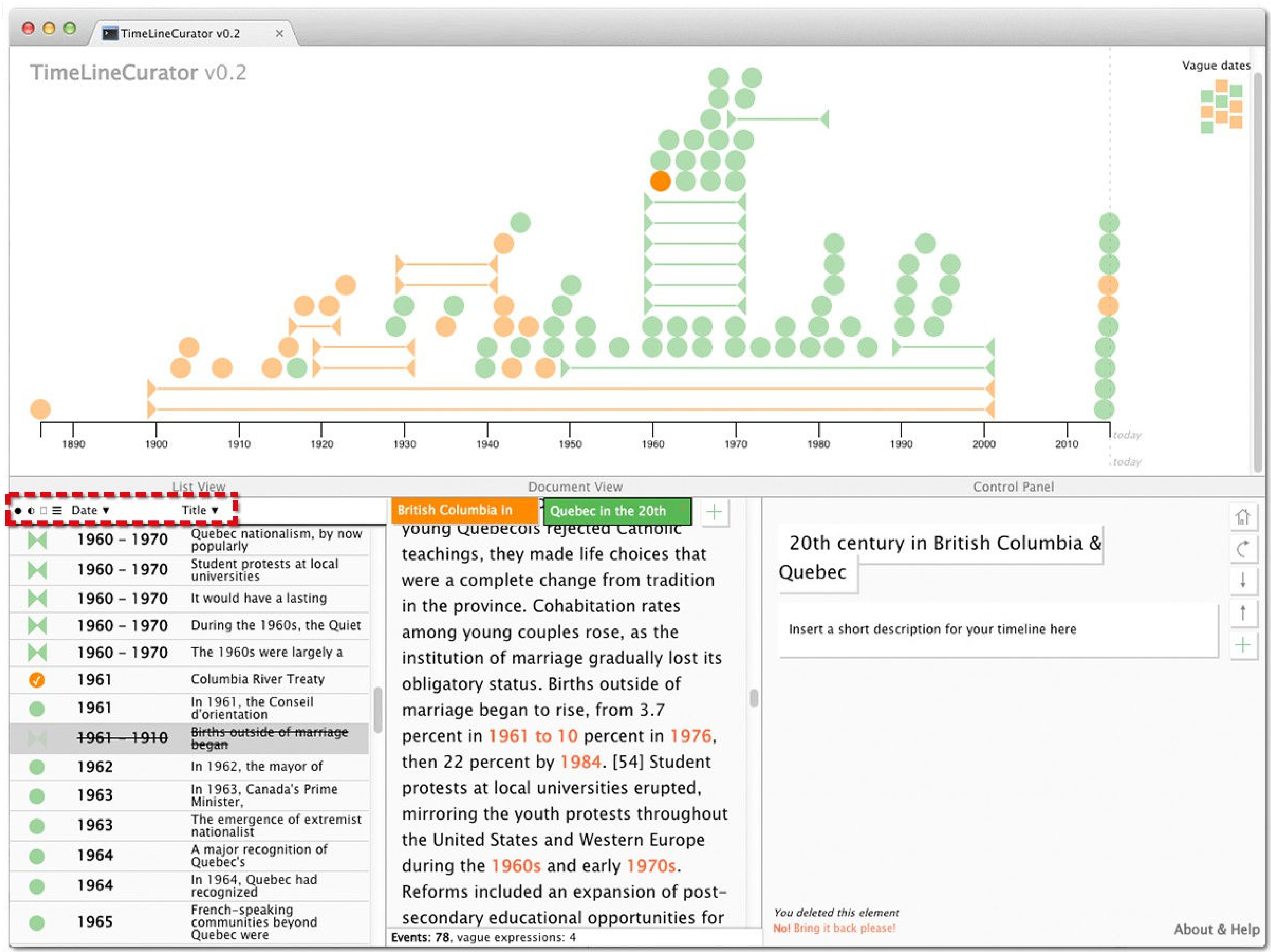


Timeline Curator




TimeLineCurator v0.2
TimeLineCurator v0.2
Vague dates
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
today
List View
Document View
Control Panel
Date ▾
Title ▾
1886 In 1886, a head tax
1900 – 2000 The early part of the
1900 – 2000 During the 20th century, many
1903 The first major sympathy strike
1904 In 1886, a head tax
1908 Sikhs had to face an
1914 was in 1914 when the
1916 A referendum in 1916 asked
1917 – 1921 Prohibition Alcohol was prohibited
1918 Canada's first general strike occurred
1920 – 1930 A lull in industrial tensions
1920 – 1930 U.S. prohibition in the 1920s
1921 However, by 1921 the failures
1923 By 1923 the government passed
British Columbia in the 20th
Since the days of the fur trade, British Columbia's economy has been based on natural resources, particularly fishing, logging and mining. From the canneries to the mills and mines, B.C.'s resource sector was increasingly the domain of large commercial interests.
With industrialization and economic growth, workers arrived to join in the seemingly boundless prosperity. Increasingly, these workers came from Asia as well as Europe. The mix of cultures and diversity was a source of strength, but also, often, of conflict. The
Events: 26, vague expressions: 5
British Columbia in the 20th century
Insert a short description for your timeline here
About & Help

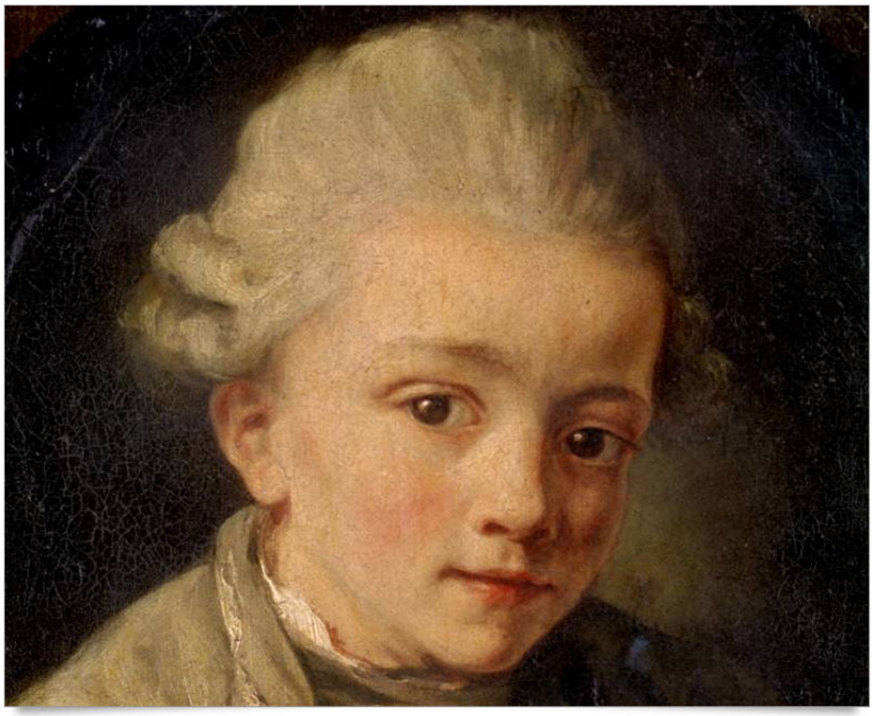

TimeLineCurator v0.2
TimeLineCurator v0.2
Vague dates
1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010
today
List View
Document View
Control Panel
Date ▾
Title ▾
1935 That strike wave peaked in
1942 The Pacific Coast Militia Rangers
1942 Japanese military attacks against BC
1942 World War II contributions
1943 The Rocky Mountain Rangers sent
1945 World War II contributions
1947 By 1923 the government passed
1961 Columbia River Treaty In
19 Feb 2015 However, by 1921 the failures
19 Feb 2015 During the 20th century, many
???? Prohibition Alcohol was prohibited
???? After two tense months of
???? The Khalsa Society also kept
???? The Komagata Maru spent two
???? After two tense months of
British Columbia in the 20th
Keefer
Mahony, were awarded the Victoria Cross for actions with BC-based regiments in Italy.
Columbia River Treaty[edit]
In 1961, British Columbia ratified the Columbia River Treaty which required the building of three large dams in British Columbia in return for financial compensation related to U.S. hydroelectric power production enabled by the dams. The dams flooded large areas within British Columbia, but would prove to be a very stable and renewable source of power for the province.
Events: 26, vague expressions: 5
Columbia River Treaty
Content:
British Columbia ratified the Columbia River Treaty which required the building of three large dams in British Columbia in return for financial compensation related to U.S. hydroelectric power production enabled by the dams.
Add Media
https://upload.wikimedia.c
Wikipedia
Columbia River Treaty
Track: 1 2 3 4 5 6
About & Help



TimeLineCurator v0.2
Vague dates
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
today
List View
Document View
Control Panel
Date ▾
Title ▾
1960 – 1970 Quebec nationalism, by now popularly
1960 – 1970 Student protests at local universities
1960 – 1970 It would have a lasting
1960 – 1970 During the 1960s, the Quiet
1960 – 1970 The 1960s were largely a
1961 Columbia River Treaty
1961 In 1961, the Conseil d'orientation
~~1961 – 1910 Births outside of marriage began~~
1962 In 1962, the mayor of
1963 In 1963, Canada's Prime Minister,
1963 The emergence of extremist nationalist
1964 A major recognition of Quebec's
1964 In 1964, Quebec had recognized
1965 French-speaking communities beyond Quebec were
British Columbia in
Quebec in the 20th
young Québécois rejected Catholic teachings, they made life choices that were a complete change from tradition in the province. Cohabitation rates among young couples rose, as the institution of marriage gradually lost its obligatory status. Births outside of marriage began to rise, from 3.7 percent in 1961 to 10 percent in 1976, then 22 percent by 1984. [54] Student protests at local universities erupted, mirroring the youth protests throughout the United States and Western Europe during the 1960s and early 1970s. Reforms included an expansion of post-secondary educational opportunities for
Events: 78, vague expressions: 4
20th century in British Columbia & Quebec
Insert a short description for your timeline here
You deleted this element
No! Bring it back please!
About & Help

Timeline Curator – Exported Timeline

1760
First musical compositions

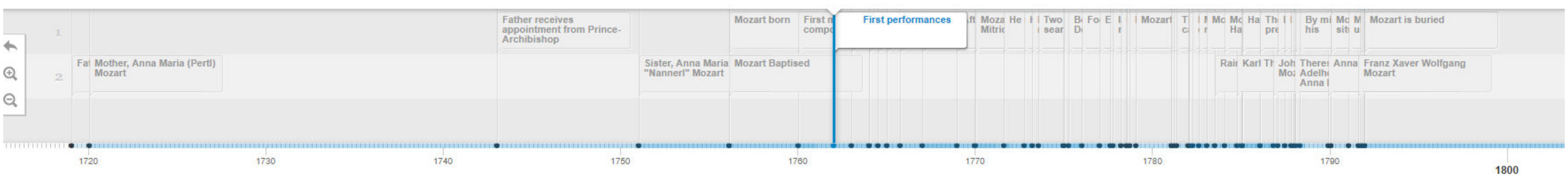


Mozart painted by Greuze, 1764

1762

First performances

These began with an exhibition, in 1762, at the court of the Prince-elector Maximilian III of Bavaria in Munich, and at the Imperial Court in Vienna and Prague.



Topic Modelling?!

"In machine learning and natural language processing, a topic model is a type of statistical model for **discovering the abstract "topics"** that occur in a **collection of documents**. "

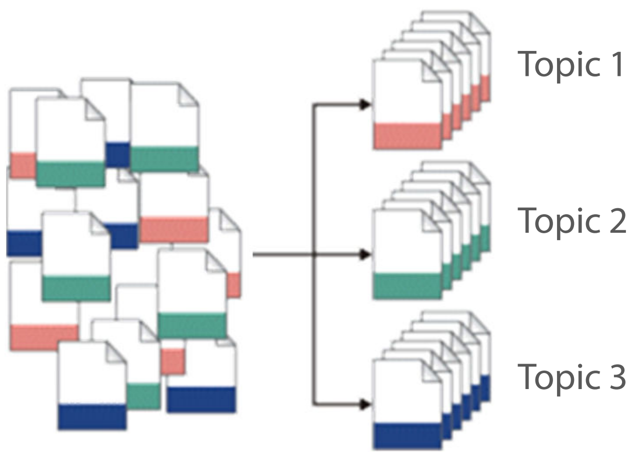
- Wikipedia

Topic Modelling = Clustering of documents into topics

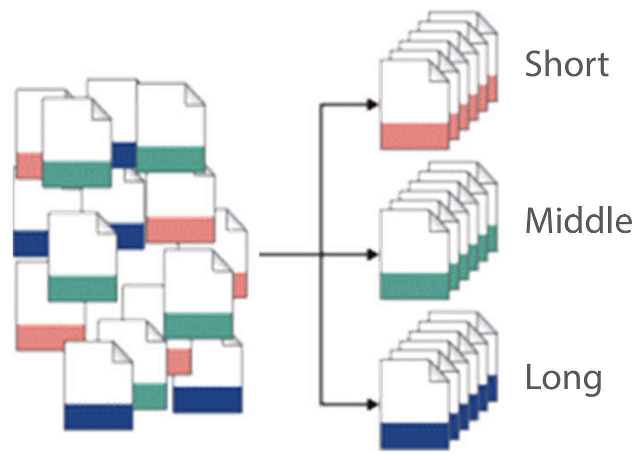
- Hierarchical **vs.** Non-hierarchical
- Partial **vs.** Agglomerative
- Deterministic **vs.** Probabilistic
- Incremental **vs.** Batch
- Supervised **vs.** Unsupervised

Text Clustering $\supsetneq$ Topic Modeling

Clustering = finding groups of *similar objects* in the data



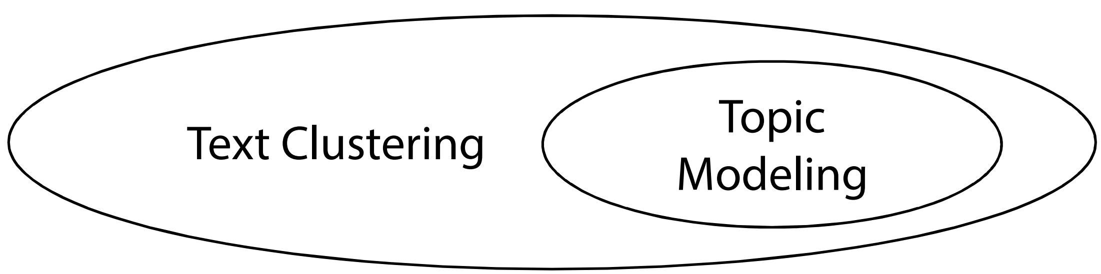
Similarity: Document Content



Similarity: Document Length

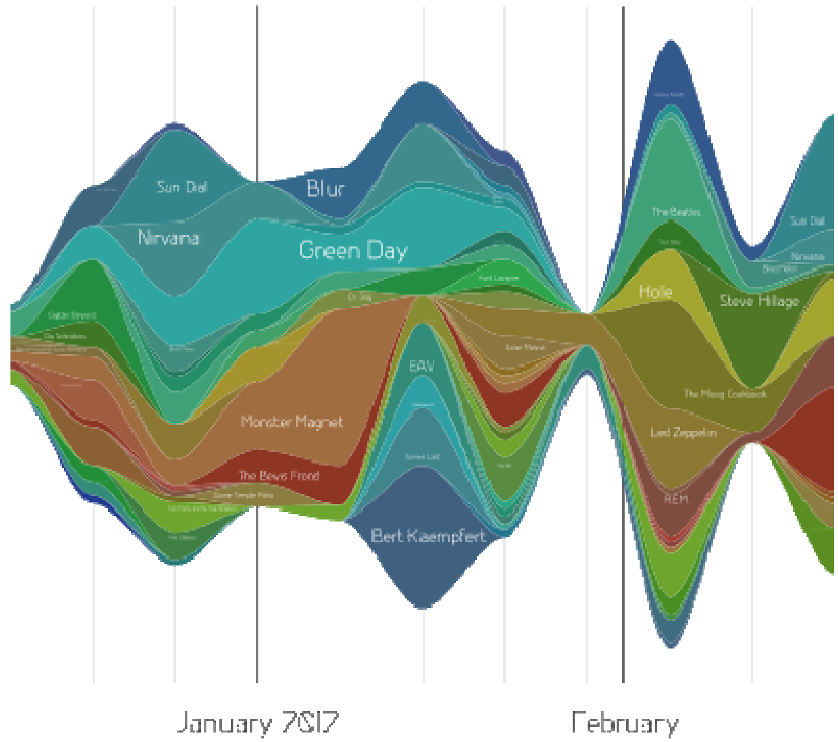
Topic Modeling
= discovering the abstract topics that occur in a documents collection

= clustering documents (with respect to their content) into topics



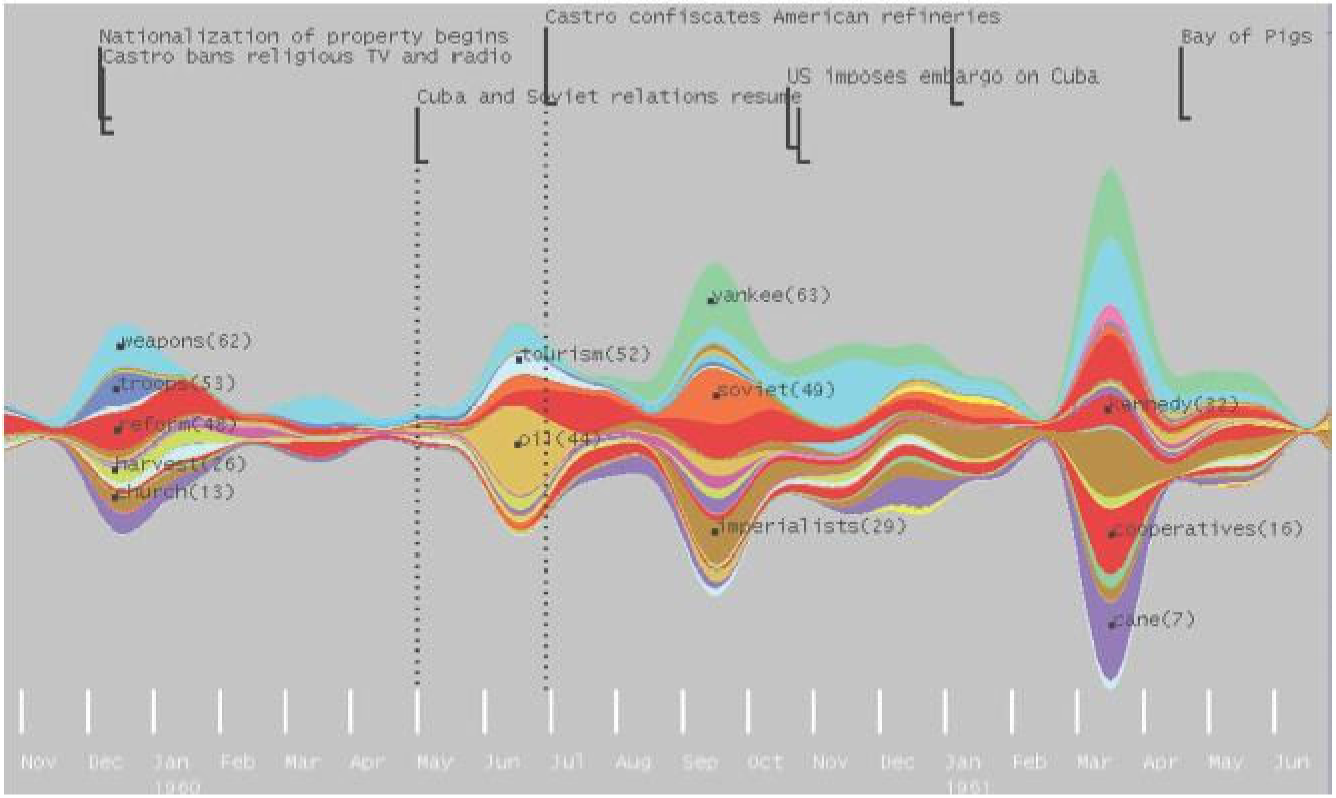
Aside: Stacked area graph / Stream graph

- Streams represent data element prominence over time
- Overall flow shows data volume



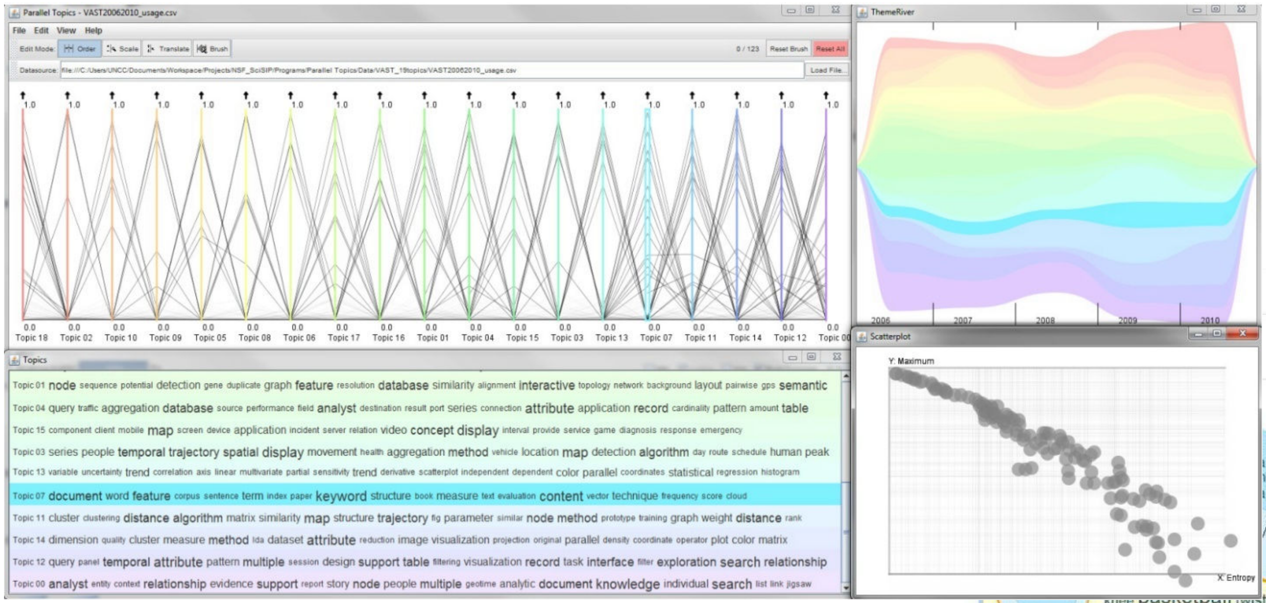
Theme River

- Canonical example of stream graphs
- News themes over time

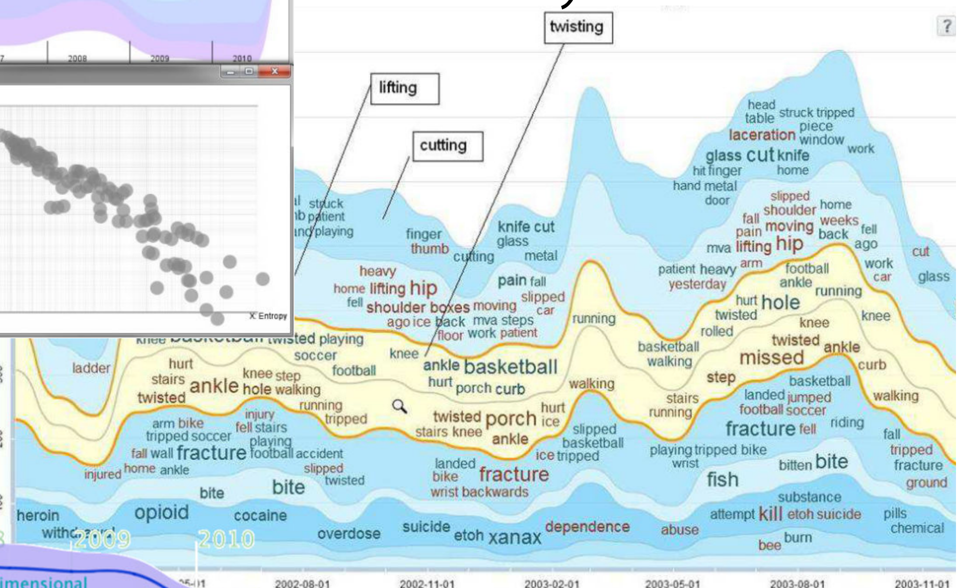


Susan Havre, IEEE Computer Society, *ThemeRiver: Visualizing Thematic Changes in Large Document Collections*, Vol.8. No.1., January-March 2002

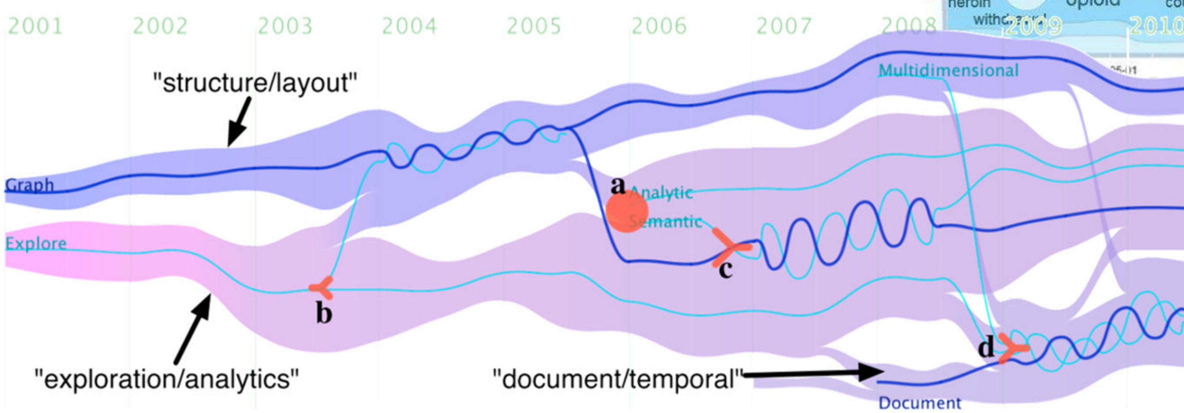
Visualizing Topic Modeling



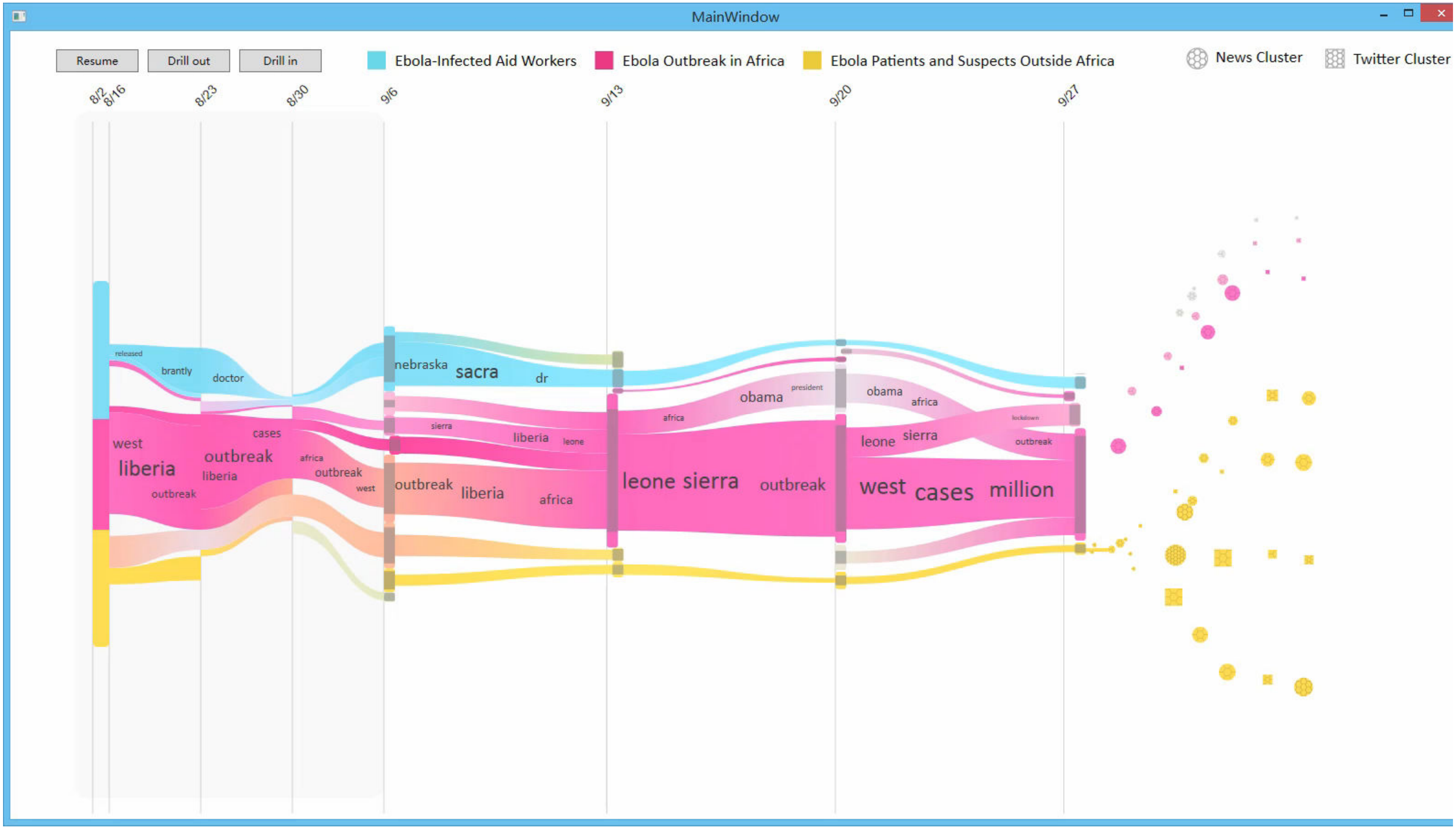
[3] **Paralleltopics: A probabilistic approach to exploring document collections** *by Dou et al.*



[1] **Tiara: a visual exploratory text analytic system** *by Wei et al.*

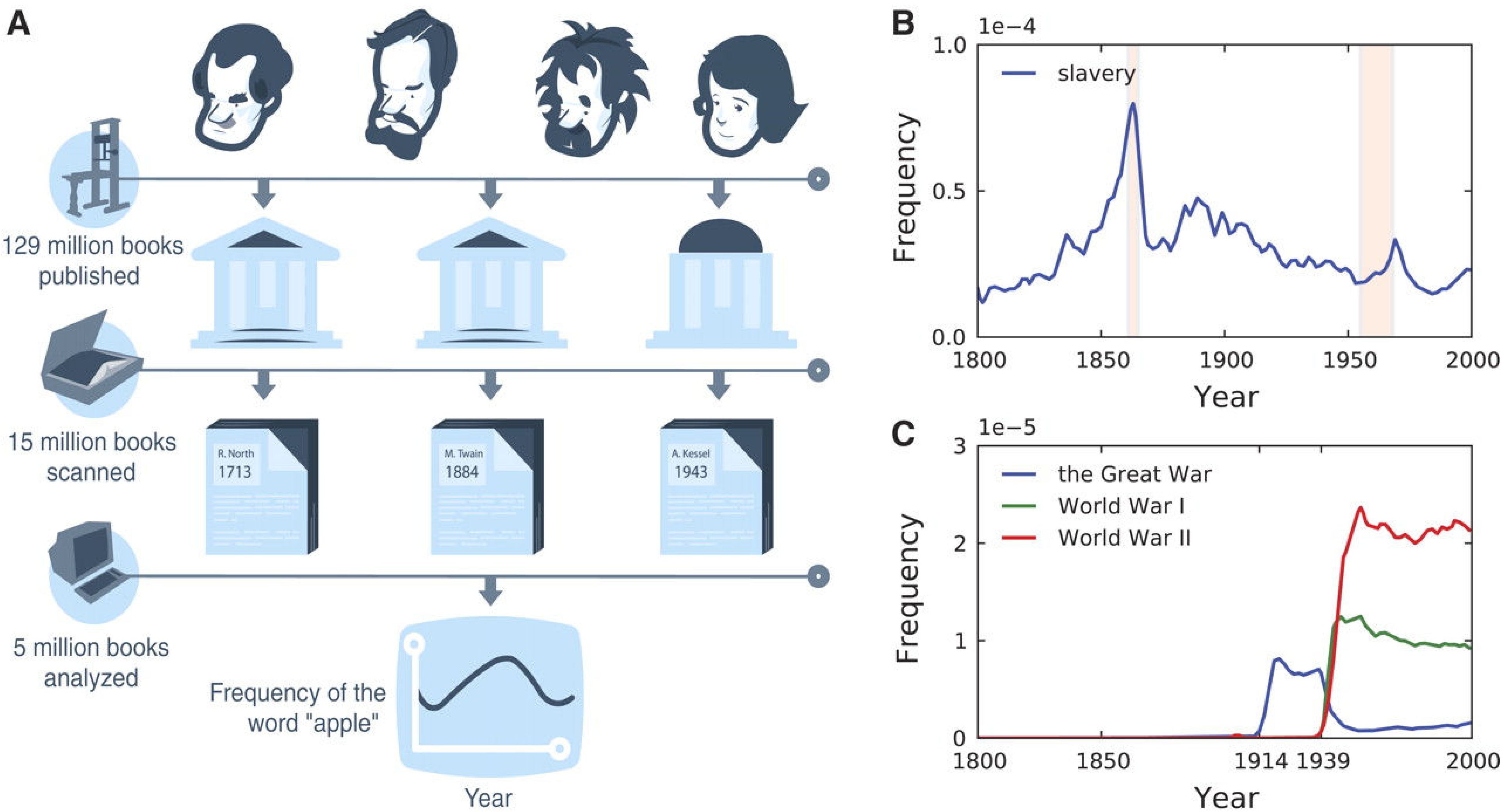


[2] **Textflow: Towards better understanding of evolving topics in text** *by Cui et al.*



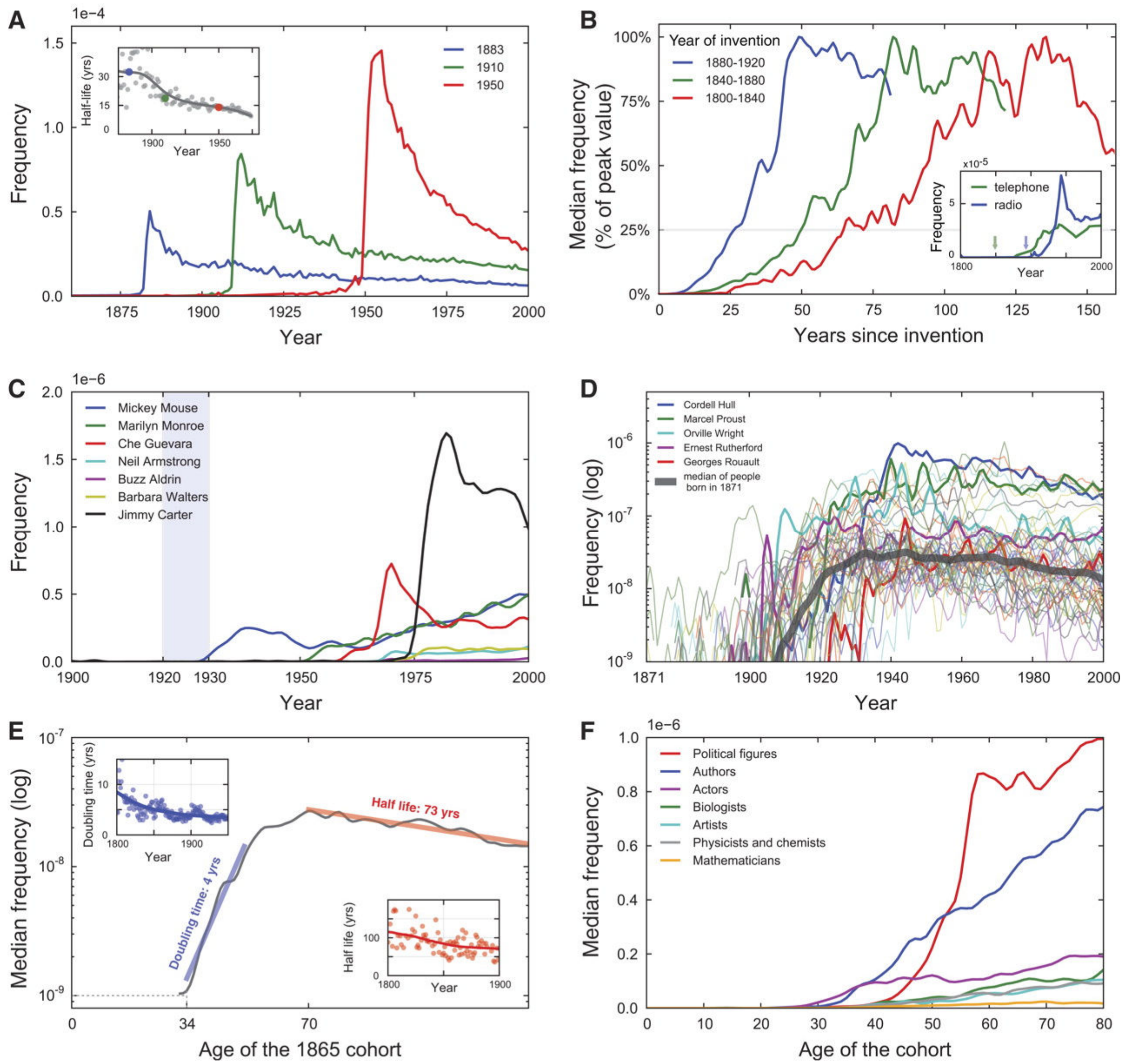
S. Liu, J. Yin, X. Wang, W. Cui, K. Cao and J. Pei, "Online Visual Analytics of Text Streams," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 11, pp. 2451-2466, Nov. 1 2016.

Large Scale Corpus Analysis: Culturomics



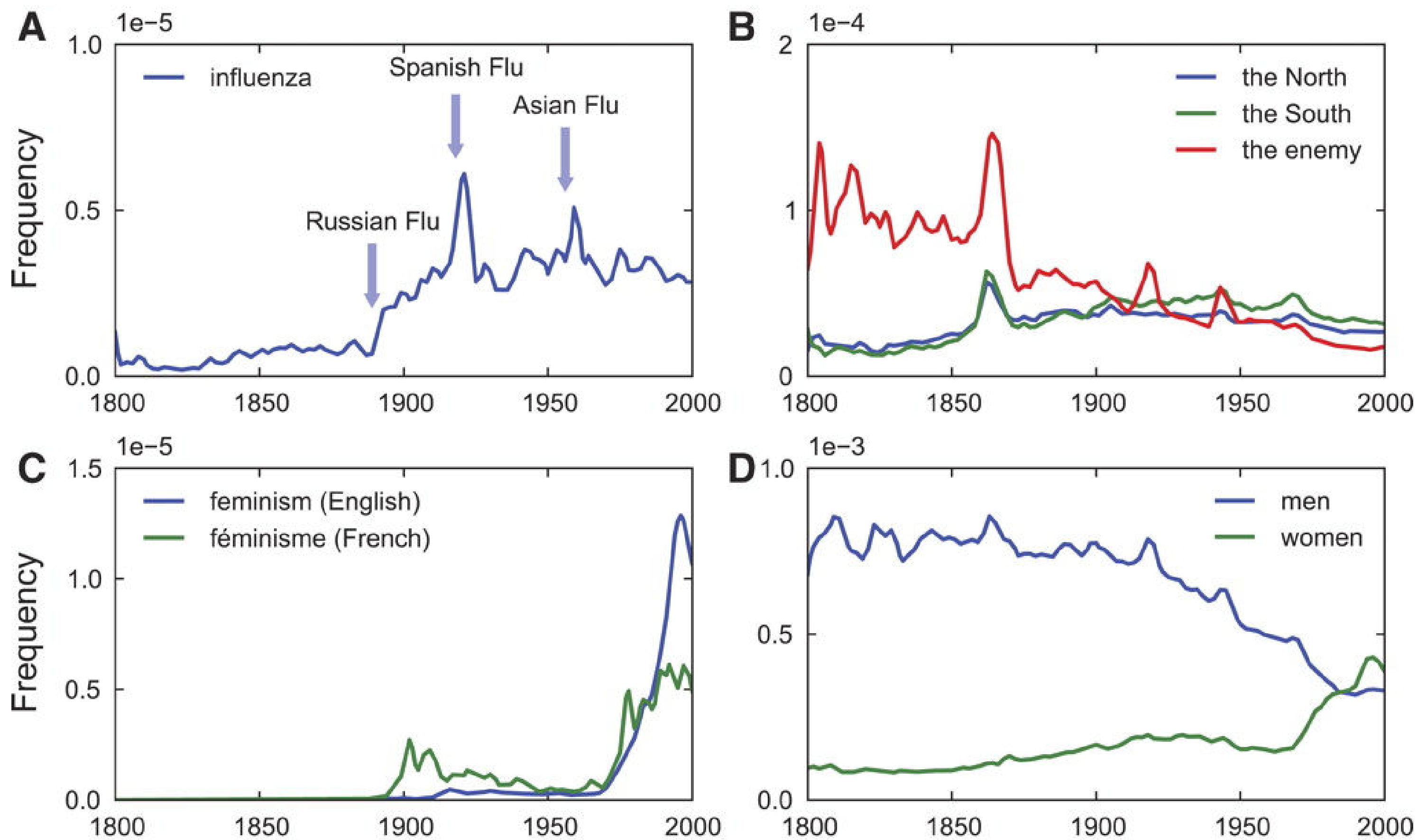
Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331, no. 6014 (January 14, 2011): 176–182.

Culturomics

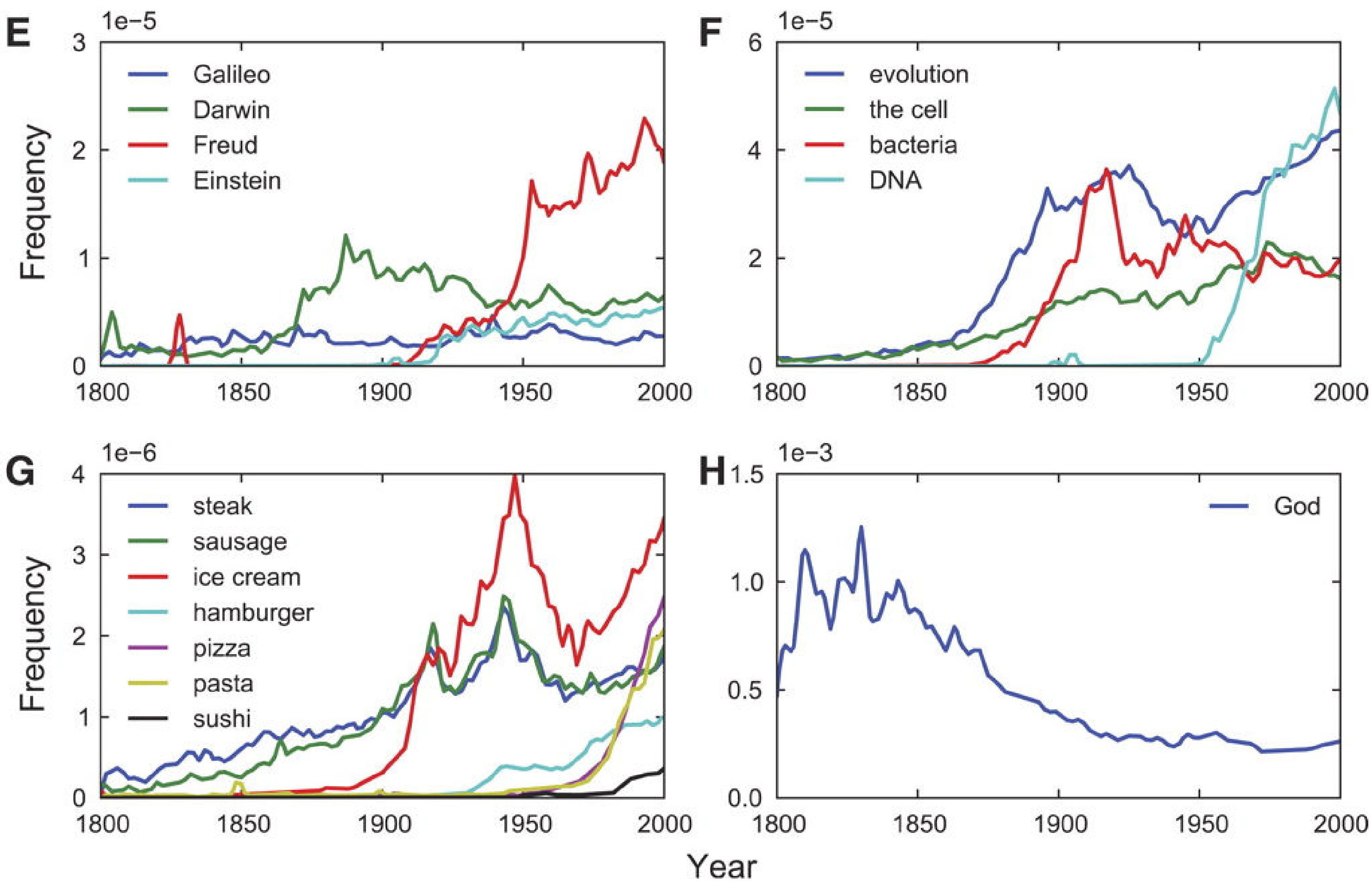


Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* 331, no. 6014 (January 14, 2011): 176–182.

Culturomics



Culturomics



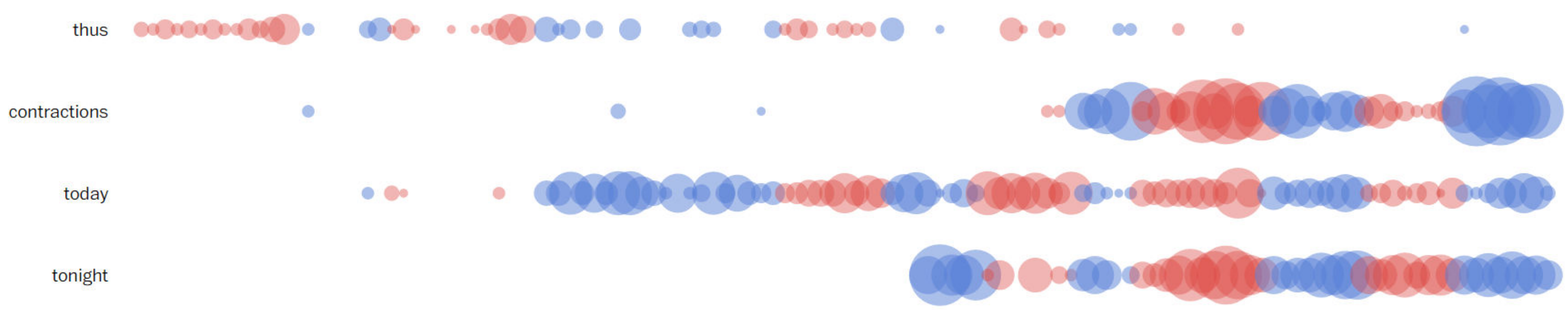
History of the President's Words

Daily Lexicon

With the rise of contractions and the fall of "thus," Liberman said this is illustrative of a larger, secular trend. "This obviously interacts with genre, and in particular with formality," he said, stating that "won't" appears much more frequently in magazines than "will not," while less of a gap exists in academic papers.

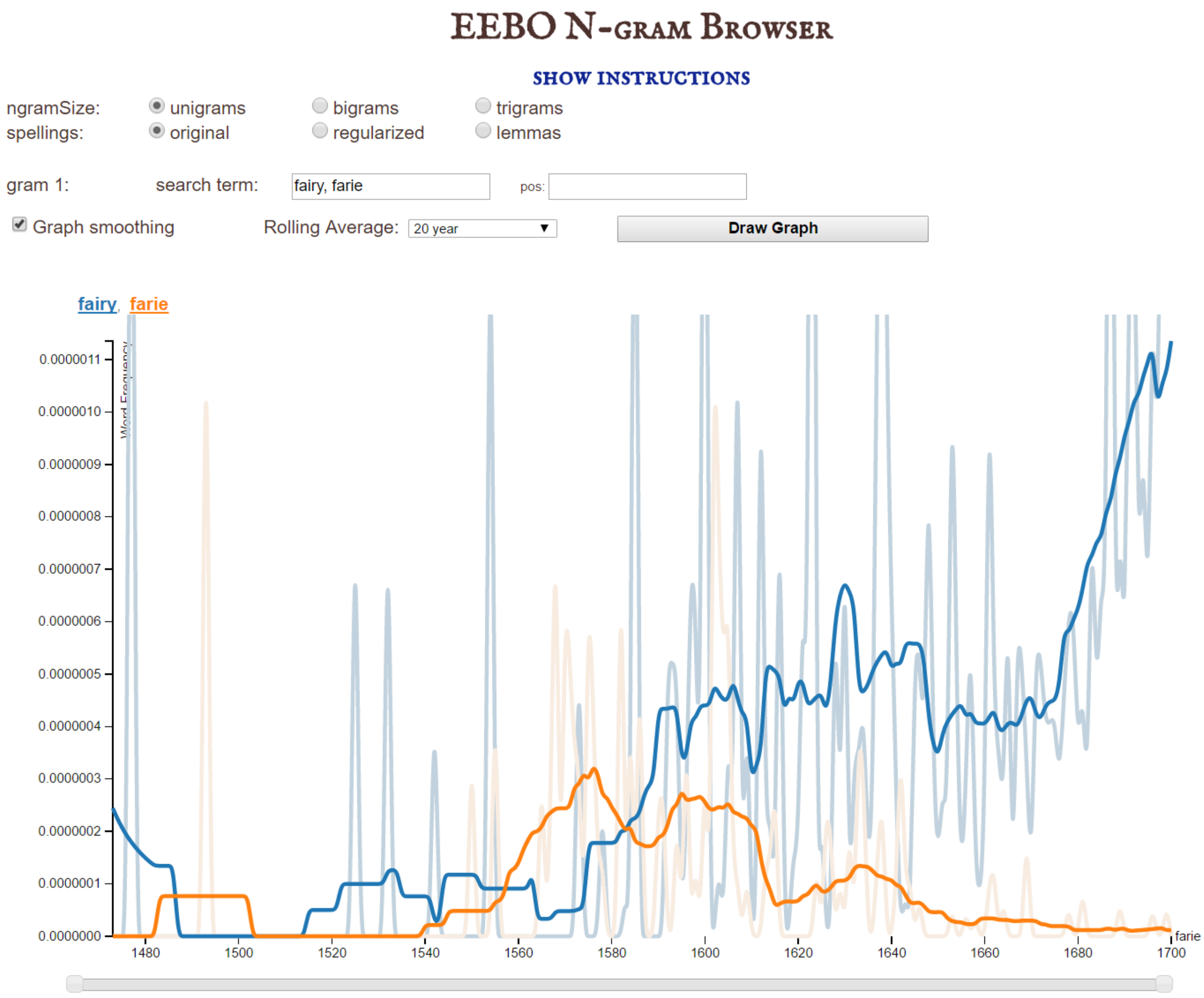
State of the Union addresses incorporated contractions even later than secular dialogue, which hints at the level of formality presidents wish to project. Liberman suggested that the spoken version of the addresses may be more relaxed and contraction-laden than the much more prescribed written documents let on. "Perhaps Truman, Eisenhower and Kennedy were using contractions in their performances that were not written as such in their texts," he said. The increased use of contractions suggests a growing laxity in the written documents rather than in spoken words.

The rise of "tonight," Fields clarified, was simply the result of the State of the Union addresses occurring in the evening time for the first time with Lyndon B. Johnson's 1965 address.



https://www.washingtonpost.com/graphics/politics/2016-sotu/language/

Culturomics: Language Change



http://earlyprint.wustl.edu/

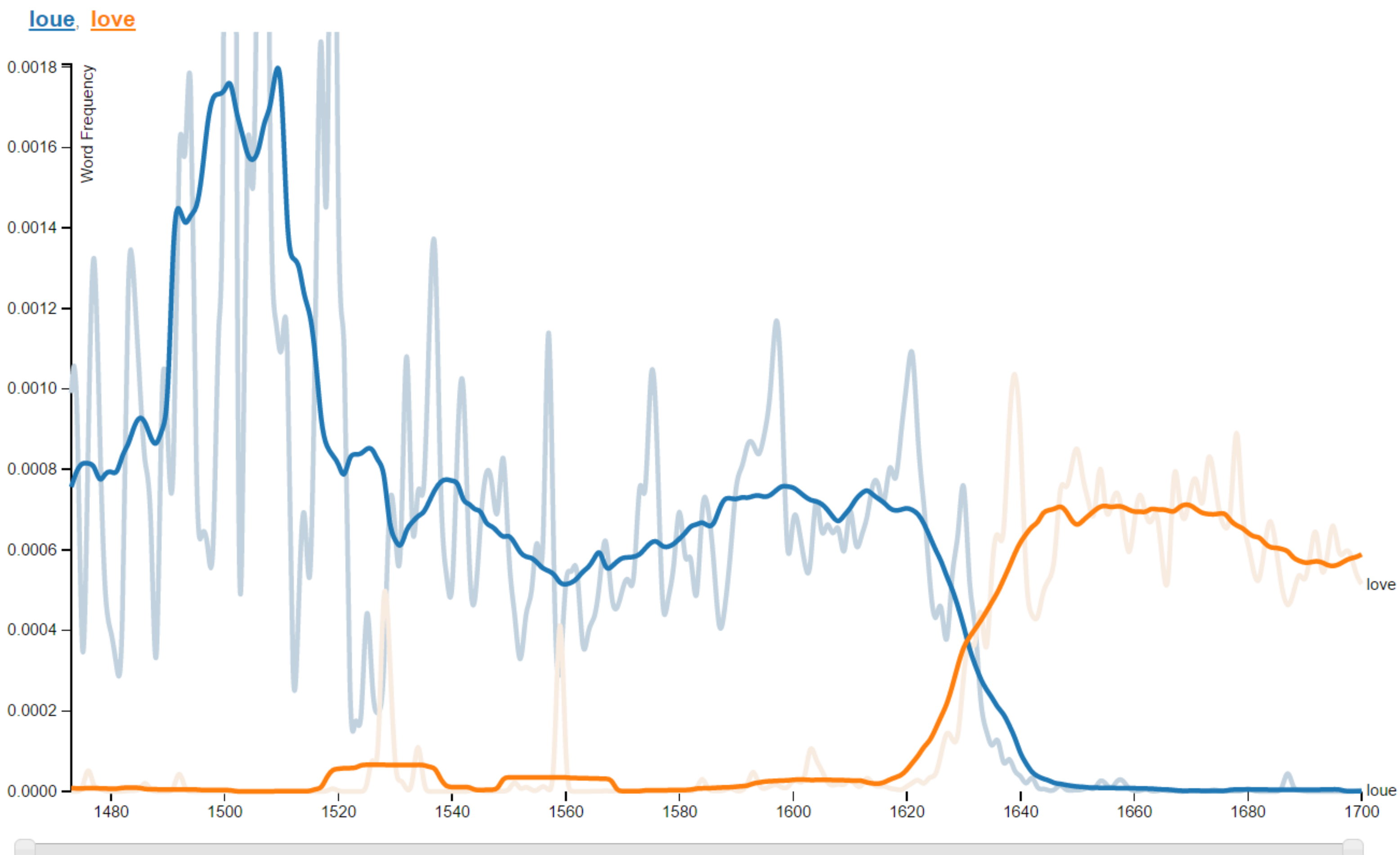
EEBO N-gram Browser

SHOW INSTRUCTIONS

ngramSize: unigrams bigrams trigrams
spellings: original regularized lemmas

gram 1: search term: love,loue pos:

Graph smoothing Rolling Average: 20 year ▼ **Draw Graph**



Metatation

Interdisciplinary Research Case Study

Hrim Mehta, Adam Bradley, Mark Hancock, Christopher Collins

Project Overview

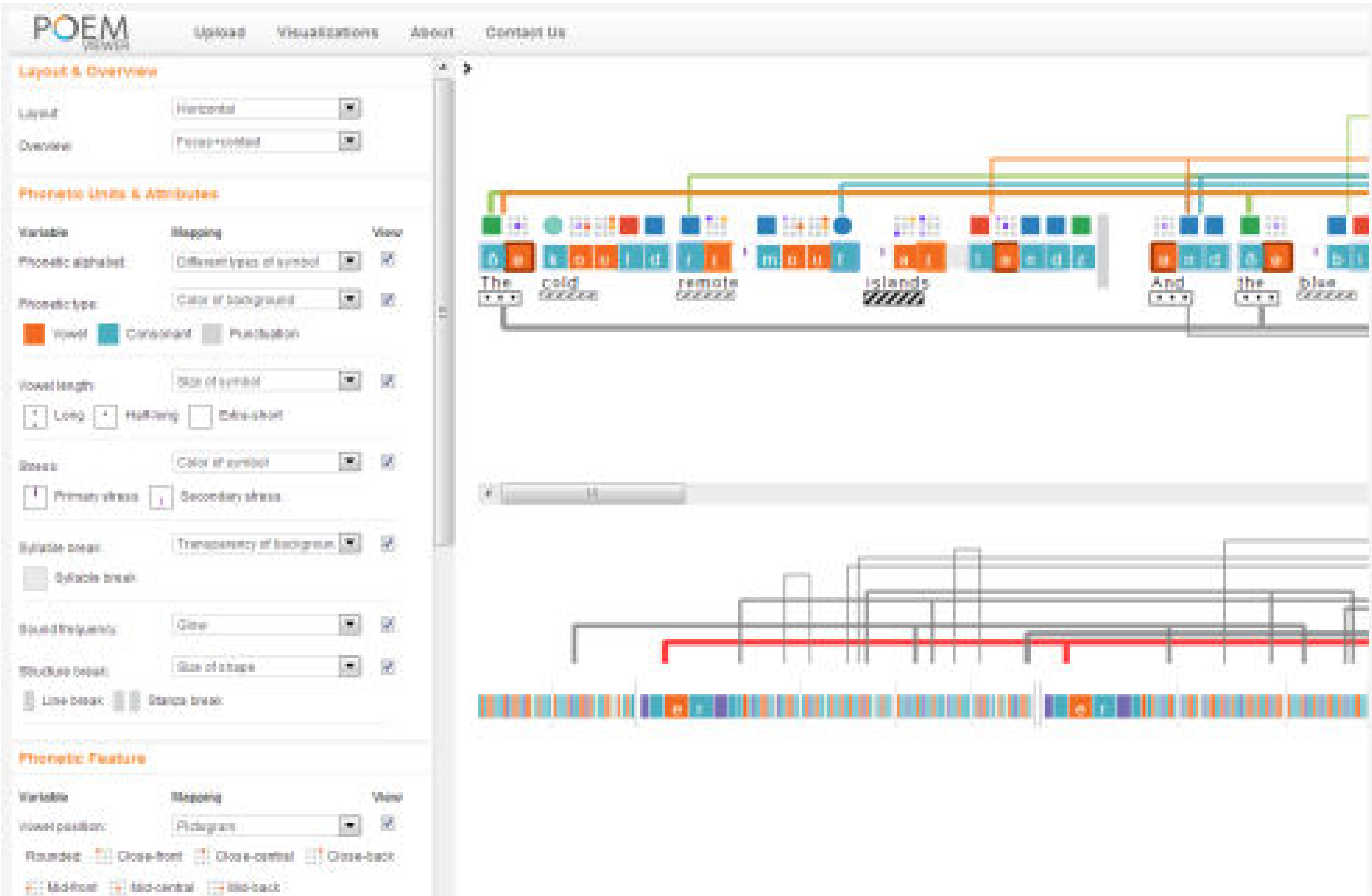
- Team Expertise:
  - Human-Computer Interaction
  - Computational Linguistics
  - English Literature and Poetry
- Methodology:
  - Grounding Study + Analysis
  - Derive Requirements
  - Prototyping & Design
  - Expert Feedback

Free-form annotations

as

implicit interactions

Existing Tools for Literary Analysis



PoemViewer by Oxford Visual Informatics Lab

Limitations of Existing Literary Analysis Tools

- Disconnected from workflow
- Premature data presentation

Understanding the Context

Observational study of
annotation practice of poetry critics
as they analyse a poem

Observational Study

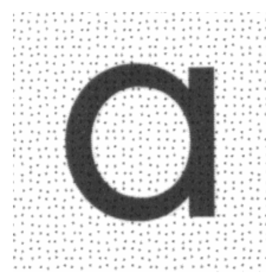
Participants

14 literary critics (3 PhD students & 11 university professors)

Task

Close reading of assigned pairs of poems printed on Anoto paper with Livescribe Anoto pen for annotations, if required

Analysis session followed by retrospective think aloud



Dataset

14 poems in total, 7 pairs of poems assigned to 7 participants

anoto.com

Categorisation of Annotations

Annotation Form

Ellipses, Underlines, Connectors, Polygons, Brackets, Text, Miscellaneous notations

Cognitive Purpose

CO (Computational Offloading)

EML (Externalizing to reduce Memory Load)

EML + CO

Space on Page

Word space, White space, Margin

Results

- Pen-and-paper over digital tools
- Experiential cognition
- Access of external resources
- Polymorphism of annotations
  - Annotation form and annotation function **do not** have a one-to-one relation

Results

- Pen-and-paper over digital tools
- Experiential cognition
- Access of external resources
- Polymorphism of annotations
  - Annotation form and annotation function **do not** have a one-to-one relation
  - Consistency in participant activity but the form, space & time vary

Results

- Pen-and-paper over digital tools
- Experiential cognition
- Access of external resources
- Polymorphism of annotations
  - Annotation form and annotation function **do not** have a one-to-one relation
  - Consistency in participant activity but the form, space & time vary
  - Characteristics of CO & EML

CO Characteristics

making a short, unhealthy life the shorter.
I kill it, and another instant's added
to the horrifying mortmain of
ephemera: keys, drift, sea-urchin shells,
you packrat off with joy . . . a dead fly swept

We stood by a pond that winter day,
And the sun was white, as though chidden of God,
And a few leaves lay on the starving sod;
– They had fallen from an ash, and were gray.

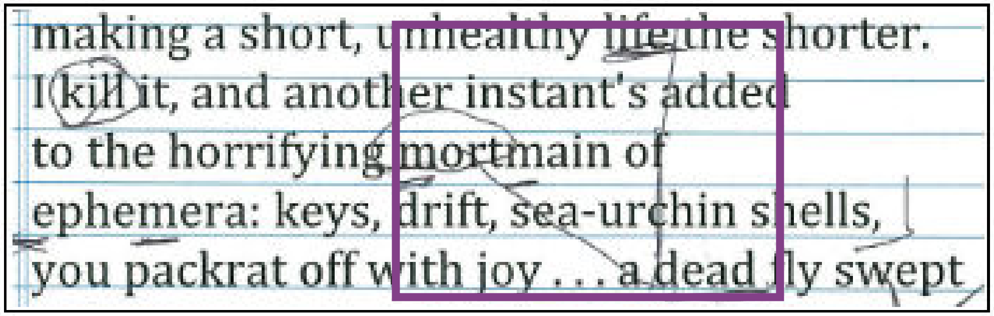
Your eyes on me were as eyes that rove
Over tedious riddles of years ago;
And some words played between us to and fro
On which lost the more by our love.

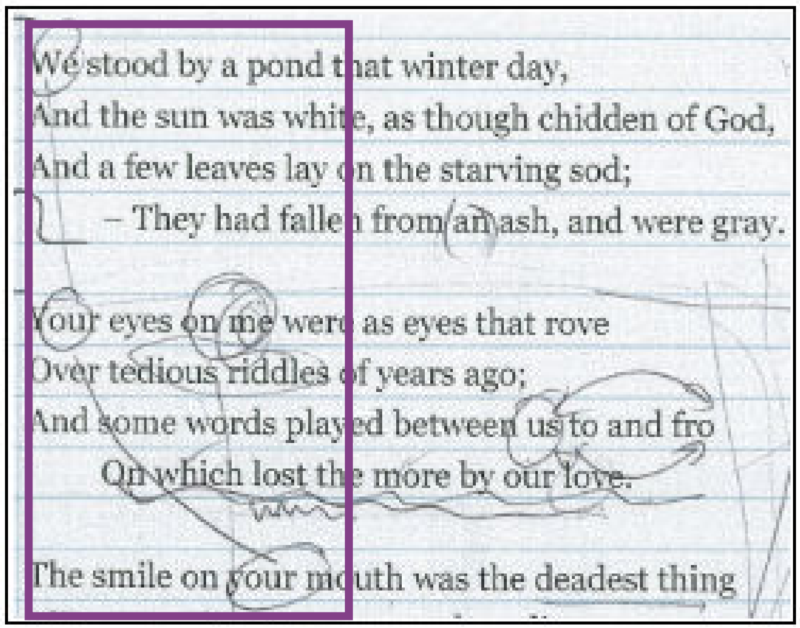
The smile on your mouth was the deadest thing

CO Characteristics

Had, having, and in quest to have, extreme;
A bliss in proof, and proved, a very woe;
Before, a joy proposed; behind, a dream.

CO Characteristics

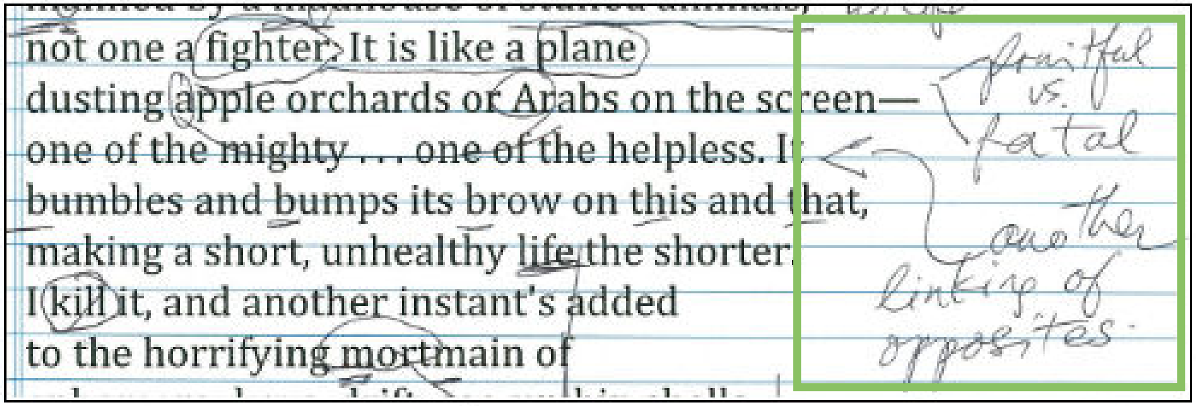

making a short, unhealthy life the shorter.
I kill it, and another instant's added
to the horrifying mortmain of
ephemera: keys, drift, sea-urchin shells,
you packrat off with joy . . . a dead fly swept


We stood by a pond that winter day,
And the sun was white, as though chidden of God,
And a few leaves lay on the starving sod;
– They had fallen from an ash, and were gray.

Your eyes on me were as eyes that rove
Over tedious riddles of years ago;
And some words played between us to and fro
On which lost the more by our love.

The smile on your mouth was the deadest thing

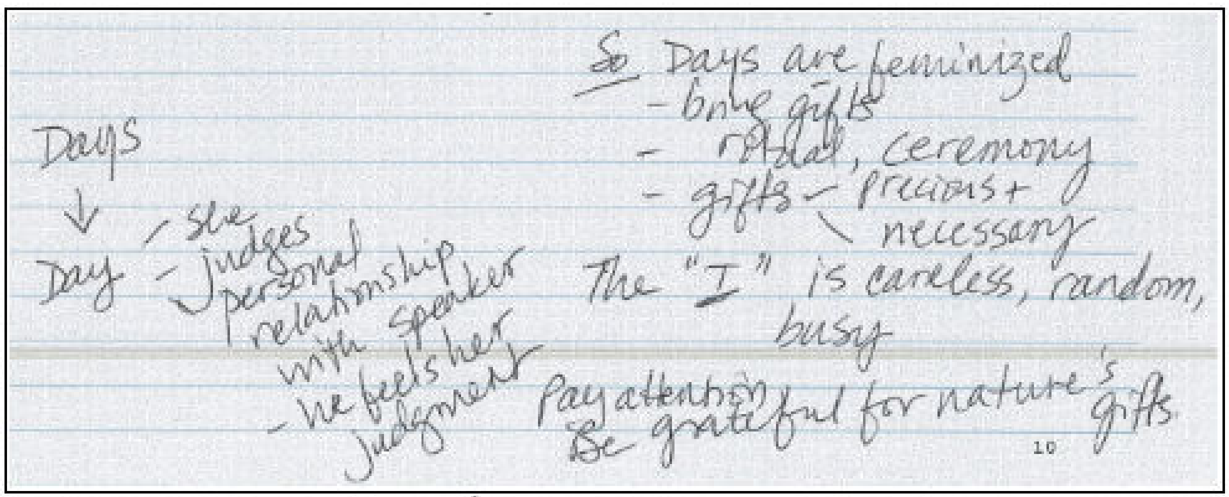
EML Characteristics



not one a fighter. It is like a plane
dusting apple orchards or Arabs on the screen—
one of the mighty . . . one of the helpless. It
bumbles and bumps its brow on this and that,
making a short, unhealthy life the shorter.
I kill it, and another instant's added
to the horrifying mortmain of

fruitful
vs.
fatal

another
linking of
opposites.



Days
↓
Day
- she judges
personal
relationship
with speaker
- he feels her
judgment

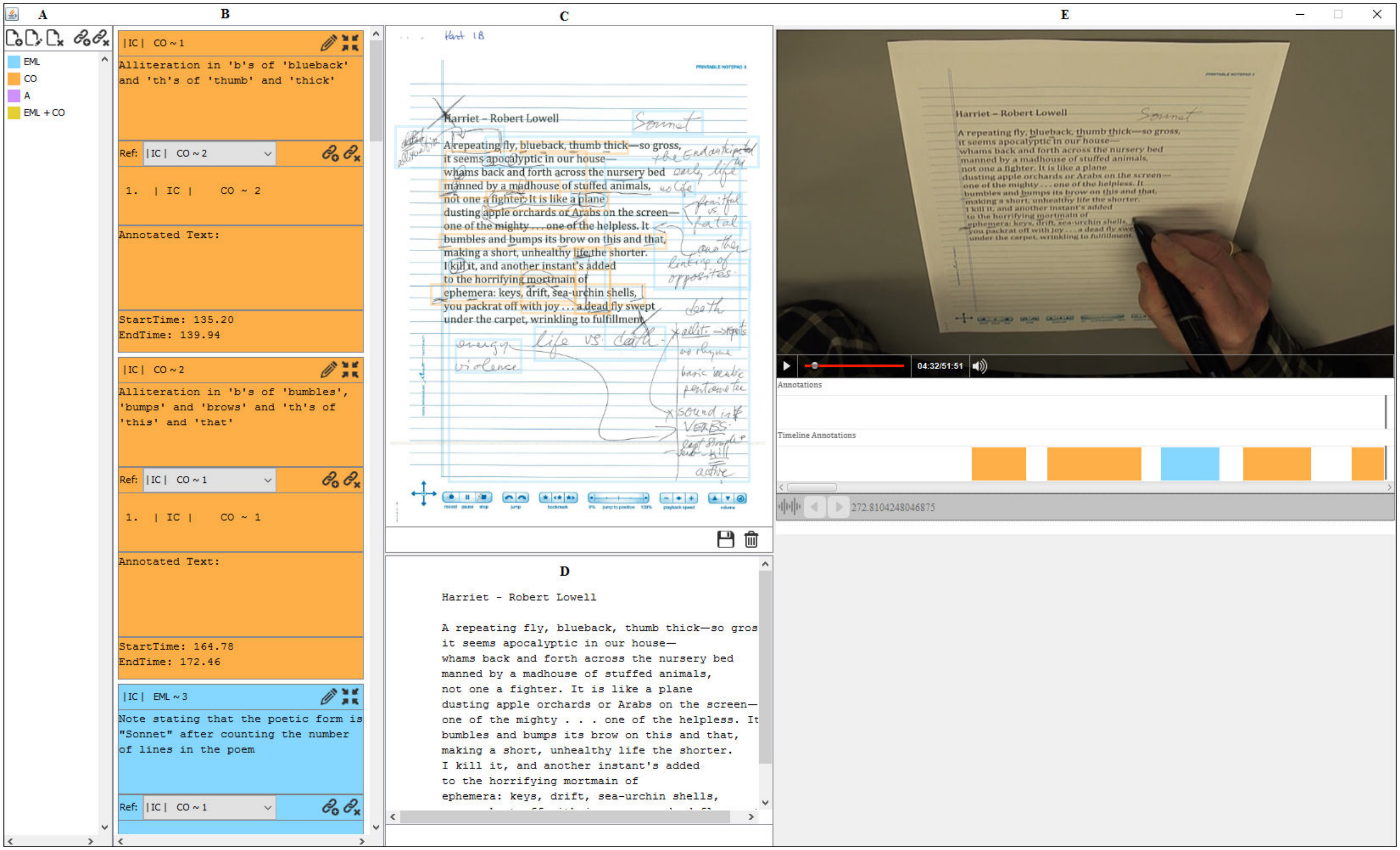
So Days are feminized
- bring gifts
- ritual, ceremony
- gifts — precious +
necessary

The "I" is careless, random,
busy

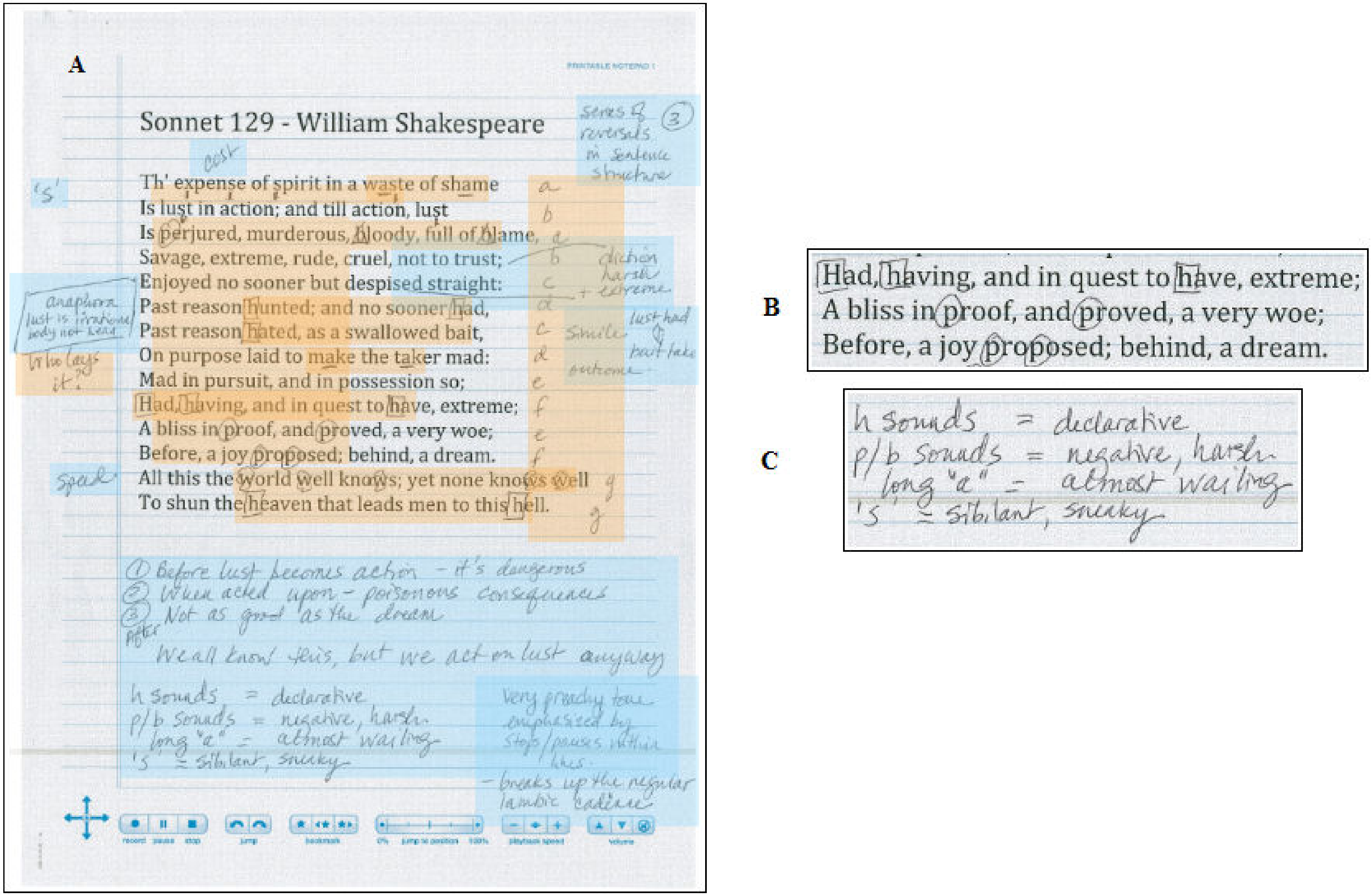
Pay attention
Be grateful for nature's gifts

10

Data Analysis Tool



Data Coding Example



Implications for Design

- Support for free-form annotations
- Permitting experiential cognition
- Minimal interruption to work flow
- Invoking analytic assistance via annotations

Metatation

Desktop-based application + physical paper

- Support for free-form annotations

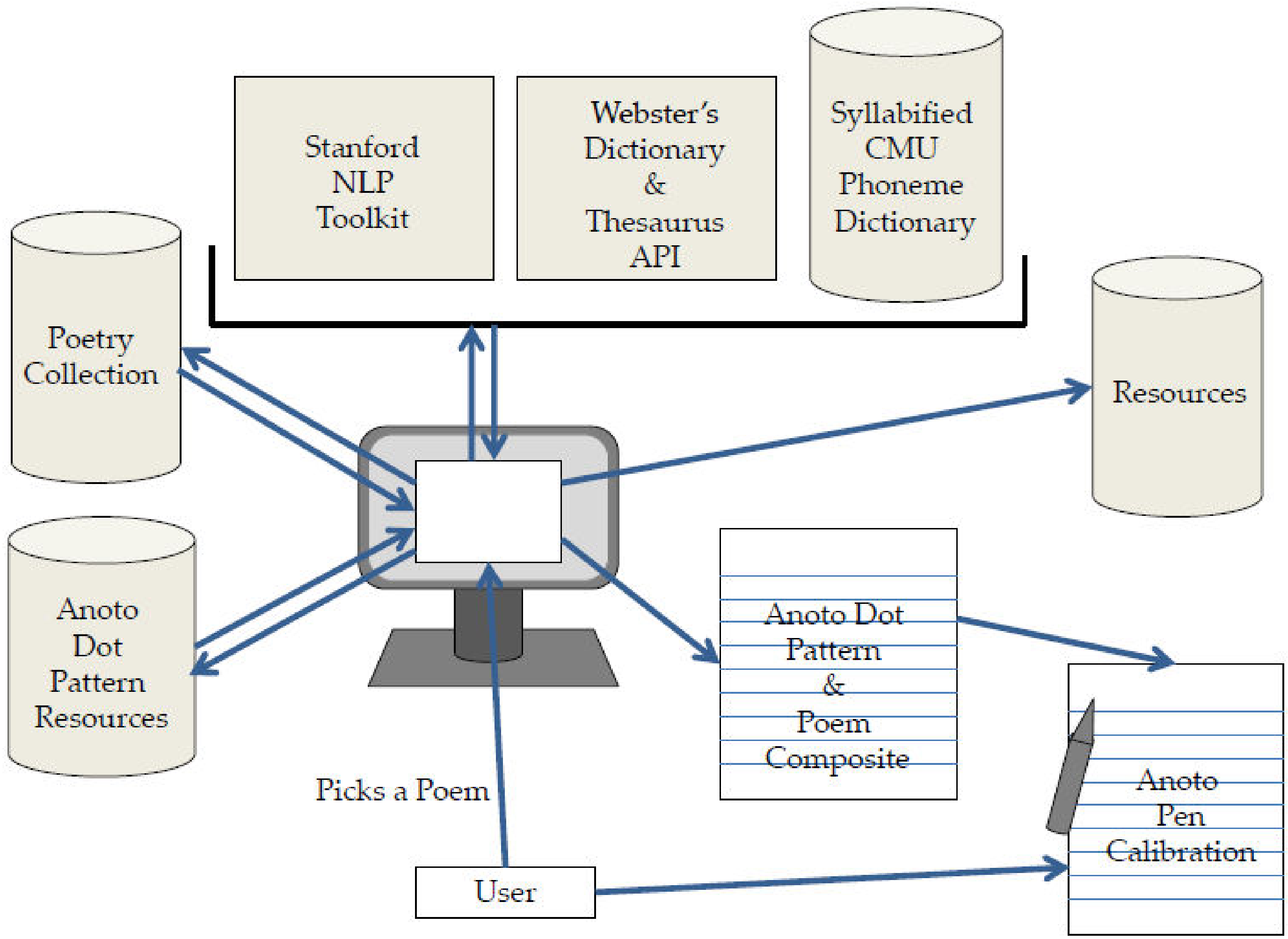
CO annotations for generating supplementary data

- Permitting experiential cognition
- Invoking analytic assistance via annotations

Analytic assistance presented on peripheral display

- Minimal interruption to work flow

Metatation: Preprocessing



Assonance, Consonance & Alliteration

Assonance

- Repetition of vowel sounds
- Rude: (R/**UW1**/D) & Cruel: (K/R/**UW1**/L)

Consonance

- Repetition of consonant sounds
- Extreme: (IH0/K)(S/T/**R**/IY1/M) & Rude: (**R**/UW1/D)

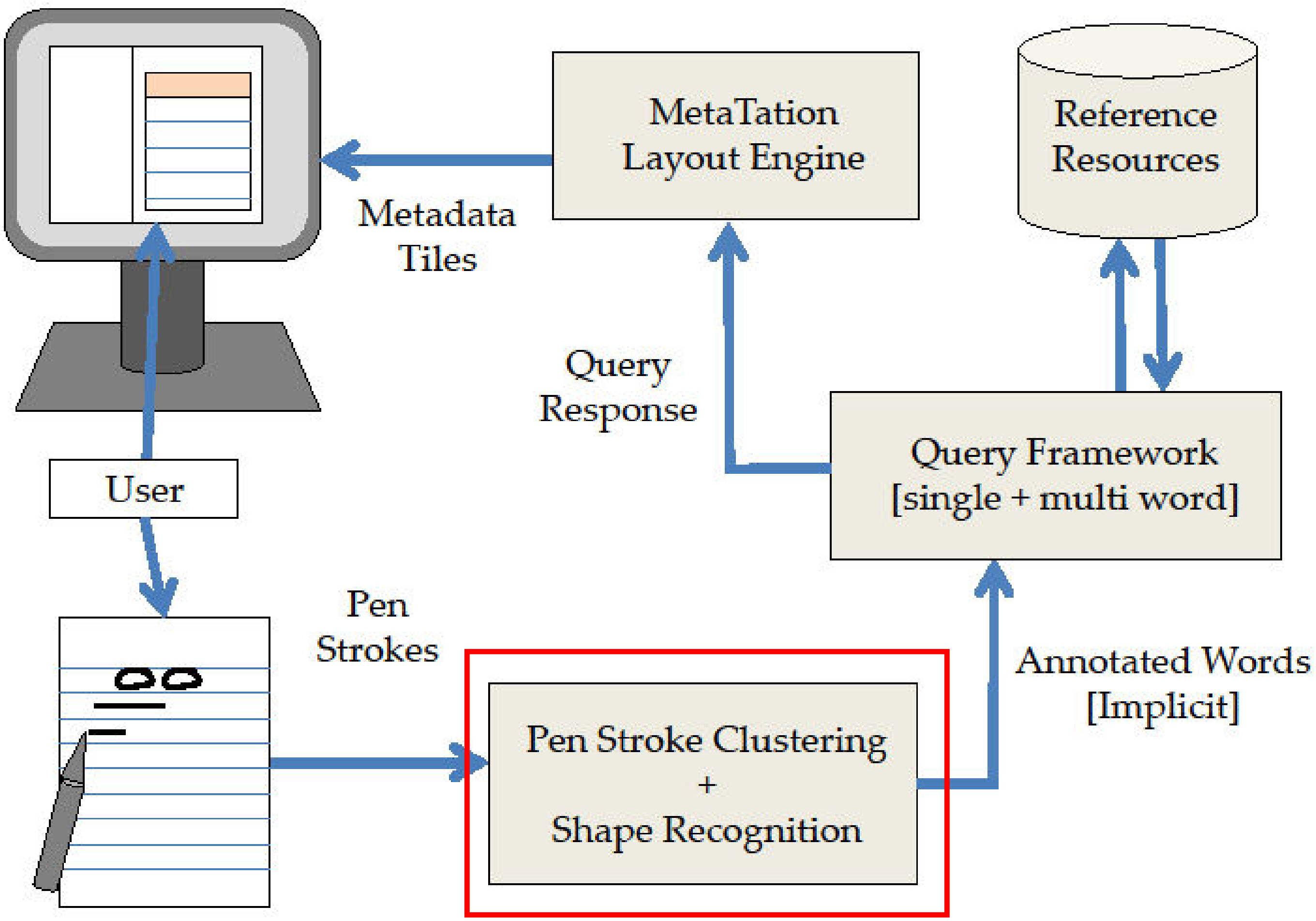
Alliteration

- Repetition of consonant sounds in stressed syllable
- Proof: (**P**/R/UW1/F), Proved: (**P**/R/UW1/V/D) & Proposed: (P/R/AH0)(**P**/OW1/Z/D)

End Rhyme

- If common vowel sound in stressed syllable, all prior sounds dissimilar & all following sounds similar
- Shame: (SH/**EY1**/M) & Blame: (BL/**EY1**/M)

Metatation: Stroke Clustering & Shape Recognition



Metatation: Stroke Clustering & Shape Recognition

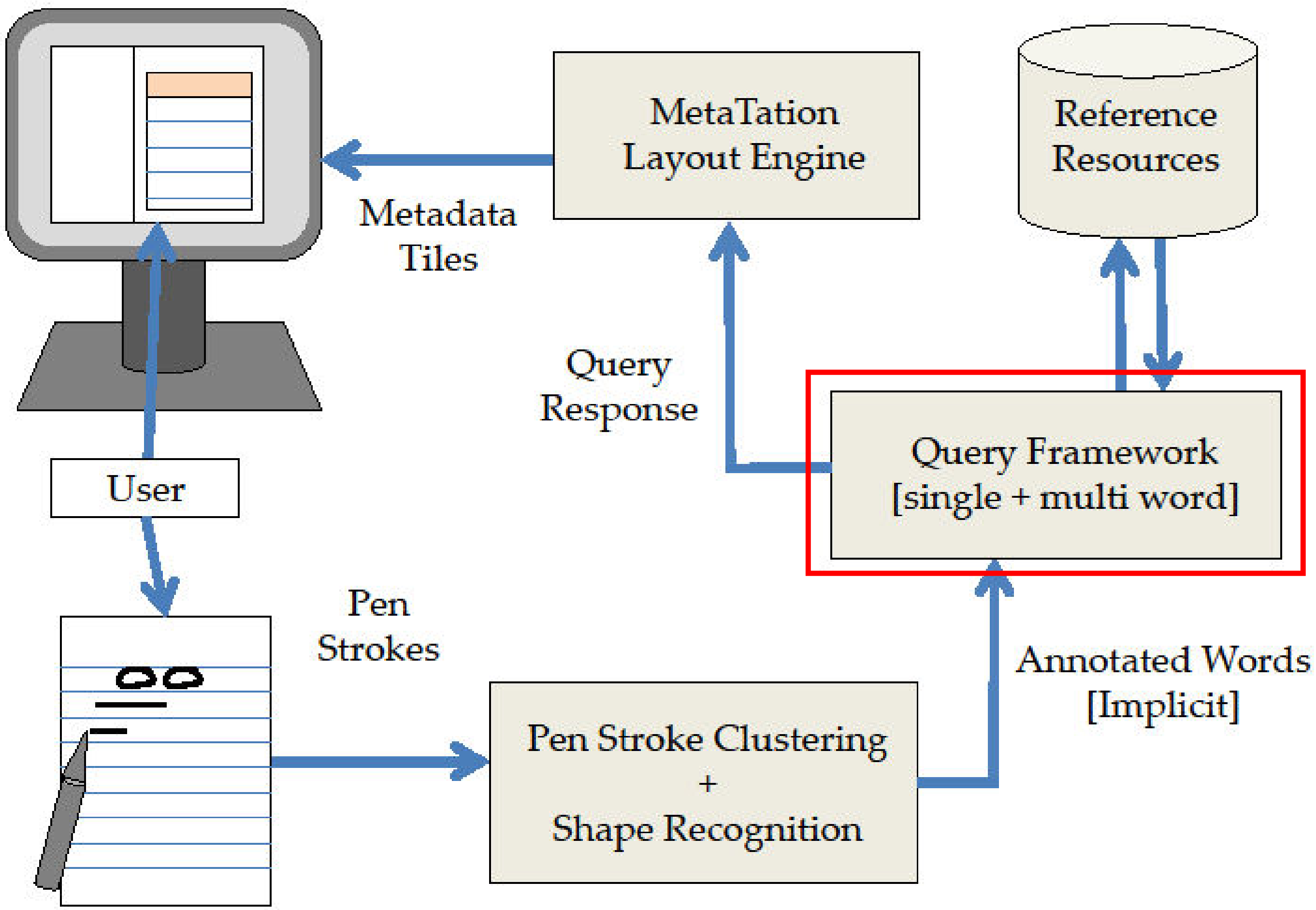
Stroke clustering

- Hierarchical agglomerative clustering
- Spatiotemporal distance between pen strokes as distance metric for the clustering process (CO characteristics)

Shape Recognition

- Geometric recognizer for detecting <u>underlines</u>, ellipses and connectors

Metatation: Query Framework



Metatation: Query Framework

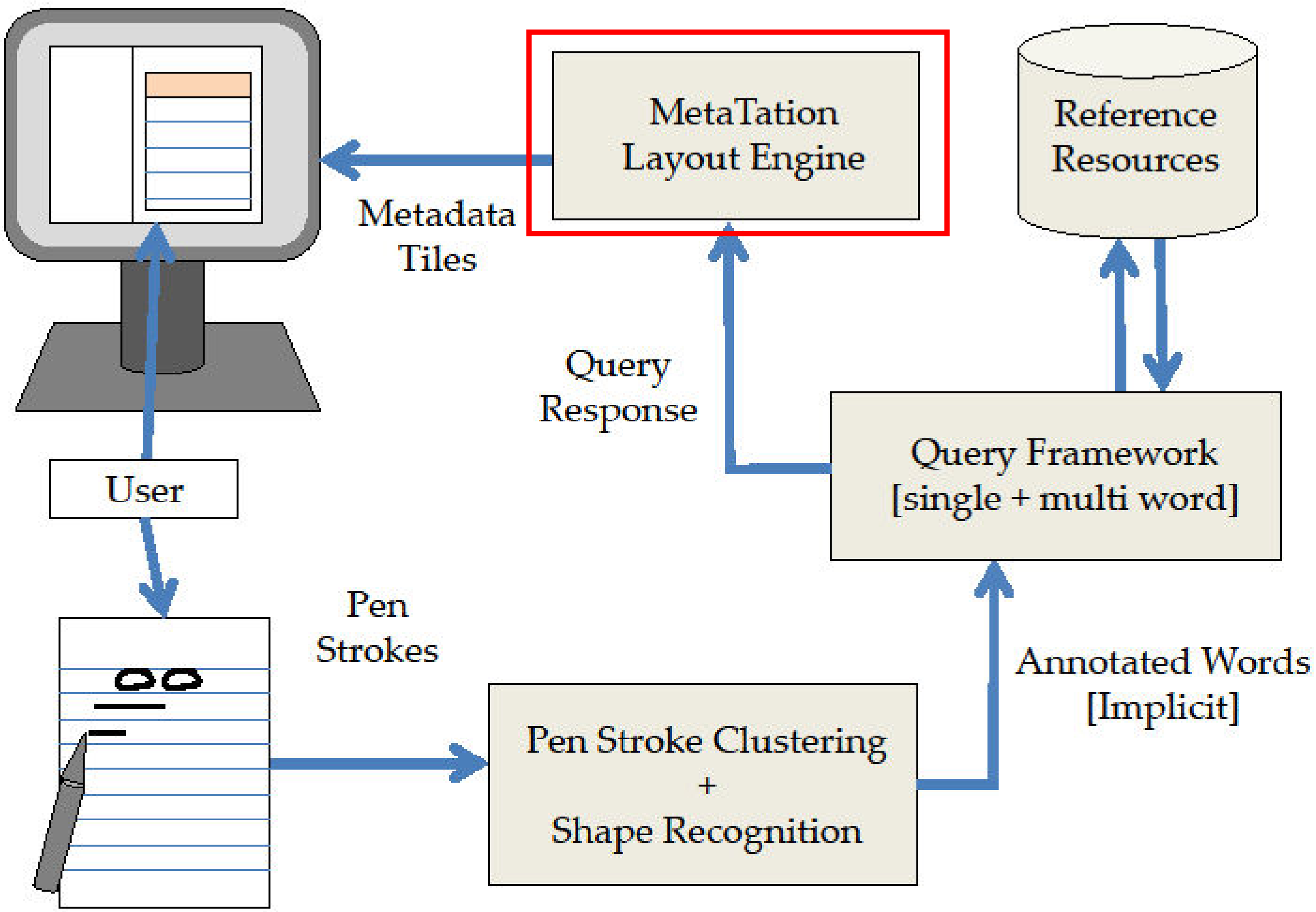
Semantic Relations

- Word details, Synonyms, Antonyms
- Source: Merriam Webster Dictionary & Thesaurus

Phonetic Relations

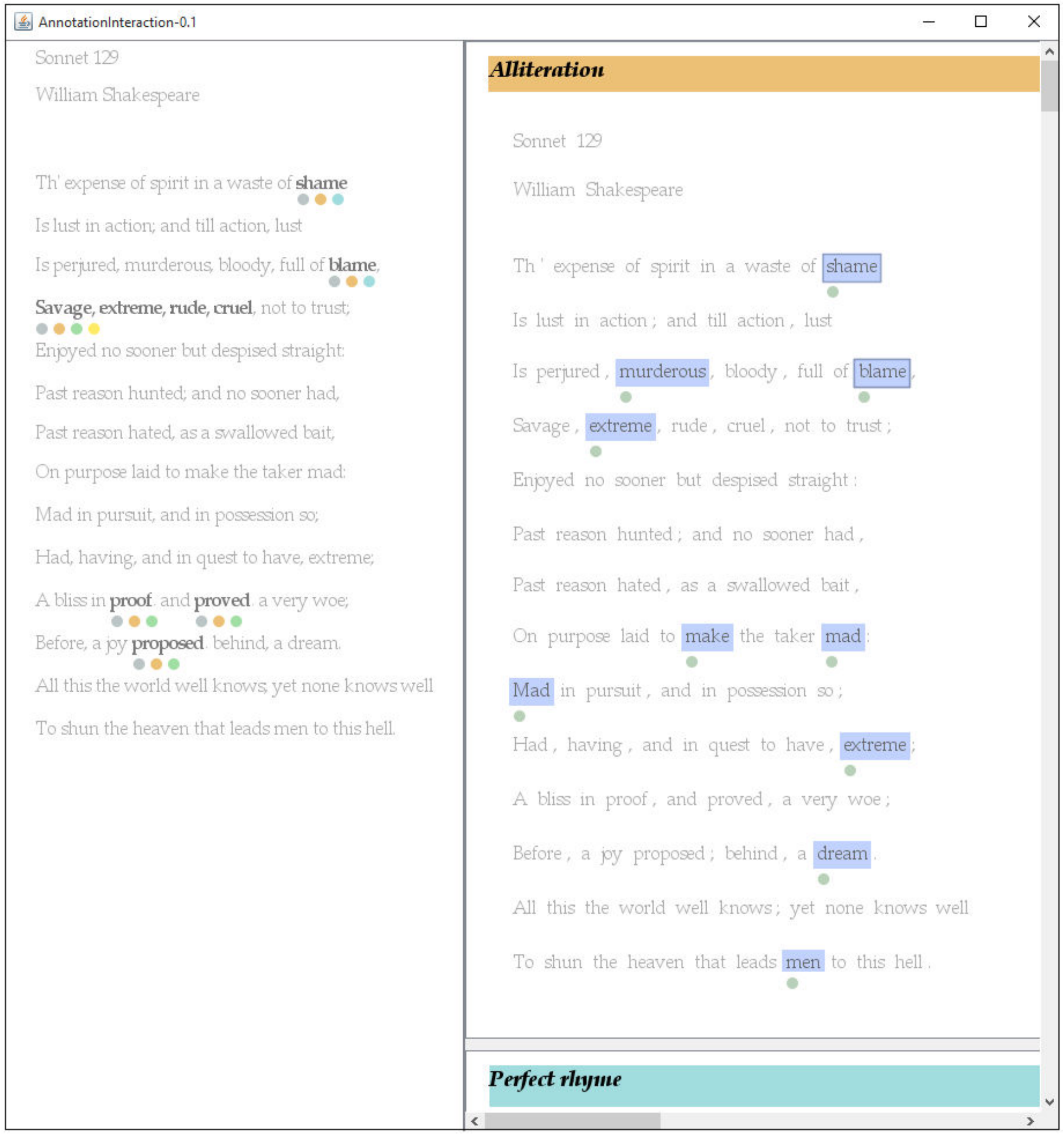
- Assonance, Consonance, Alliteration, End rhyme
- Sources:
  - CMU Phoneme Dictionary (in NLTK),
  - Broken into syllables with algorithm in "On the syllabification of phonemes" by Bartlett et al. (Proc. HLT-ACL 2009)

Metatation: Interface



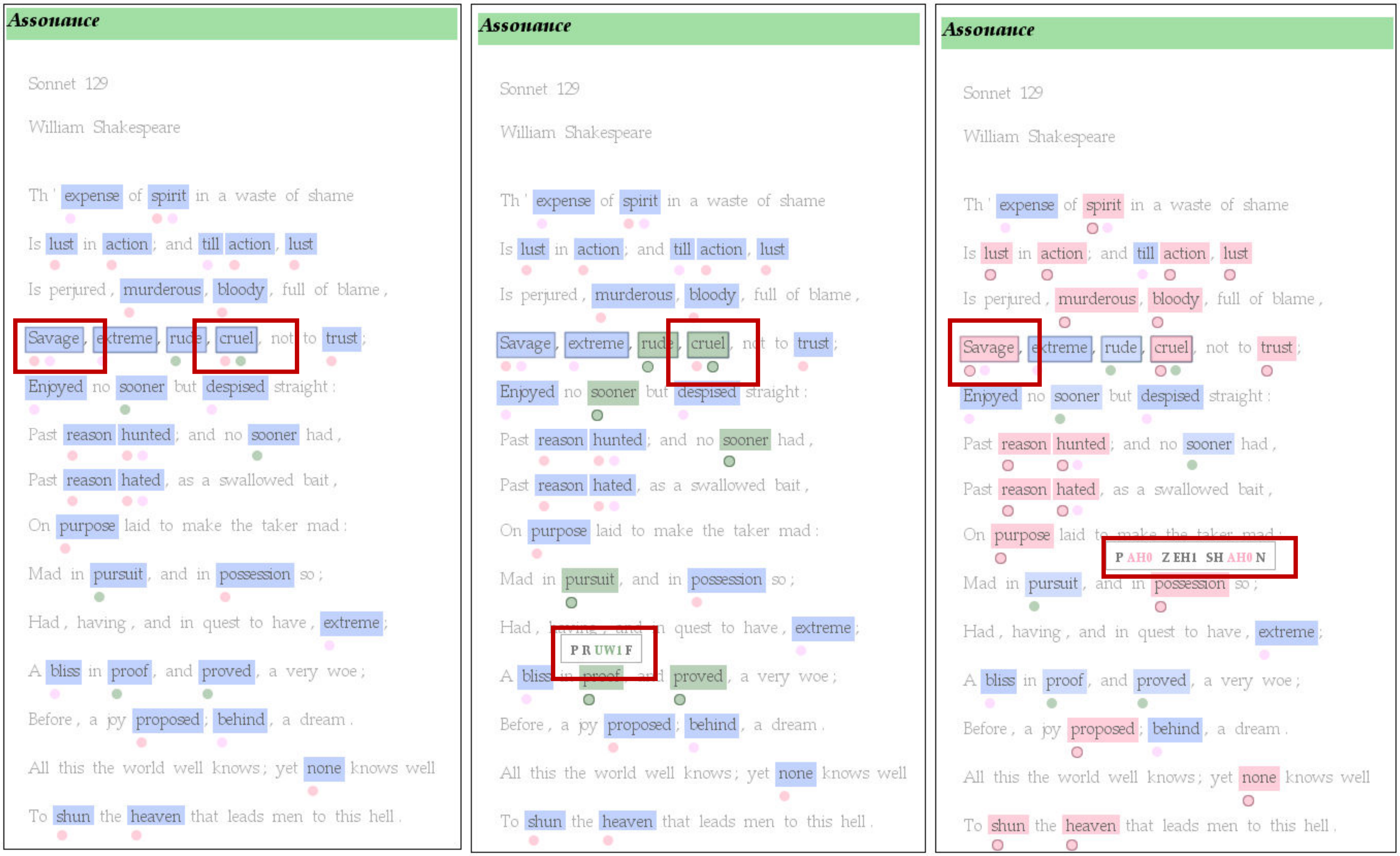
Metatation: Interface

Worksheet Viewer Panel

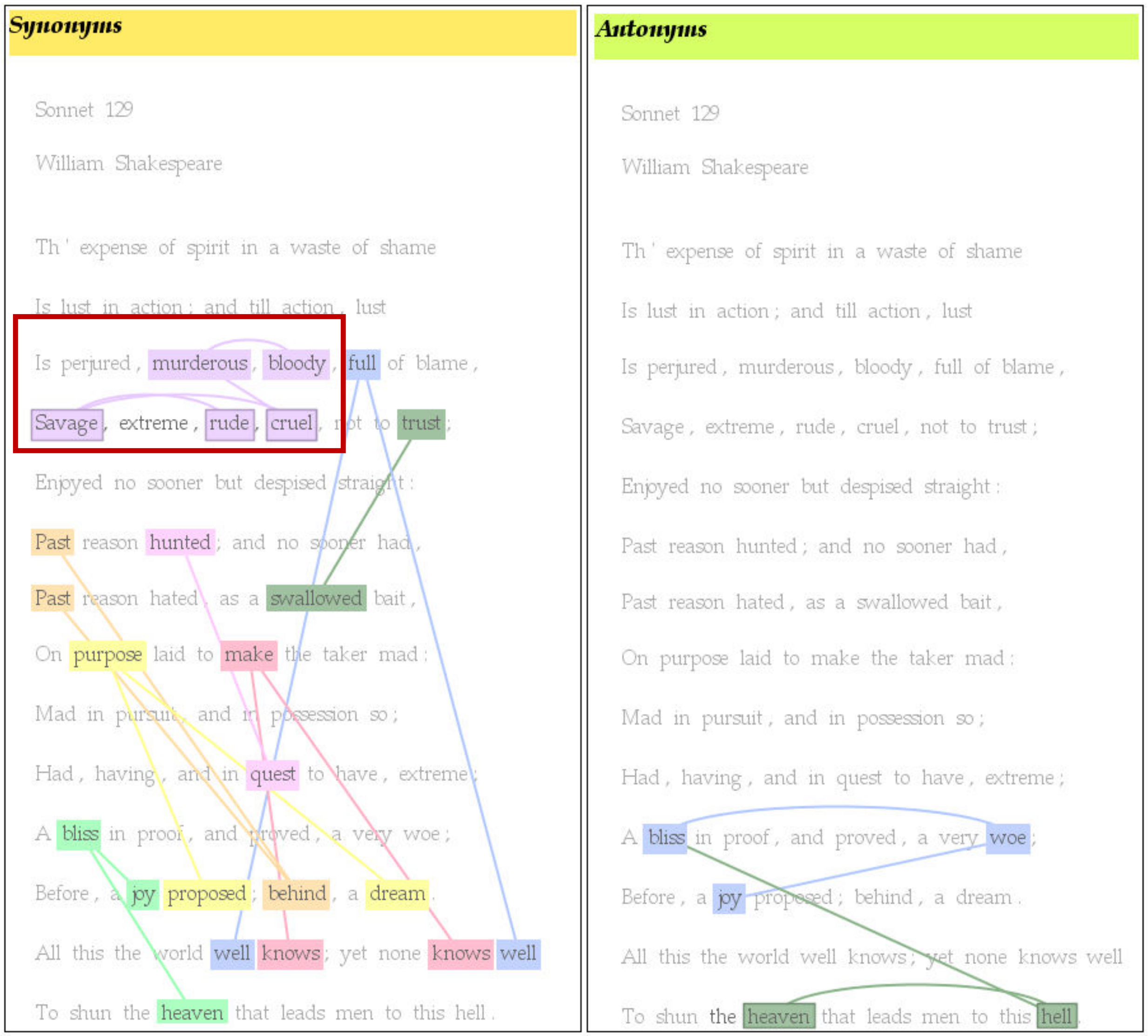


Metadata Tile Stream Panel

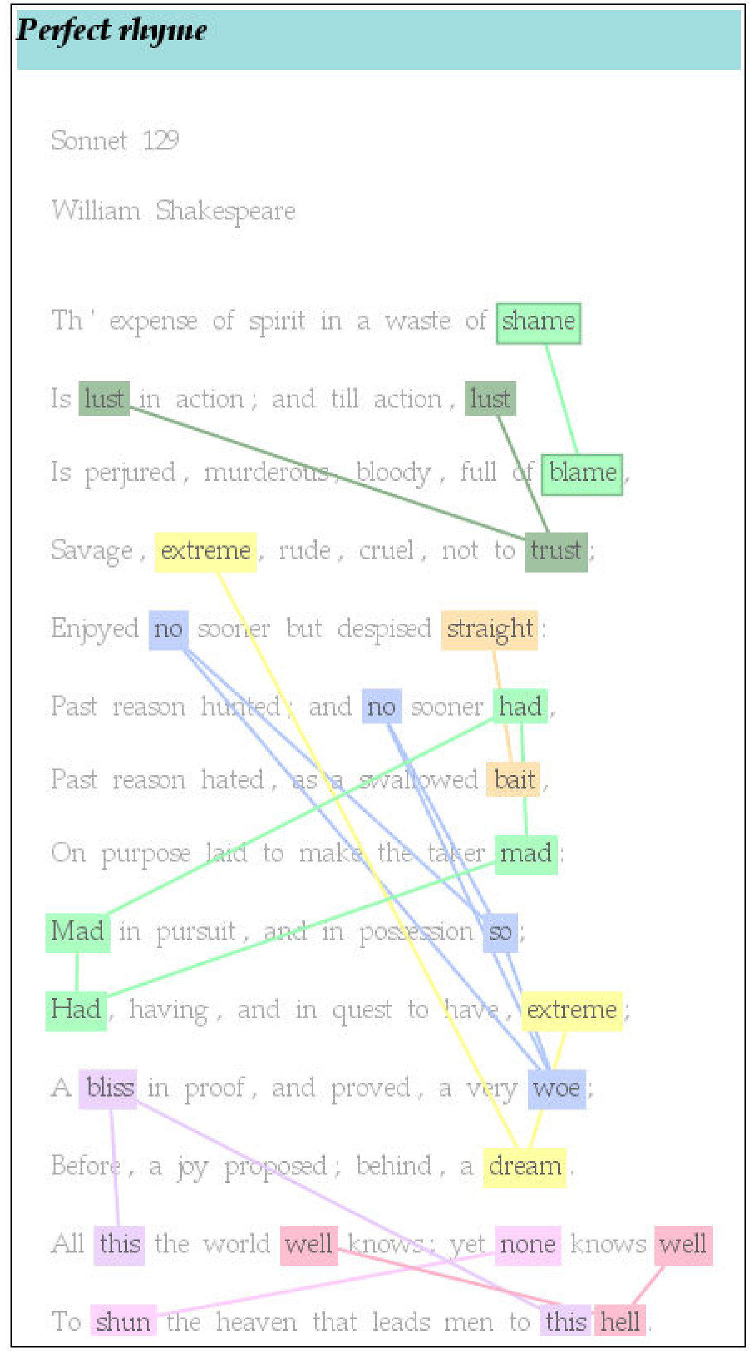
Assonance, Consonance, Alliteration Metadata Tile



Synonyms & Antonyms Metadata Tile



End Rhyme Metadata Tile



Word Details Metadata Tile

shame

Origin

Middle English, from Old English [scamu;] akin to Old High German [scama] shame

Pronuciations

(SH.EY1.M)

Noun

First recorded use

before 12th century

Senses

Sense 1

a painful emotion caused by consciousness of guilt, shortcoming, or impropriety

the susceptibility to such emotion
Usage examples: have you no [shame]?

Sense 2

a condition of humiliating disgrace or disrepute :[ignominy]
Usage examples: the [shame] of being arrested

Sense 3

something that brings censure or reproach also something to be regretted :[pity]
Usage examples: it's a [shame] you can't go

a cause of feeling shame

Verb

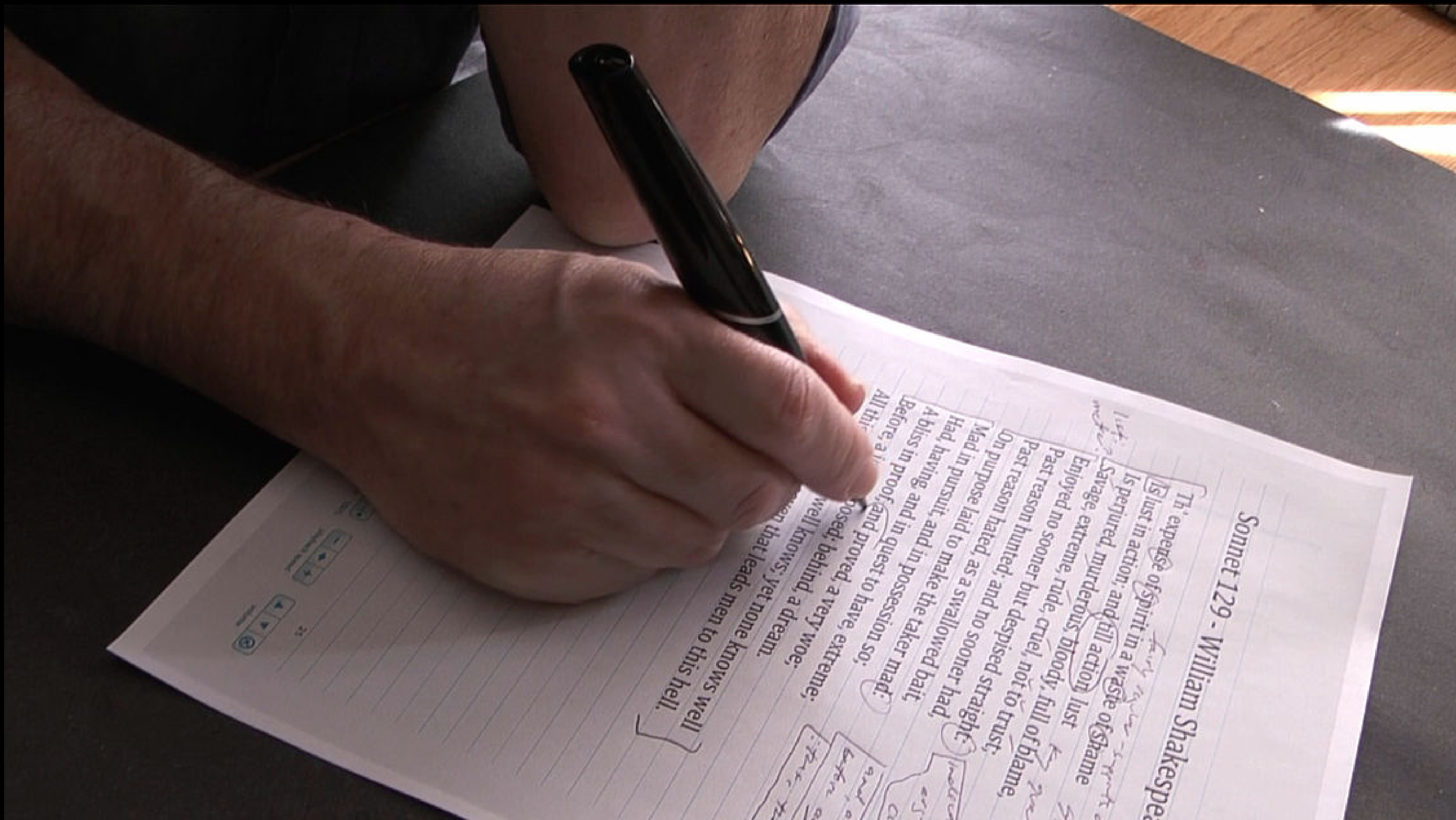
First recorded use

13th century

Senses

Sense 1

to bring shame to :[disgrace]
Usage examples: [shamed] the family name


Sonnet 129 - William Shakespea
Th' expense of spirit in a waste of shame
Is lust in action; and till action, lust
Is perjured, murderous, bloody, full of blame
Savage, extreme, rude, cruel, not to trust,
Enjoyed no sooner but despised straight,
Past reason hunted; and no sooner had
Past reason hated, as a swallowed bait,
On purpose laid to make the taker mad;
Mad in pursuit, and in possession so;
Had, having, and in quest to have, extreme;
A bliss in proof, and proved, a very woe;
Before, a
osed; behind, a dream.
All thi
well knows; yet none knows well
en that leads men to this hell.

Preliminary Evaluation

Evaluation Design

- 2 of the domain experts from the previous study explored the tool as they performed a reading of an assigned poem

Qualitative Results

- 2 different modes of use of MetaTation based on when the reader chooses to interact with the system
- MetaTation for teaching vs. research based analysis

Contributions

- Results of an observational study of poetry critics
- Design guidelines for the development of digital tools for supporting linguistic inquiry
- Design and implementation of Metatation
- Results of preliminary expert review of our tool

VisArgue

Interdisciplinary Research Case Study

Visual Analytics
(Daniel Keim)

Linguistics
(Miriam Butt)

Political science
(Katharina Holzinger)

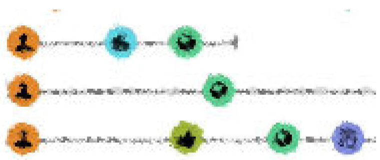
VisArgue

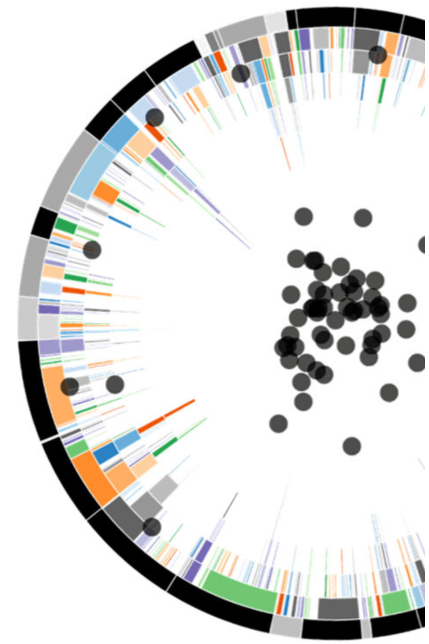
Visualizing the linguistic structure
of deliberative communication

Katharina Holzinger

Valentin Gold

Universität Konstanz





Miriam Butt

Annette Hautli-Janisz

Tina Bögel

Maike Müller

Daniel Keim

Menna El-Assady

Wolfgang Jentner

Rita Sevastjanova

Carmela Acevedo

Visual Analytics Linguistics Political science

Goal: To develop an instrument
for the automatic measurement
of the deliberative quality of communication

“Deliberation is a **communicative process** that aims at **taking a decision** (or recommendation) on collectively **binding rules** or public projects. The substantive goal is to **achieve the common good** and universality of rules.”

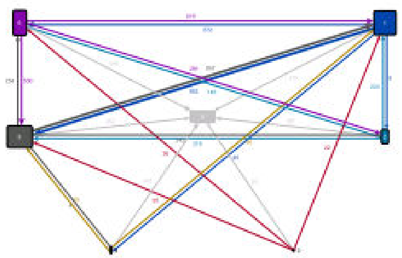
Visual Analytics | Linguistics | Political science

Goals:

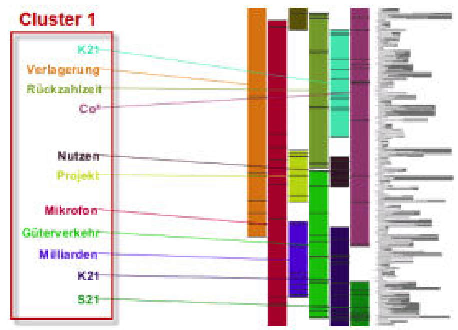
- Deep automatic **syntactic parsing** of the utterances to identify deliberative patterns.
- Define negotiation vs. argumentation through the analysis of the **speaker stands** and **attitude**.
- Identification of different roles of utterances by the usage of **connectors** and **function words**.

Visual Analytics | Linguistics | Political science

Participation

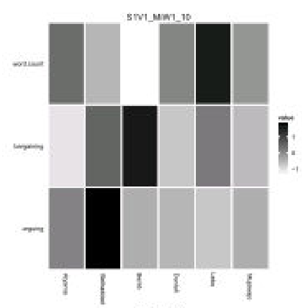


Content

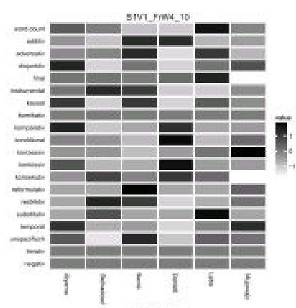


Sentiments

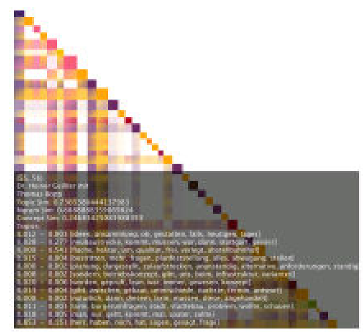
Intentions



Argumentativeness



Interactions



(. . . and some more!)

Visual Analytics | Linguistics | Political science

Research questions

- What factors make deliberative argumentation successful?
- Can we detect these factors both by shallow statistical and deep linguistic approaches?
- How can visual analytics support the analysis of deliberation in large corpora?

Content Overview using Lexical Episode Plots

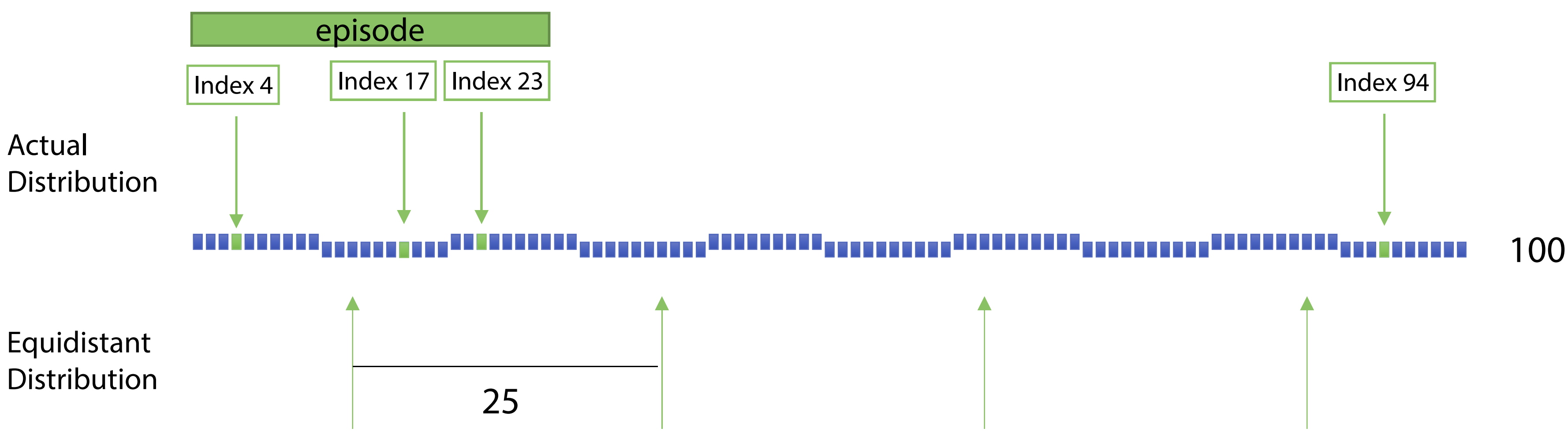
First Debate

Second Debate

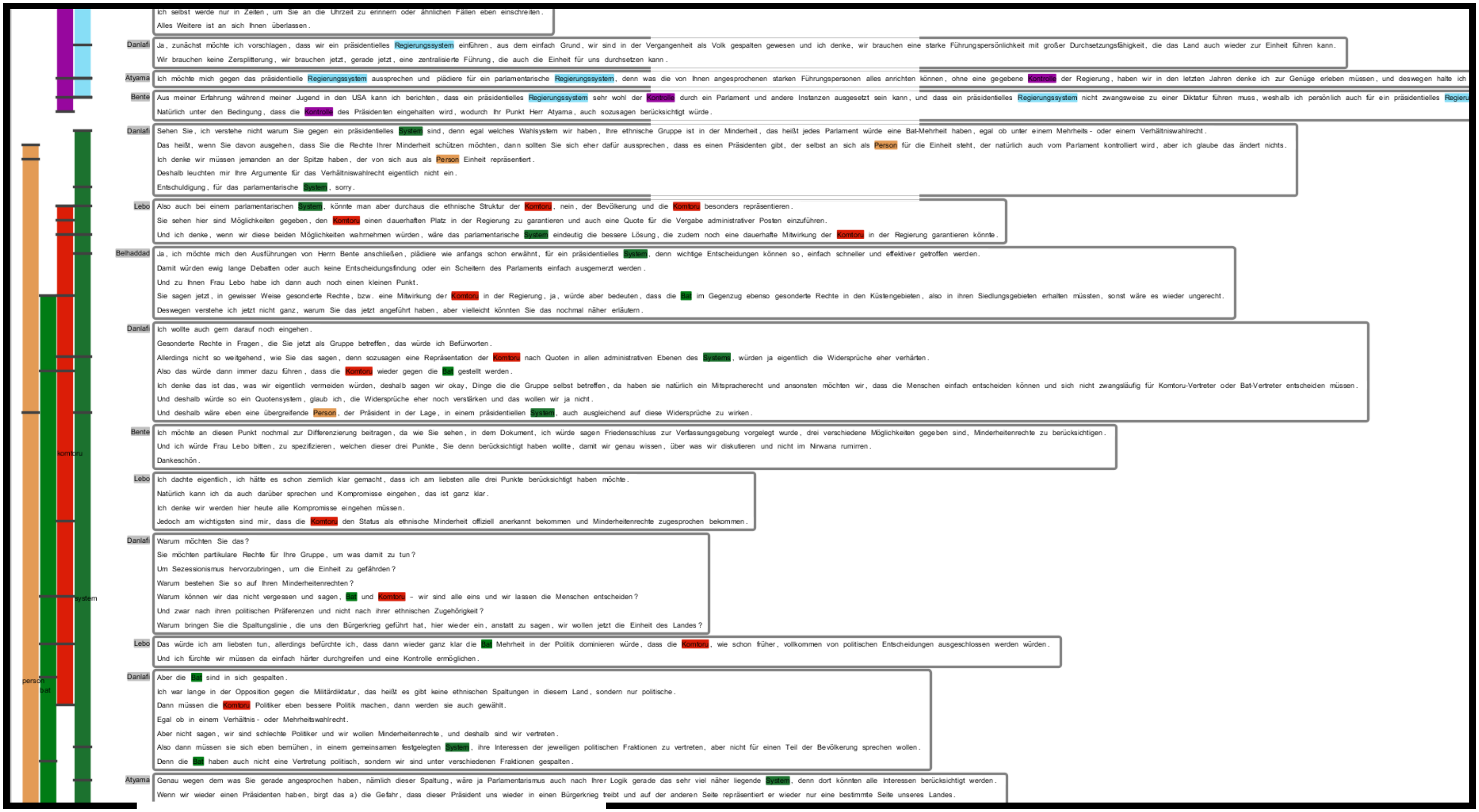
Third Debate

Lexical Episodes

= portion within the word sequence of a corpus where a certain word appears more densely than expected from its frequency in the whole text.

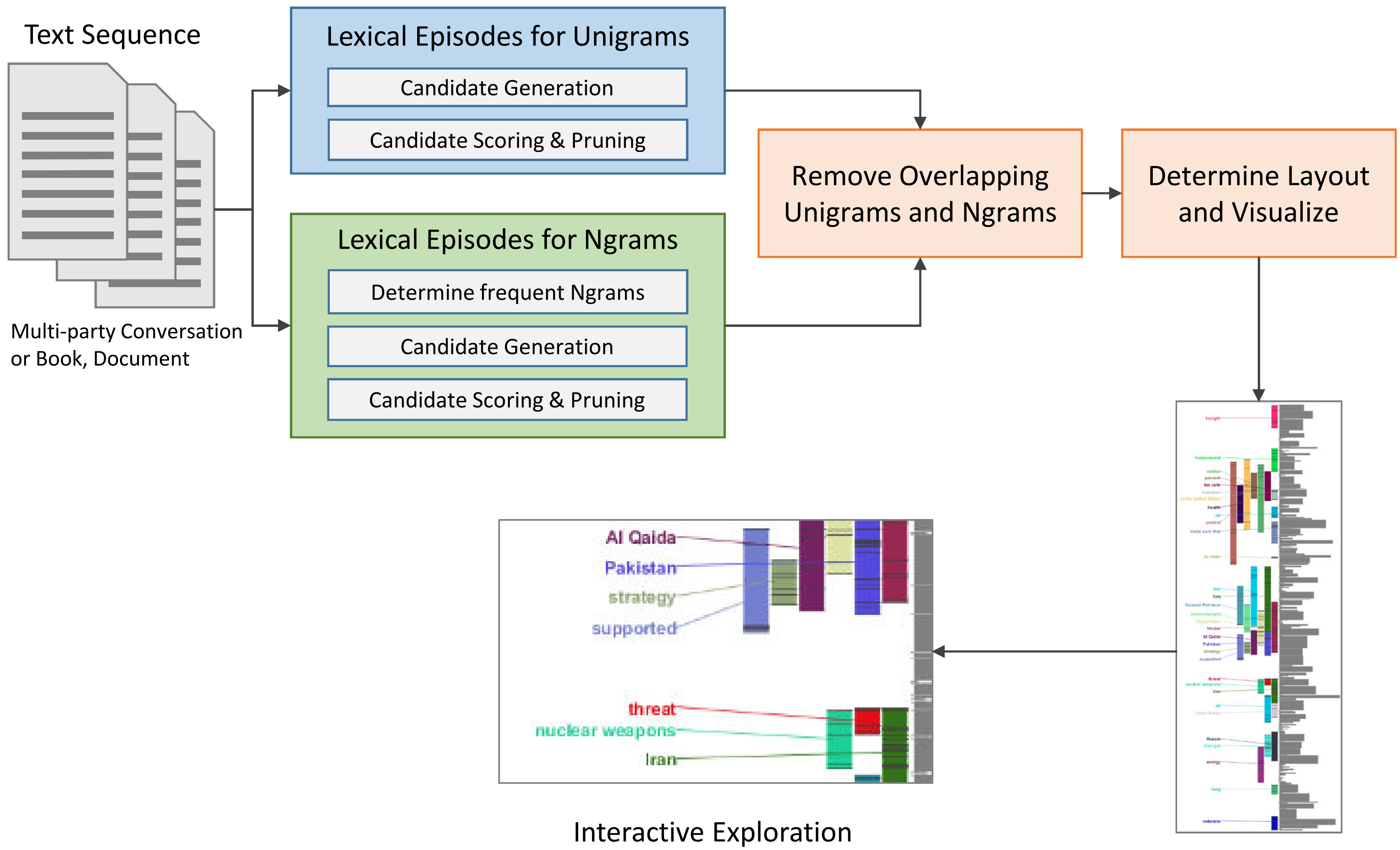


Lexical Episodes Plots

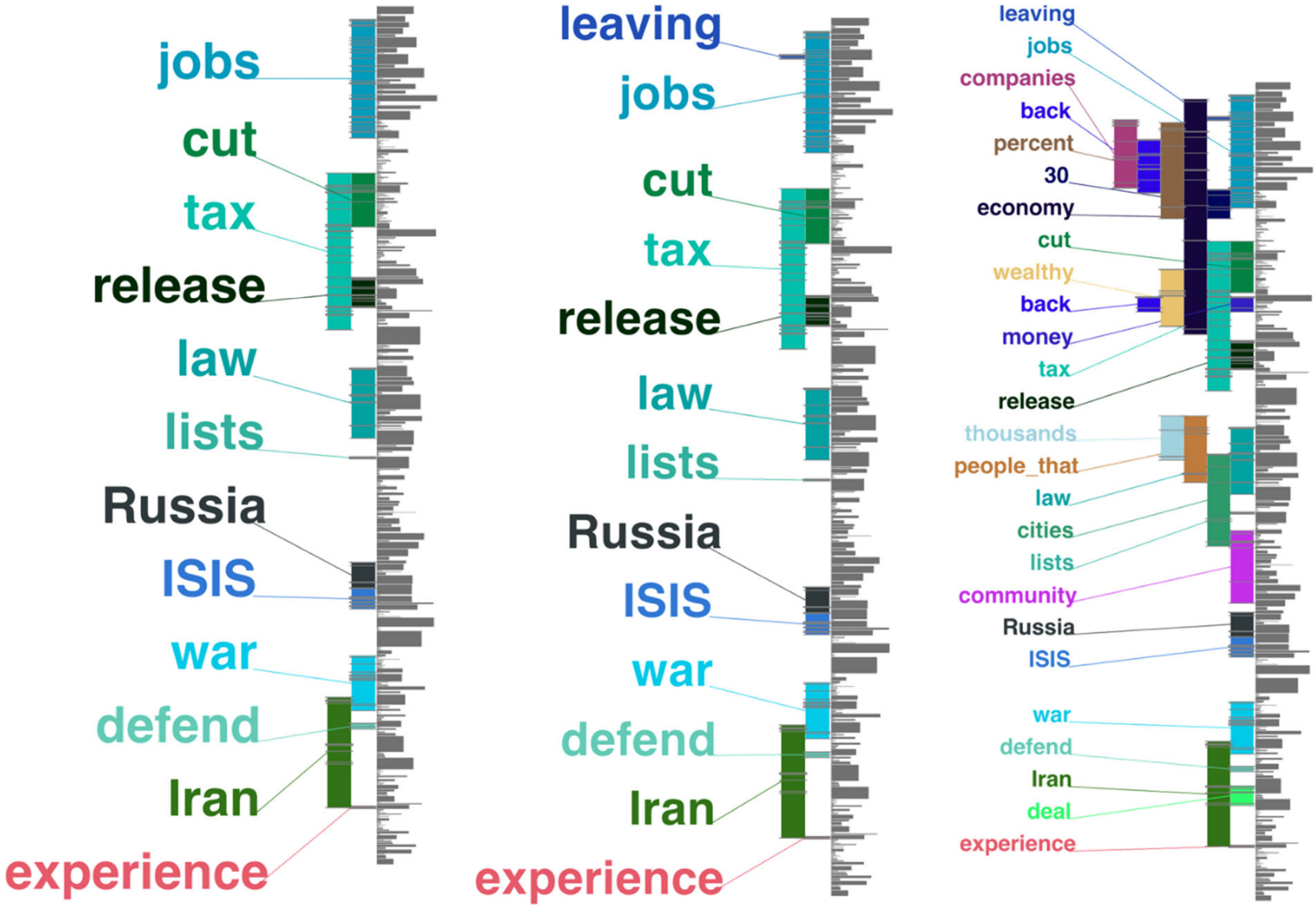


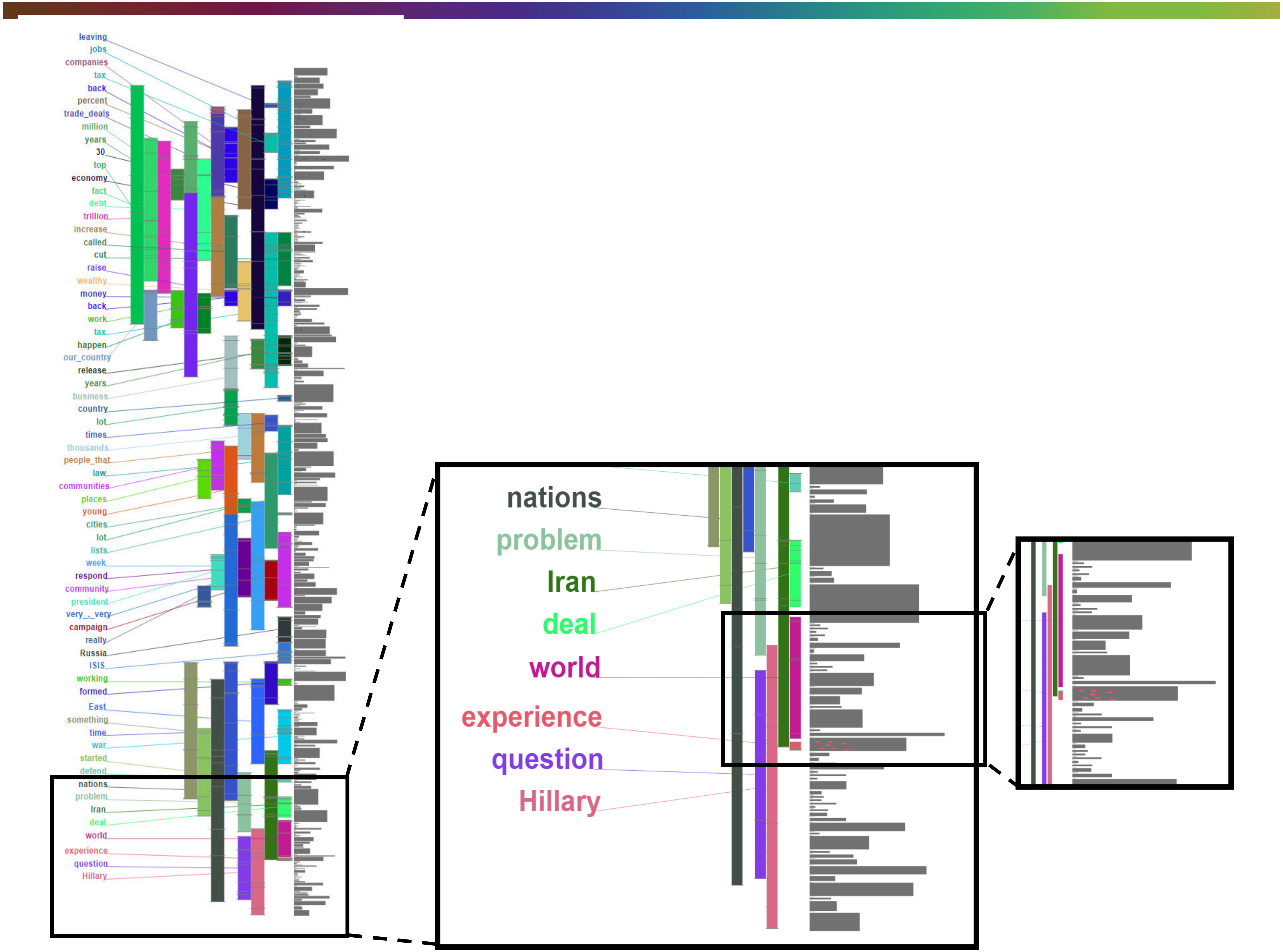
Close Reading

Lexical Episodes - Processing Pipeline

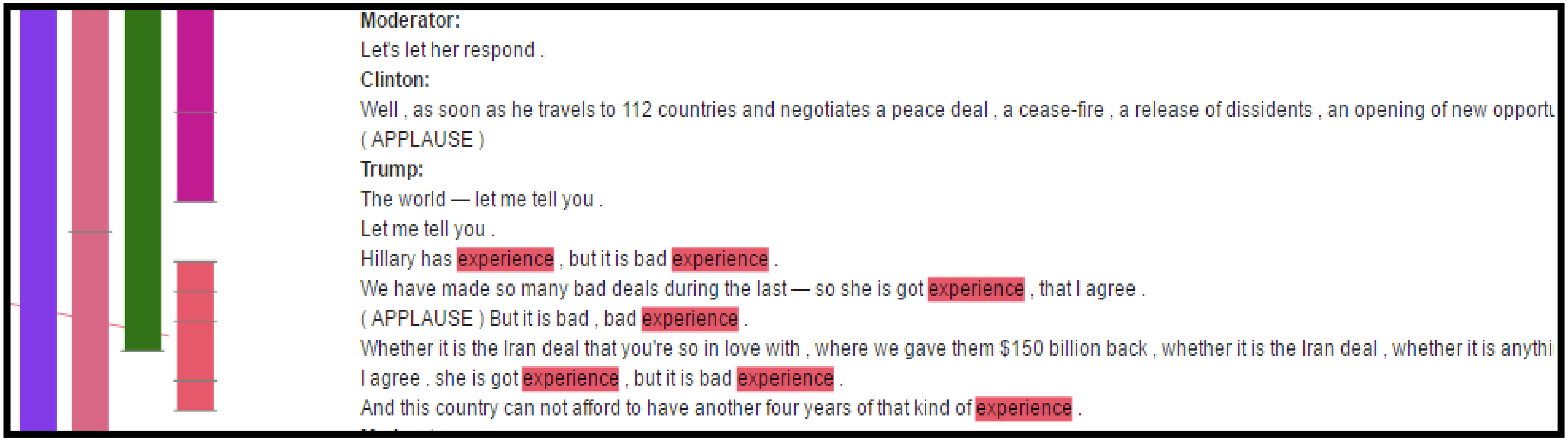


First Debate

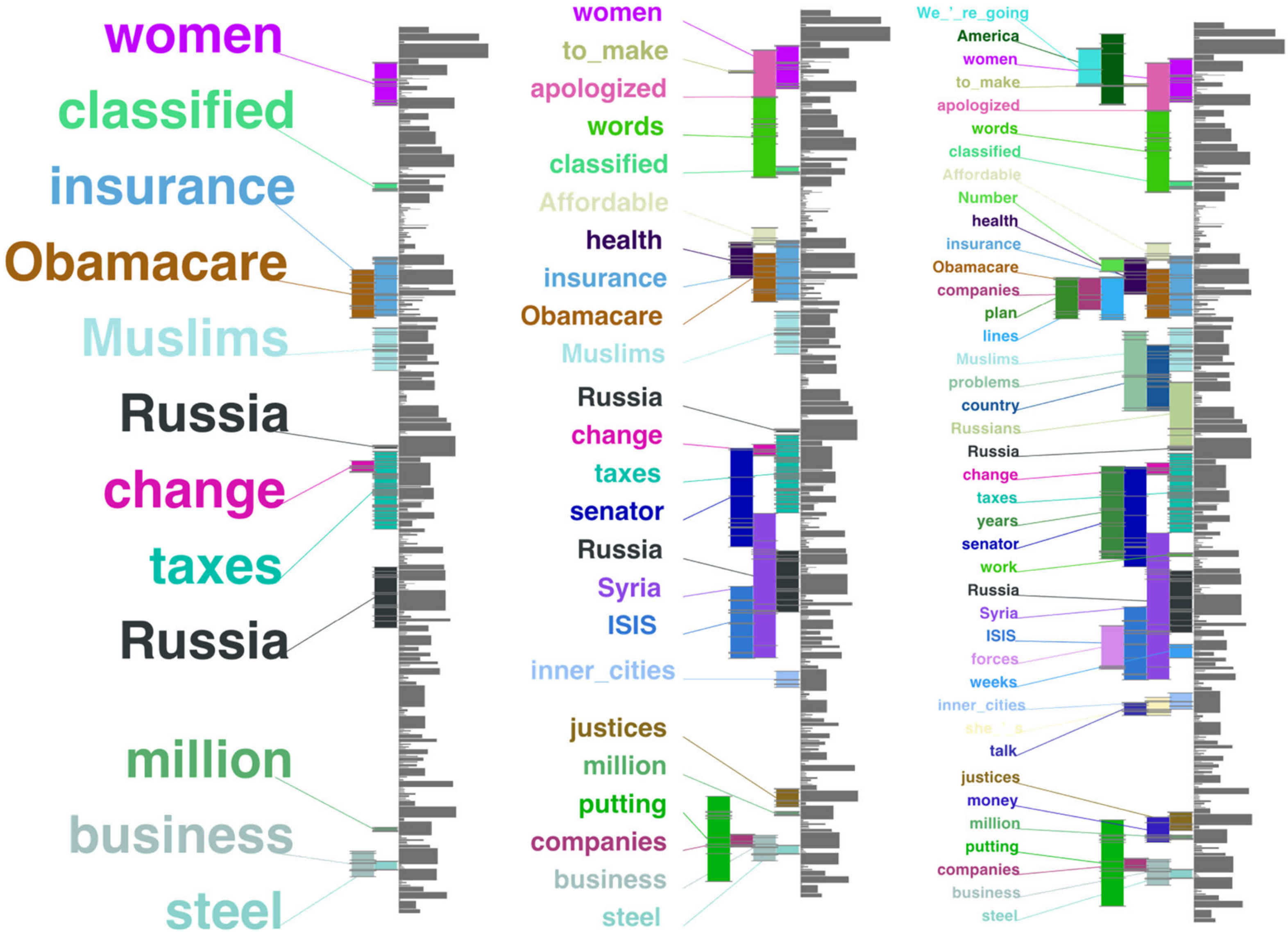


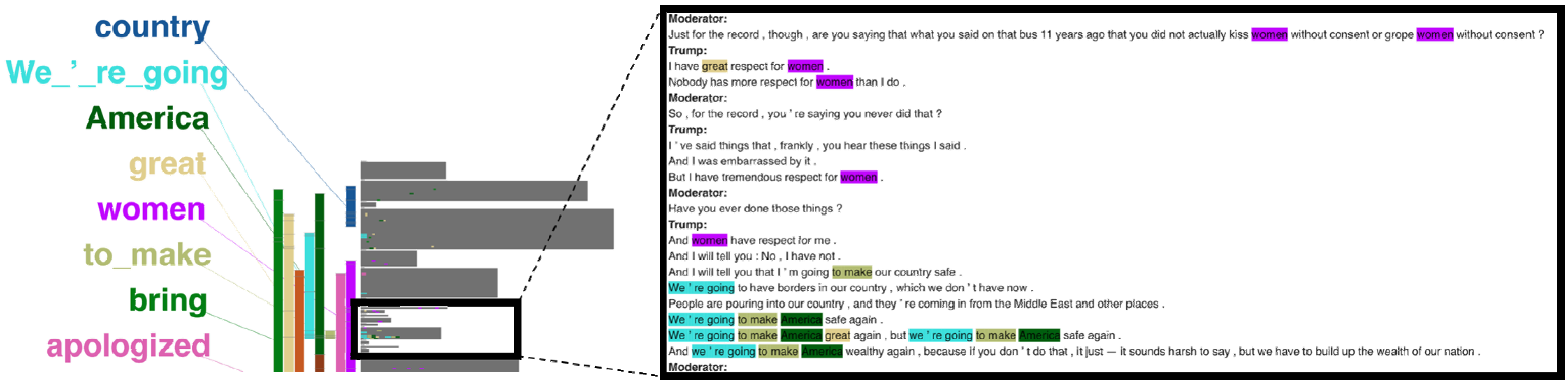


leaving
jobs
companies
tax
back
percent
trade_deals
million
years
30
top
economy
fact
debt
trillion
increase
called
cut
raise
wealthy
money
back
work
tax
happen
our_country
release
years
business
country
lot
times
thousands
people_that
law
communities
places
young
cities
lot
lists
week
respond
community
president
very_,_very
campaign
really
Russia
ISIS
working
formed
East
something
time
war
started
defend
nations
problem
Iran
deal
world
experience
question
Hillary


Moderator:
Let's let her respond .
Clinton:
Well , as soon as he travels to 112 countries and negotiates a peace deal , a cease-fire , a release of dissidents , an opening of new opportu
(APPLAUSE)
Trump:
The world — let me tell you .
Let me tell you .
Hillary has experience , but it is bad experience .
We have made so many bad deals during the last — so she is got experience , that I agree .
(APPLAUSE) But it is bad , bad experience .
Whether it is the Iran deal that you're so in love with , where we gave them $150 billion back , whether it is the Iran deal , whether it is anythi
I agree . she is got experience , but it is bad experience .
And this country can not afford to have another four years of that kind of experience .

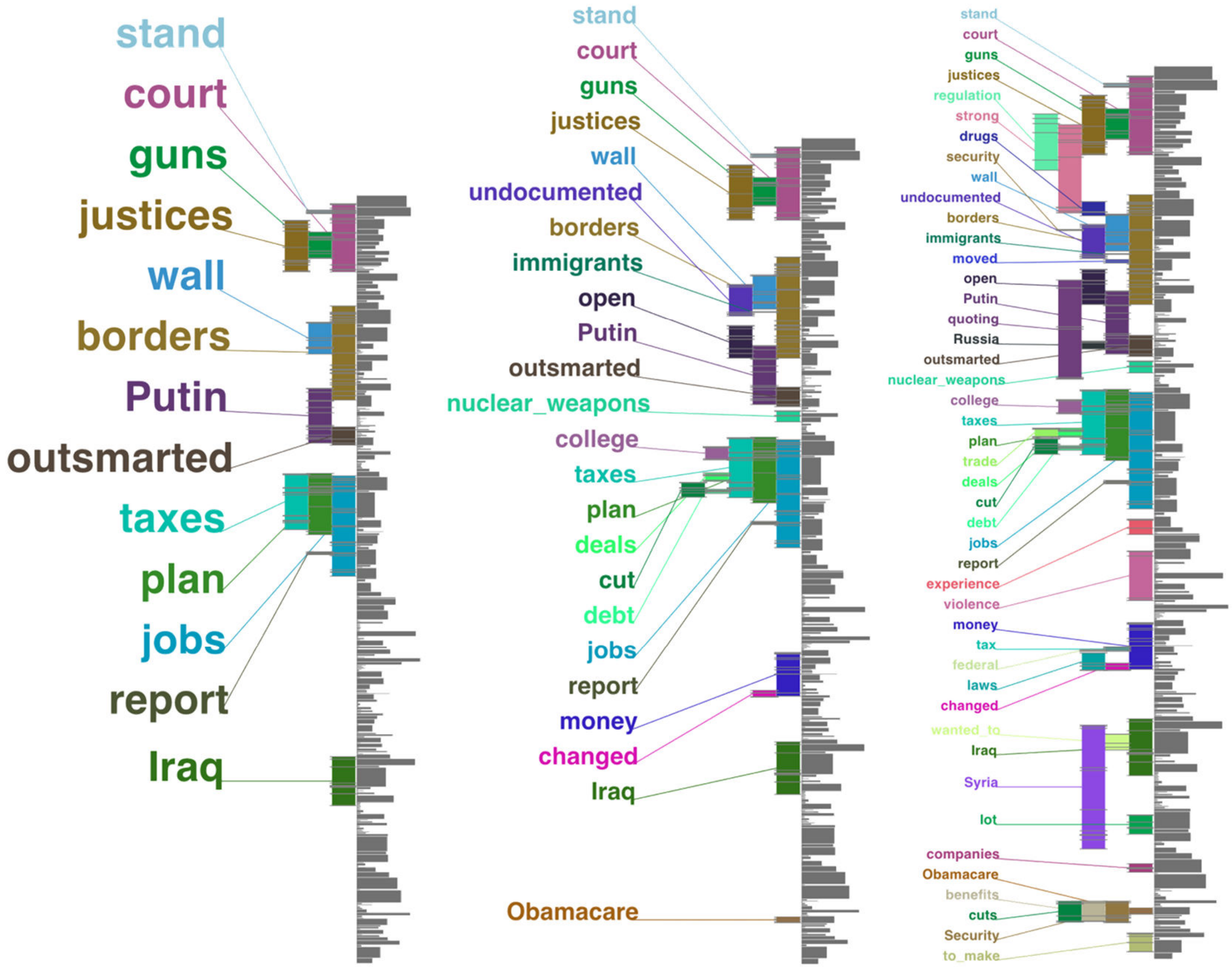
Second Debate

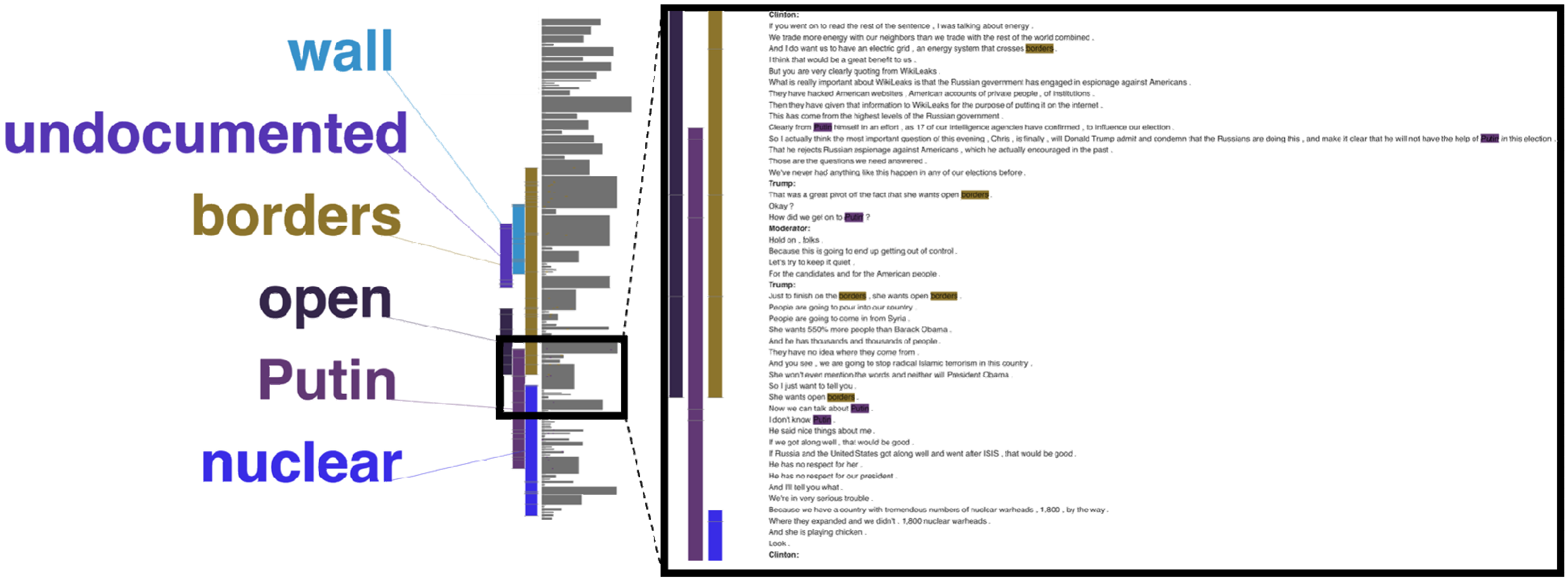




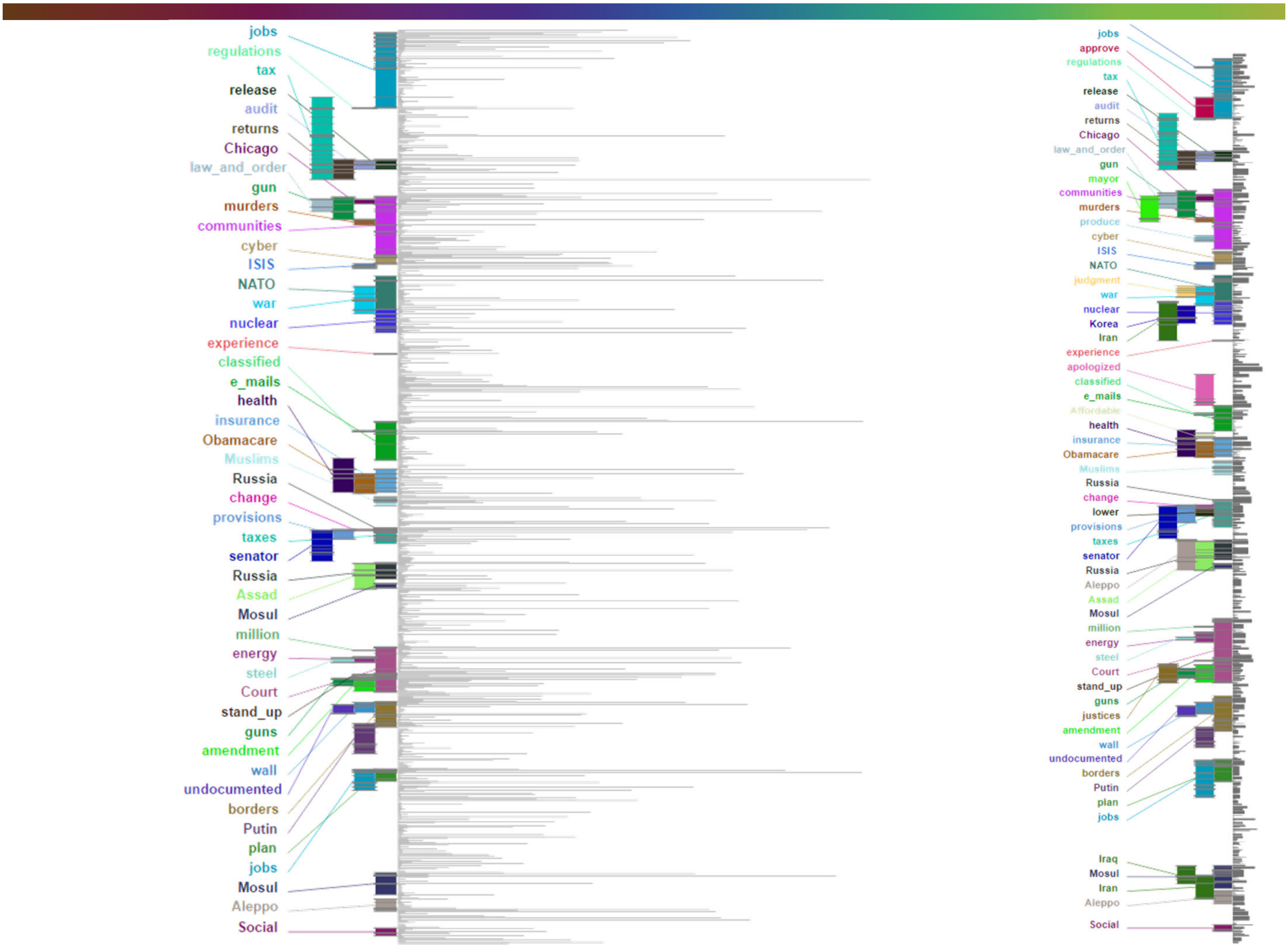
country
We_’_re_going
America
great
women
to_make
bring
apologized
Moderator:
Just for the record , though , are you saying that what you said on that bus 11 years ago that you did not actually kiss women without consent or grope women without consent ?
Trump:
I have great respect for women .
Nobody has more respect for women than I do .
Moderator:
So , for the record , you ' re saying you never did that ?
Trump:
I ' ve said things that , frankly , you hear these things I said .
And I was embarrassed by it .
But I have tremendous respect for women .
Moderator:
Have you ever done those things ?
Trump:
And women have respect for me .
And I will tell you : No , I have not .
And I will tell you that I ' m going to make our country safe .
We ' re going to have borders in our country , which we don ' t have now .
People are pouring into our country , and they ' re coming in from the Middle East and other places .
We ' re going to make America safe again .
We ' re going to make America great again , but we ' re going to make America safe again .
And we ' re going to make America wealthy again , because if you don ' t do that , it just — it sounds harsh to say , but we have to build up the wealth of our nation .
Moderator:

Third Debate



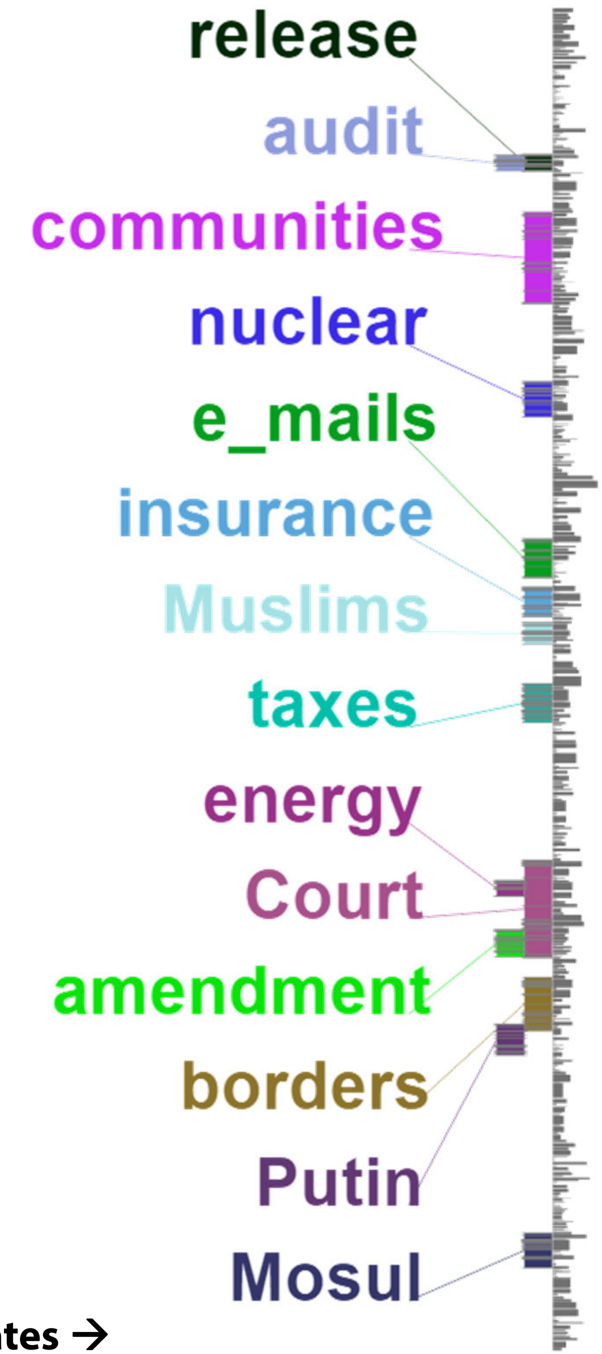


wall
undocumented
borders
open
Putin
nuclear
Clinton:
If you went on to read the rest of the sentence , I was talking about energy .
We trade more energy with our neighbors than we trade with the rest of the world combined .
And I do want us to have an electric grid , an energy system that crosses borders .
I think that would be a great benefit to us .
But you are very clearly quoting from WikiLeaks .
What is really important about WikiLeaks is that the Russian government has engaged in espionage against Americans .
They have hacked American websites , American accounts of private people , of institutions .
Then they have given that information to WikiLeaks for the purpose of putting it on the internet .
This has come from the highest levels of the Russian government .
Clearly from Putin himself in an effort , as 17 of our intelligence agencies have confirmed , to influence our election .
So I actually think the most important question of this evening , Chris , is finally , will Donald Trump admit and condemn that the Russians are doing this , and make it clear that he will not have the help of Putin in this election .
That he rejects Russian espionage against Americans , which he actually encouraged in the past .
Those are the questions we need answered .
We've never had anything like this happen in any of our elections before .
Trump:
That was a great pivot off the fact that she wants open borders .
Okay ?
How did we get on to Putin ?
Moderator:
Hold on , folks .
Because this is going to end up getting out of control .
Let's try to keep it quiet .
For the candidates and for the American people .
Trump:
Just to finish on the borders , she wants open borders .
People are going to pour into our country .
People are going to come in from Syria .
She wants 550% more people than Barack Obama .
And he has thousands and thousands of people .
They have no idea where they come from .
And you see , we are going to stop radical Islamic terrorism in this country .
She won't even mention the words and neither will President Obama .
So I just want to tell you .
She wants open borders .
Now we can talk about Putin .
I don't know Putin .
He said nice things about me .
If we got along well , that would be good .
If Russia and the United States got along well and went after ISIS , that would be good .
He has no respect for her .
He has no respect for our president .
And I'll tell you what .
We're in very serious trouble .
Because we have a country with tremendous numbers of nuclear warheads , 1,800 , by the way .
Where they expanded and we didn't . 1,800 nuclear warheads .
And she is playing chicken .
Look .
Clinton:


jobs
regulations
tax
release
audit
returns
Chicago
law_and_order
gun
murders
communities
cyber
ISIS
NATO
war
nuclear
experience
classified
e_mails
health
insurance
Obamacare
Muslims
Russia
change
provisions
taxes
senator
Russia
Assad
Mosul
million
energy
steel
Court
stand_up
guns
amendment
wall
undocumented
borders
Putin
plan
jobs
Mosul
Aleppo
Social
jobs
approve
regulations
tax
release
audit
returns
Chicago
law_and_order
gun
mayor
communities
murders
produce
cyber
ISIS
NATO
judgment
war
nuclear
Korea
Iran
experience
apologized
classified
e_mails
Affordable
health
insurance
Obamacare
Muslims
Russia
change
lower
provisions
taxes
senator
Russia
Aleppo
Assad
Mosul
million
energy
steel
Court
stand_up
guns
justices
amendment
wall
undocumented
borders
Putin
plan
jobs
Iraq
Mosul
Iran
Aleppo
Social

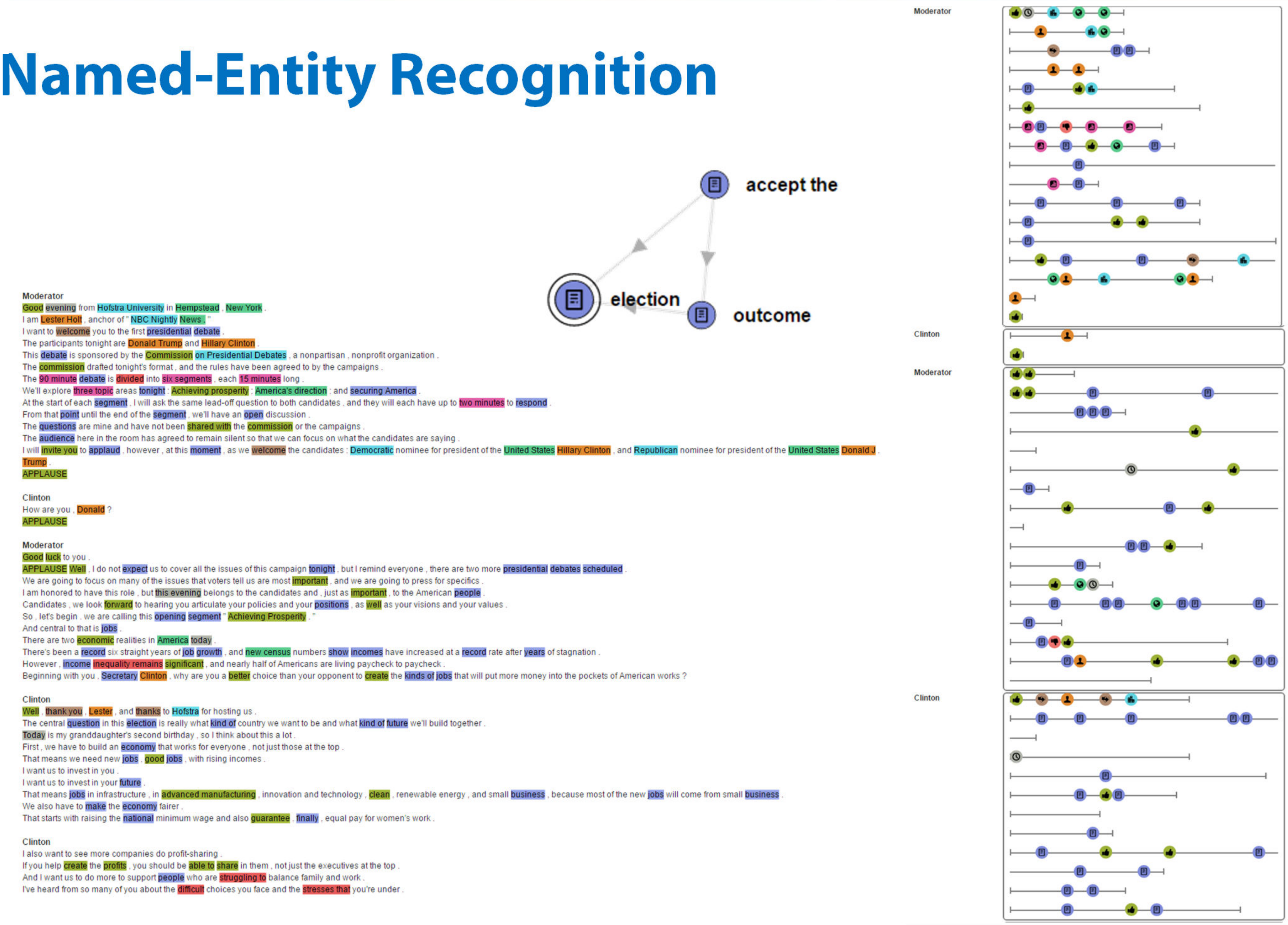
Lexical Episodes

- Generic approach for generating text overviews using lexical chaining
- Text-type and language independent
- Distant Reading vs. Close Reading
- Interactivity through Zooming and Highlighting
- Different levels of detail by steering the significance level



All Debates →

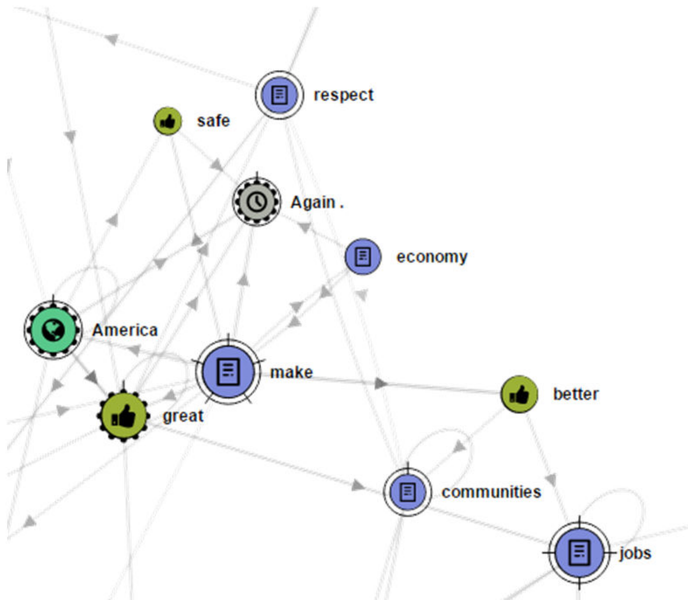
Named-Entity Recognition



Concept Relations

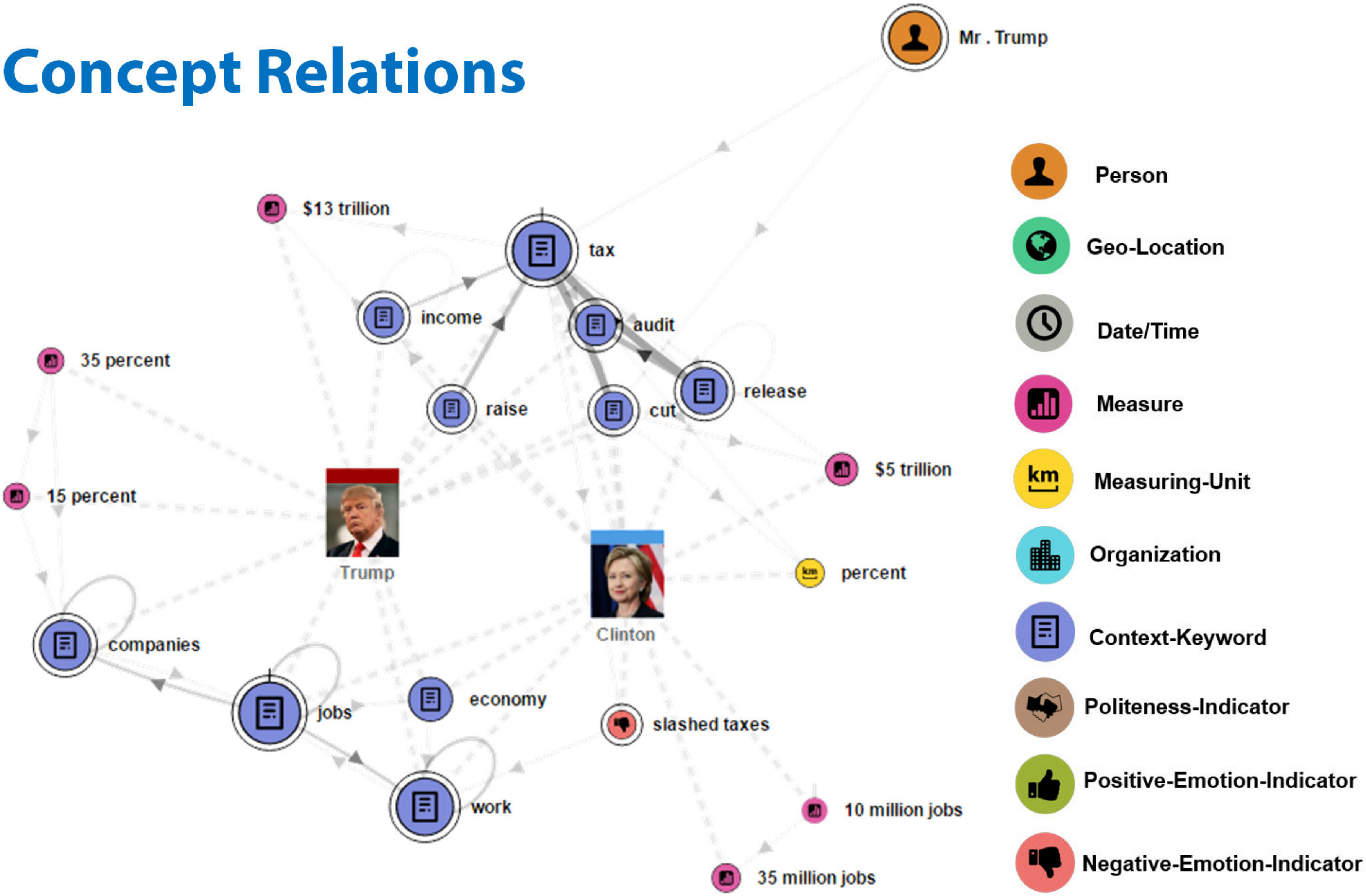
"I believe that we can cut taxes by an additional $5 trillion."

Entity-Pairs: | cut | taxes | | cut | $5 trillion | | taxes | $5 trillion |

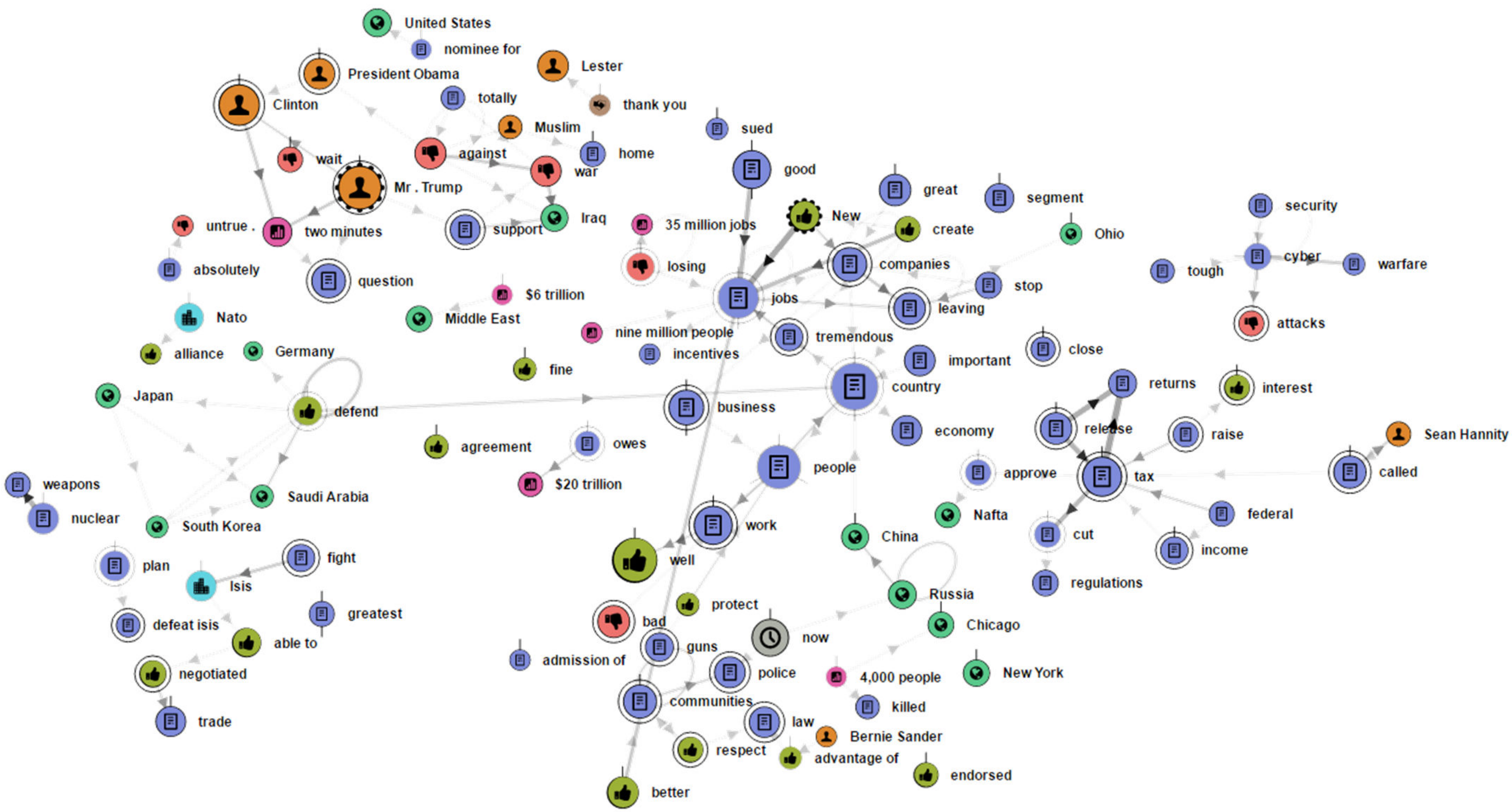


- Person
- Geo-Location
- Date/Time
- Measure
- Measuring-Unit
- Organization
- Context-Keyword
- Politeness-Indicator
- Positive-Emotion-Indicator
- Negative-Emotion-Indicator

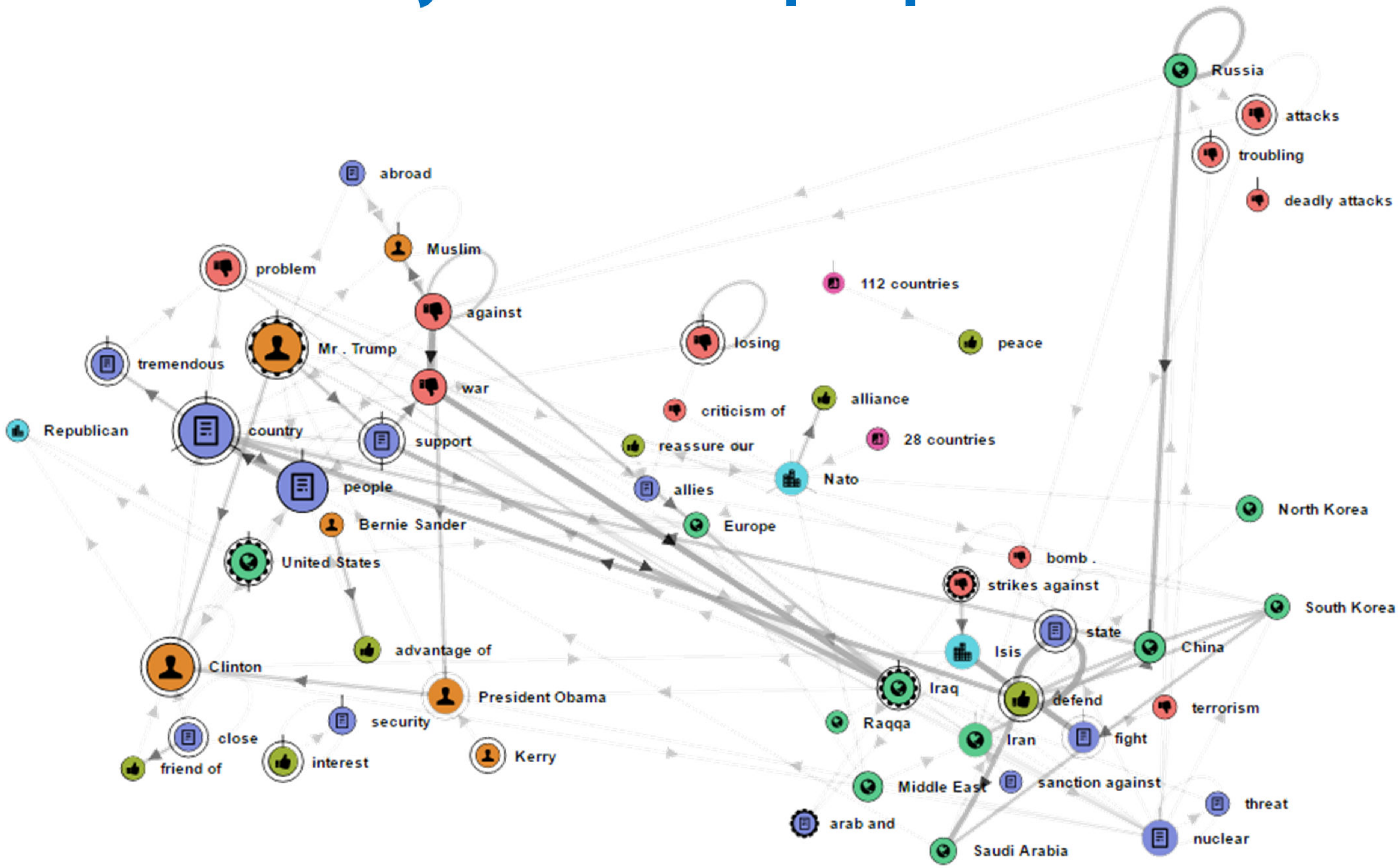
Concept Relations



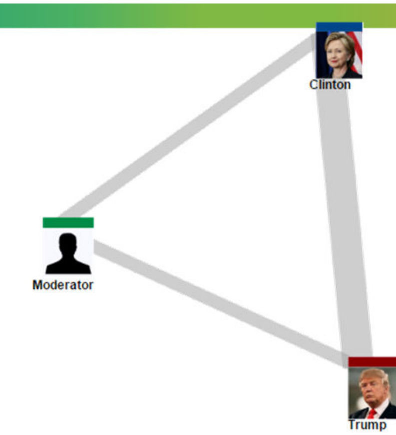
Named-Entity Relationship Exploration

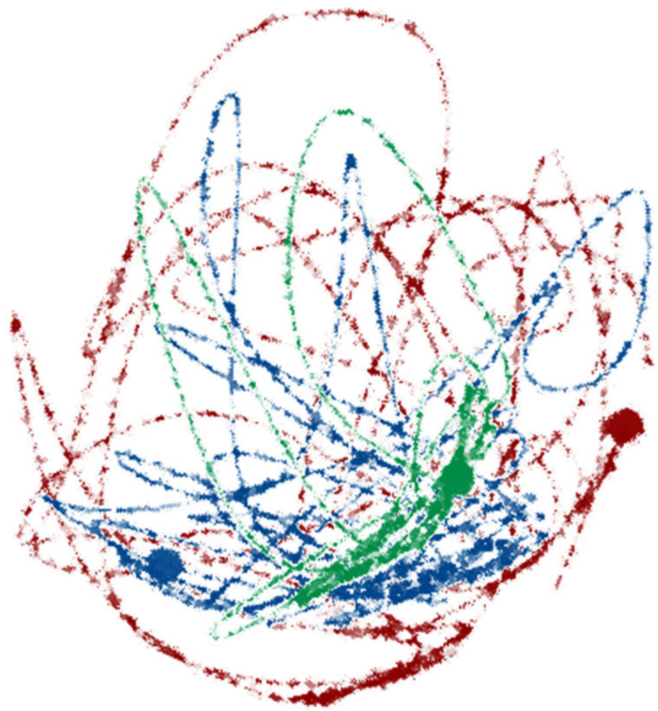


Named-Entity Relationship Exploration

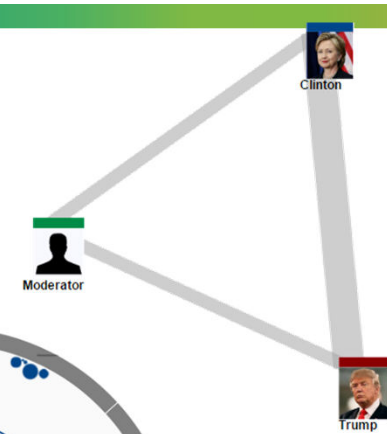


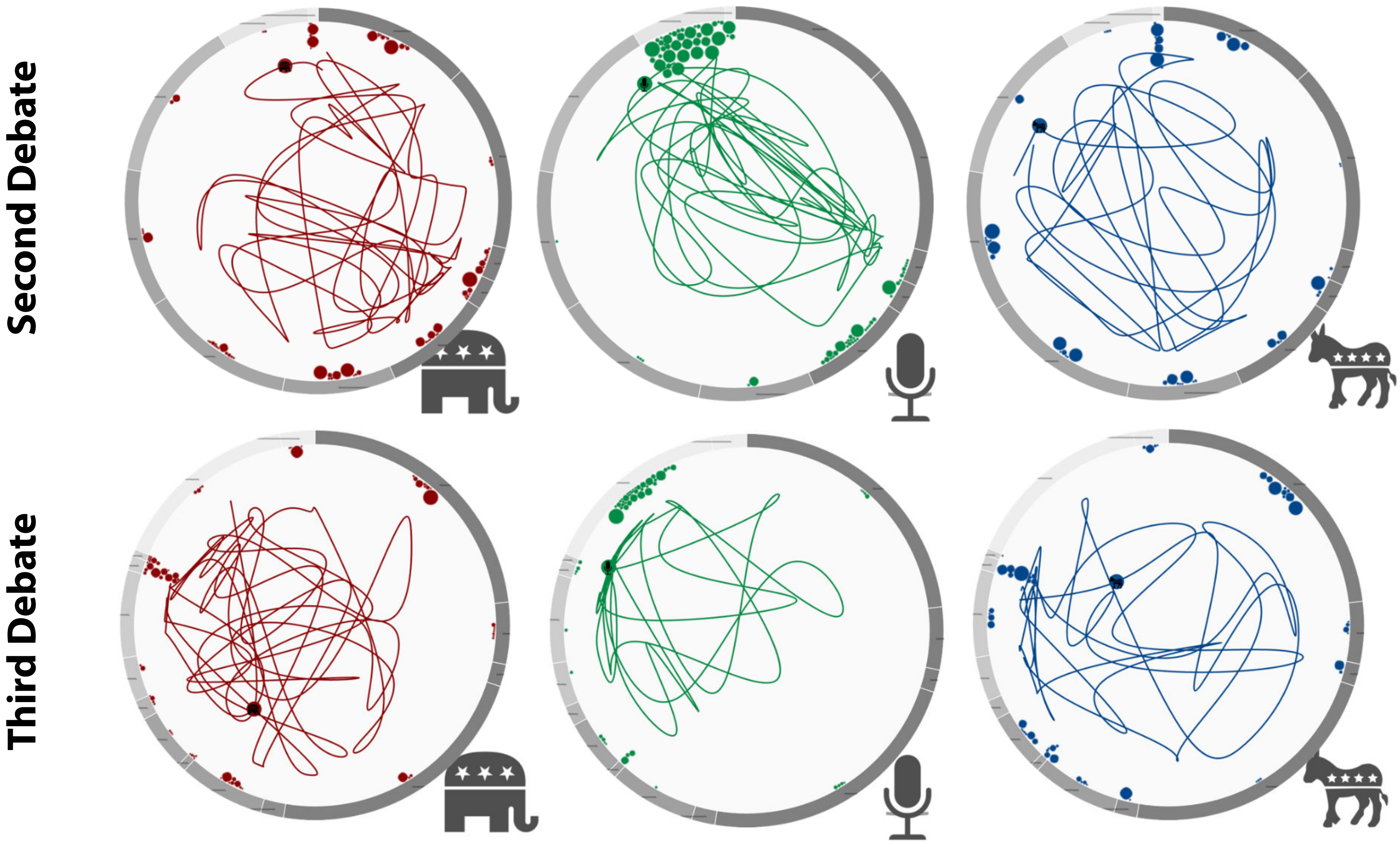
Speaker Interactions



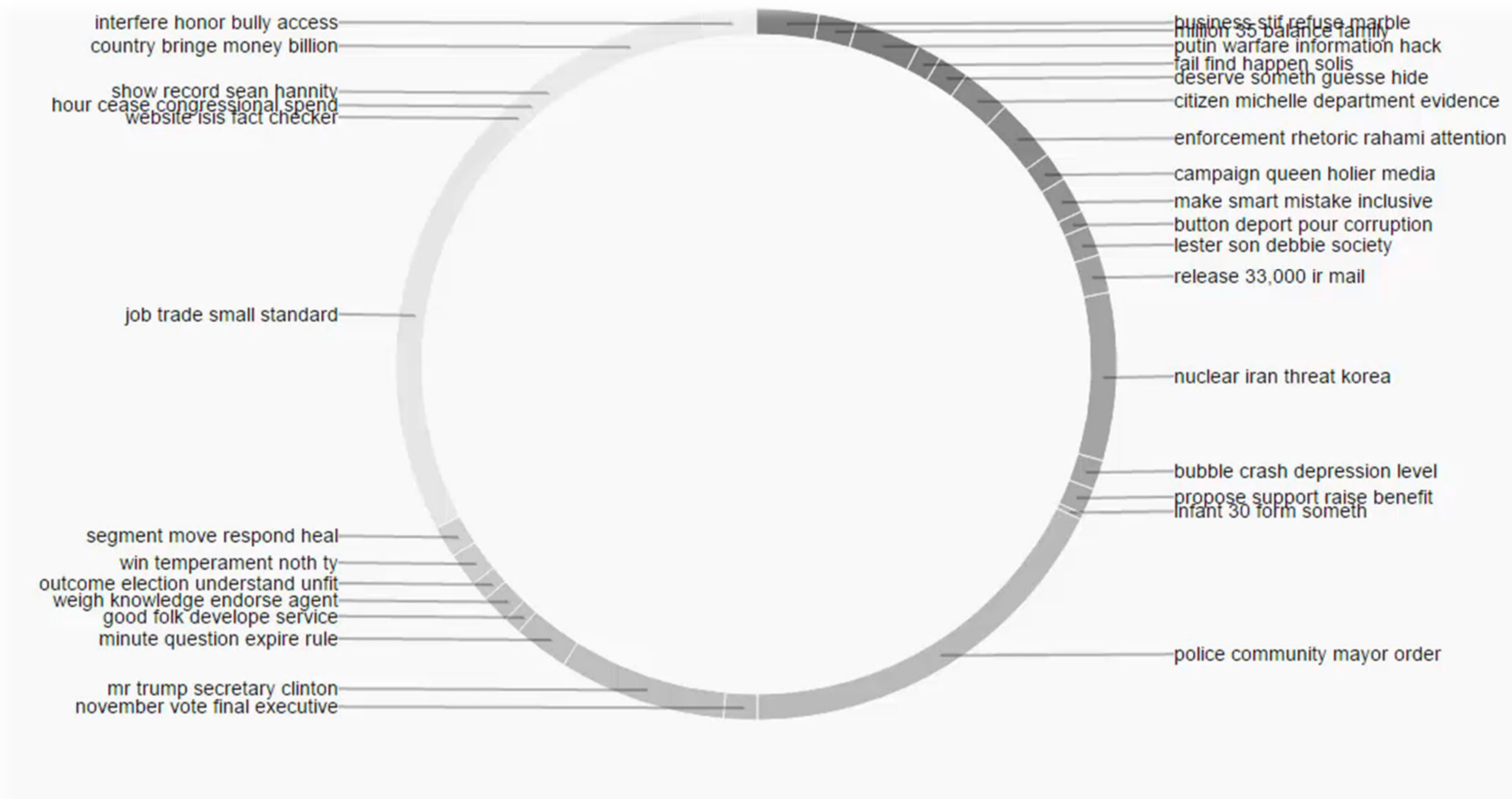


How were the speakers moving in the topic space?





How were the speakers moving in the topic space?



Summary

- Many types of visualization for text data
- Types of NLP applied differ, from counting words to topic modelling and machine learning
- User tasks differ
- Interdisciplinary teams