White Paper

Report ID: 107654

Application Number: HK-50011-12

Project Director: Marti Hearst

Institution: University of California, Berkeley

Reporting Period: 9/1/2012-5/31/2015

Report Due: 12/31/2015

Date Submitted: 2/26/2016

FINAL PERFORMANCE REPORT

NEH HK-50011-12:

WordSeer: A Text Analysis Environment for Literature Study

Marti A. Hearst,
Project Director,
University of California, Berkeley

February 26, 2016

NARRATIVE DESCRIPTION

This is the final report for The WordSeer project, NEH HK-50011-12. This project received total funding of $217,189 as an implementation grant to improve on the WordSeer digital humanities software tool for exploration of textual literature. Details about the WordSeer project can be viewed at :

http://wordseer.berkeley.edu

SUMMARY OF ACCOMPLISHMENTS

This is the final report for The WordSeer project, NEH HK-50011-12. This project received total funding of $217,189 as an implementation grant to improve on the WordSeer digital humanities software tool for exploration of textual literature. Details about the WordSeer project can be viewed at http://wordseer.berkeley.edu

Summary of Accomplishments

The first iteration of the WordSeer development was funded by an NEH startup grant, HD-5124411. This following grant, HK-50011-12, was an implementation grant. The PI was Professor Marti Hearst of UC Berkeley. The lead graduate student on the grant was Aditi Muralidharan, who received her PhD in December 2013 from EECS at UC Berkeley. Prof. Bryan Wagner of the English department at UC Berkeley was a collaborator.

As stated in our project proposal, our goals were to:

- Package WordSeer into a freely available open-source program, and to make it more accessible to humanities scholars by:
    - Building an interface that allows specification of which metadata to use from the XML file
    - Not requiring users to have access to a LAMP stack, but instead is much

easier to install on a laptop

- Develop additional digital humanities case studies with the tool
- Develop additional understanding of the interfaces and interactions required to understand the process of literature study on collections of text.

The software development goals were achieved. As described below, the latest release of WordSeer, WordSeer 4.0, has a new software infrastructure that allows users to run the software on their laptops without requiring server software. It also allows for creation of a WordSeer project without writing code, due to an innovative new visual tool for converting XML text into a JSON format that WordSeer can use directly. New software allows for the creation and management of projects, management of users, and organization of files.

The intellectual goals were also achieved. In Spring of 2014, Hearst, aided by Muralidharan and Wagner, lead a pro-seminar on digital humanities. Some of the students in the pro-seminar opted to conduct case studies using WordSeer, which cast more insight on the use of tool, and lead to a publication describing these case studies in CIKM. One of these students (Fan), used finding from the tool to aid in the writing of his dissertation in English. This resulted in the publication of a paper describing the use of the tool for scientific discovery (CICM 2013).

Another line of research was conducted about the best was to show search suggestions for syntactic relations in a search tool. This resulted in a publication in ACL 2014.

As documented in earlier reports, WordSeer has been used by an instructor from another university to help English students study Shakespeare in a digital humanities framework.

PUBLICATIONS AND TALKS

During the course of the grant, the following publications associated with the research appeared. Some had been in development during the startup phase of the grant.

- Muralidharan, Aditi, A Visual Interface for Exploring Language Use in Slave Narratives, HCIR Poster Session, 2011.
- Muralidharan, A. and Hearst, M.A., Finding Literary Themes with Relevance Feedback, HCIR 2012.
- Muralidharan, Aditi, and Marti A. Hearst. "A sensemaking environment for literature study." CHI'12 Extended Abstracts on Human Factors in Computing Systems. ACM, May 2012.
- Muralidharan, A. and Hearst, M.A., Supporting Exploratory Text Analysis in Literature Study, Literary and Linguistic Computing, 28 (2), 2013
- Muralidharan, Aditi, Marti A. Hearst, and Christopher Fan. "WordSeer: a knowledge synthesis environment for textual data." Proceedings of the 22nd ACM international Conference on information & knowledge management (Demos), 2013.
- Muralidharan, Aditi, Designing an Exploratory Text Analysis Tool for Humanities and Social Sciences Research, unpublished doctoral dissertation, UC Berkeley Technical Report, EECS-2013-203, Dec, 2013.
- Muralidharan, Aditi, and Marti A. Hearst. "Improving the Recognizability of Syntactic Relations Using Contextualized Examples." Proceedings of ACL, June 2014.

This work was presented at many venues during the course of the grant, including:

- Keynote talk, Analyzing Text at the Middle Distance between the Close Read and

Culturomics, Language in Social Media Workshop at NAACL-HLT, Montreal, June 2012.

- UC Berkeley CITRIS Data Innovation Day (Streams, Gardens and Clouds), “The Middle Distance: Digital Humanities and Visualization,” January 24, 2013.
- UC Berkeley Townsend Center for the Humanities Brown Bag Lunch Series, “Building Text-Analysis Tools for Literary Study,” March 18, 2013.
- Invited talk, UC Santa Cruz Future of Games group, Exploratory Text Analysis and the Middle Distance, June 7, 2013.
- Keynote talk, KDD IDEA Workshop, Exploratory Text Analysis and the Middle Distance, Chicago, IL, August 2013.
- Keynote talk, 4th Workshop on Controlled Natural Language (CNL 2014), Exploratory Text Analysis at the Middle Distance, Galway Ireland (delivered remotely), August 2014
- Cray Distinguished Speaker Series, Exploratory Text Analysis at the Middle Distance, University of Minnesota, April 2014.
- Invited talk, Digital Humanities Seminar, UC Berkeley, Exploratory Text Analysis and the Middle Distance, November 11, 2014

Development and Training of Students

Graduate and undergraduate students were paid during the course of this grant to work on the WordSeer software. The lead PhD student, Aditi Muralidharan, wrote most of the code, but after she completed her dissertation and graduated, several students were funded to continue the conversion of the code from PHP to python, and to fix bugs and add features. Especially notable contributions were made by masters students from the School of Information. These masters students also developed unique new functionality, including new visual views and a new tool for uploading

XML without requiring coding.

SOFTWARE PRODUCTS

The software for WordSeer is available online as open source at this location:

https://github.com/Wordseer/wordseer

The latest release, WordSeer 4.0, has several improvements over version 3.0:

- A unique interface for specifying which parts of the metadata associated with an XML file to include for processing in WordSeer. Custom code processes the XML and allows the user, via a visual view, to select which parts to include within the tool. This software is significant in its own right and could be useful for other projects within and outside the digital humanities.
- Installation scripts for both mac and windows operating systems. Users can now run the software directly on their laptops, as well as on a server.
- Software to allow users to keep track of projects, including creating and deleting new projects, along with the generated XML processing files.
- Improvements in the graphic design of the interface; the look and feel are modern and more streamlined.

The software infrastructure for WordSeer 4.0 was significantly changed to replace MySQL with SQLLite. The motivation for replacing MySQL was to make installment easier, not requiring the intervention of a system administration and allowing one-click installation. The PHP code in the backend was replaced with python using the SQLAlchemy framework to interface between the database calls and the javascript code in the front end. (The Stanford CoreNLP code that is imported uses Java, but it has a python wrapper.)

The transition was completed, and many software issues from the original code were fixed. Much of the functionality from the PHP version of the code was transferred (with some exceptions — annotations, sets, and some visualizations). However, the current version of the code does have performance problems when loaded with collections beyond a moderate size. It also retains a significant number of software bugs. It is hoped that the “many eyes” of the open source community will be able to figure out and fix the open issues.

Three new videos have been created to demonstrate the functionality of WordSeer 4.0 and posted on the project website. They explain:

1. The installation process for a Mac computer.
2. The new functionality for loading an XML-structured document directly into WordSeer.
3. How to use the features of WordSeer 4.0 after an instance has been loaded.

DETAILS OF WORDSEER 4.0

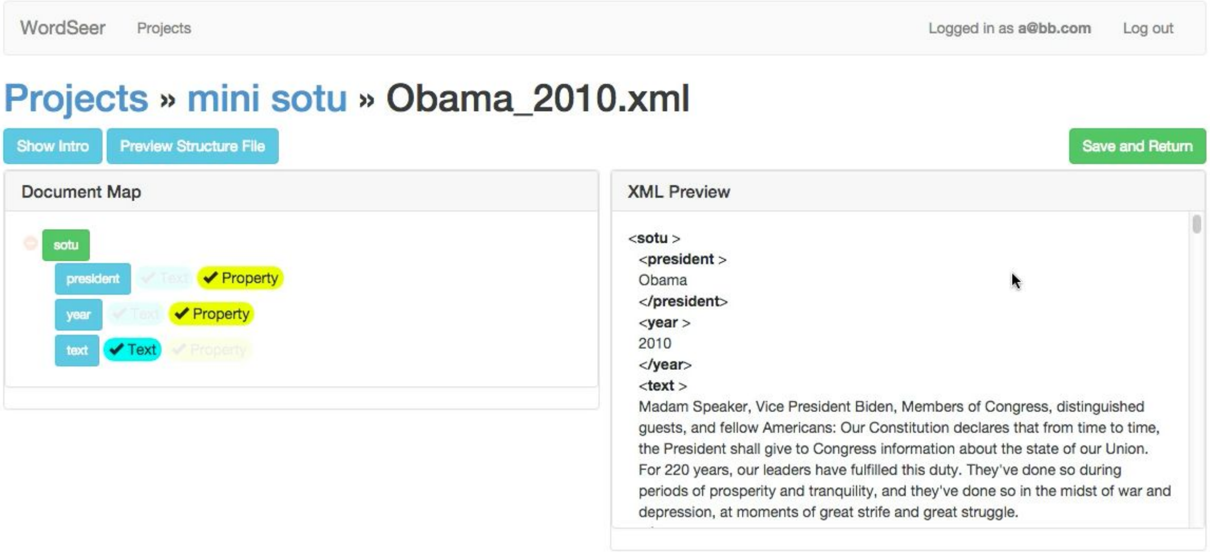
This section of the report illustrates some of the new functionality available in WordSeer 4.0. Further details can be found in the instructional videos and other material available on the WordSeer website and in the software on the github repository.

A Visual Interface for Converting XML

WordSeer 4.0 provides a unique interface for specifying which parts of the metadata associated with an XML file to include for processing. Custom code processes the XML and allows the user, via a visual view, to select which parts to include within the tool. This software is significant in its own right and could be useful for other projects within and outside the digital humanities.

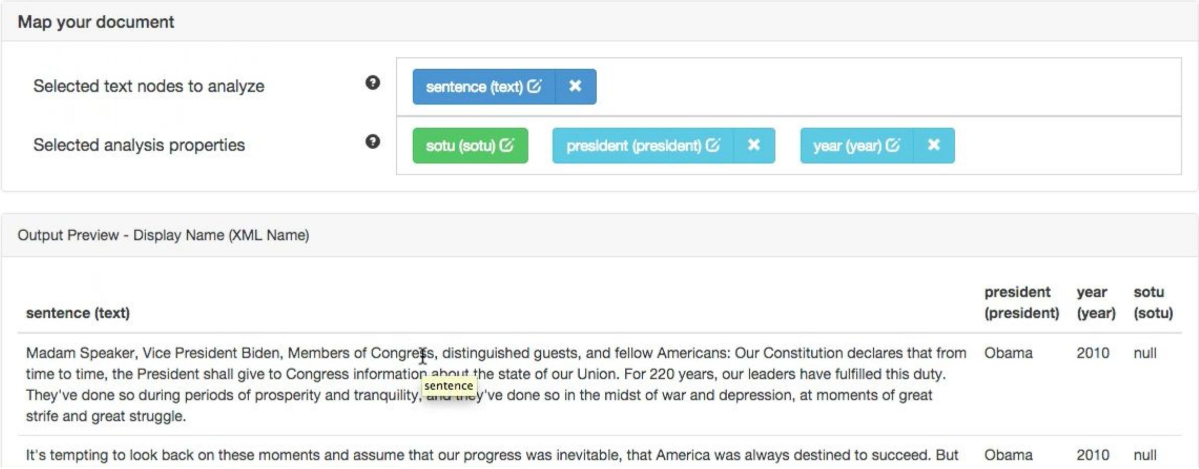
Example: The State of the Union

The screenshots below show the interface in action: after the user creates a project and uploads some files, the tool automatically parses the XML tree and allow the user to pick and choose which parts to include into the WordSeer interface.



```
<sotu >
  <president >
  Obama
  </president>
  <year >
  2010
  </year>
  <text >
  Madam Speaker, Vice President Biden, Members of Congress, distinguished
  guests, and fellow Americans: Our Constitution declares that from time to time,
  the President shall give to Congress information about the state of our Union.
  For 220 years, our leaders have fulfilled this duty. They've done so during
  periods of prosperity and tranquility, and they've done so in the midst of war and
  depression, at moments of great strife and great struggle.
```

Selecting metadata from XML files



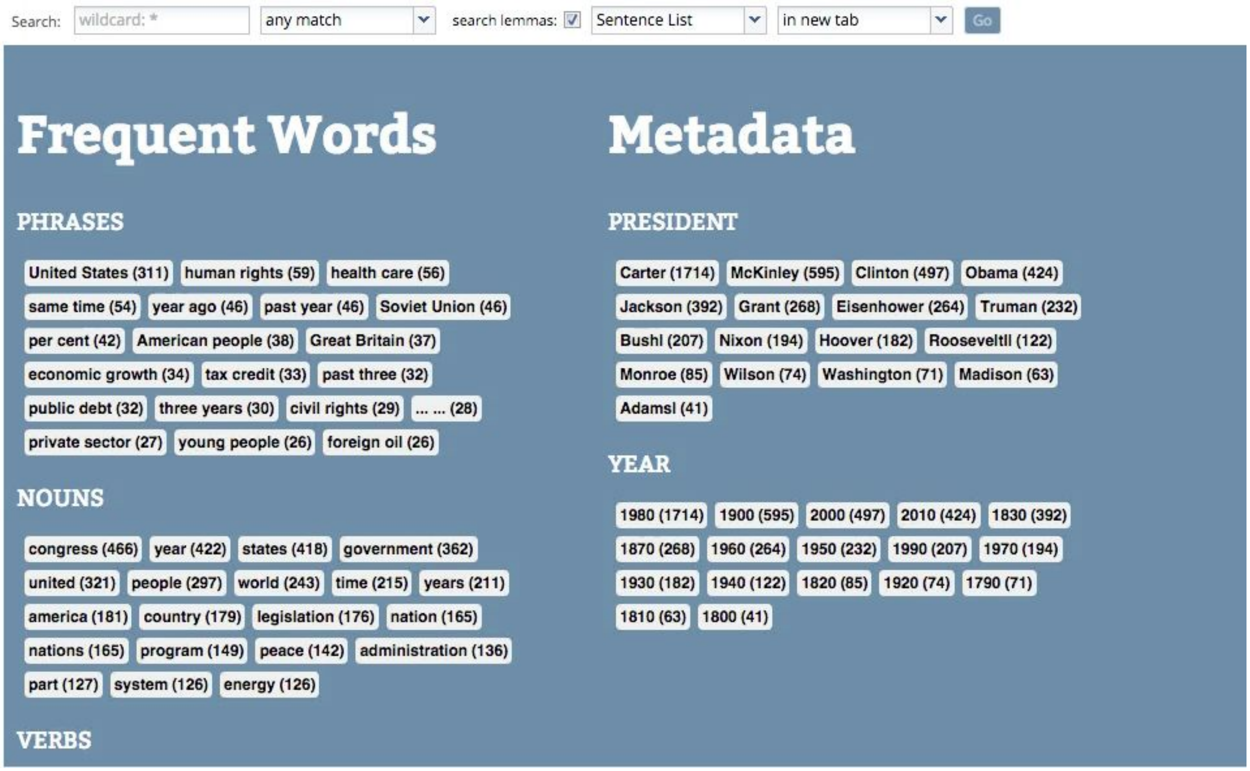
Selected text nodes to analyze: sentence (text)

Selected analysis properties: sotu (sotu), president (president), year (year)

sentence (text)	president (president)	year (year)	sotu (sotu)
Madam Speaker, Vice President Biden, Members of Congress, distinguished guests, and fellow Americans: Our Constitution declares that from time to time, the President shall give to Congress information about the state of our Union. For 220 years, our leaders have fulfilled this duty. They've done so during periods of prosperity and tranquility, and they've done so in the midst of war and depression, at moments of great strife and great struggle.	Obama	2010	null
It's tempting to look back on these moments and assume that our progress was inevitable, that America was always destined to succeed. But	Obama	2010	null

Showing a preview of the mapping from the XML structure to the selected metadata.

Below is shown the start page in WordSeer that results from this selection process. Note that no coding is required for the user to transform an XML file into the JSON file that WordSeer needs to start processing.



Search: wildcard: * | any match | search lemmas: ☑ | Sentence List | in new tab | Go

Frequent Words

PHRASES

United States (311) human rights (59) health care (56)
same time (54) year ago (46) past year (46) Soviet Union (46)
per cent (42) American people (38) Great Britain (37)
economic growth (34) tax credit (33) past three (32)
public debt (32) three years (30) civil rights (29) … … (28)
private sector (27) young people (26) foreign oil (26)

NOUNS

congress (466) year (422) states (418) government (362)
united (321) people (297) world (243) time (215) years (211)
america (181) country (179) legislation (176) nation (165)
nations (165) program (149) peace (142) administration (136)
part (127) system (126) energy (126)

VERBS

Metadata

PRESIDENT

Carter (1714) McKinley (595) Clinton (497) Obama (424)
Jackson (392) Grant (268) Eisenhower (264) Truman (232)
BushI (207) Nixon (194) Hoover (182) RooseveltII (122)
Monroe (85) Wilson (74) Washington (71) Madison (63)
AdamsI (41)

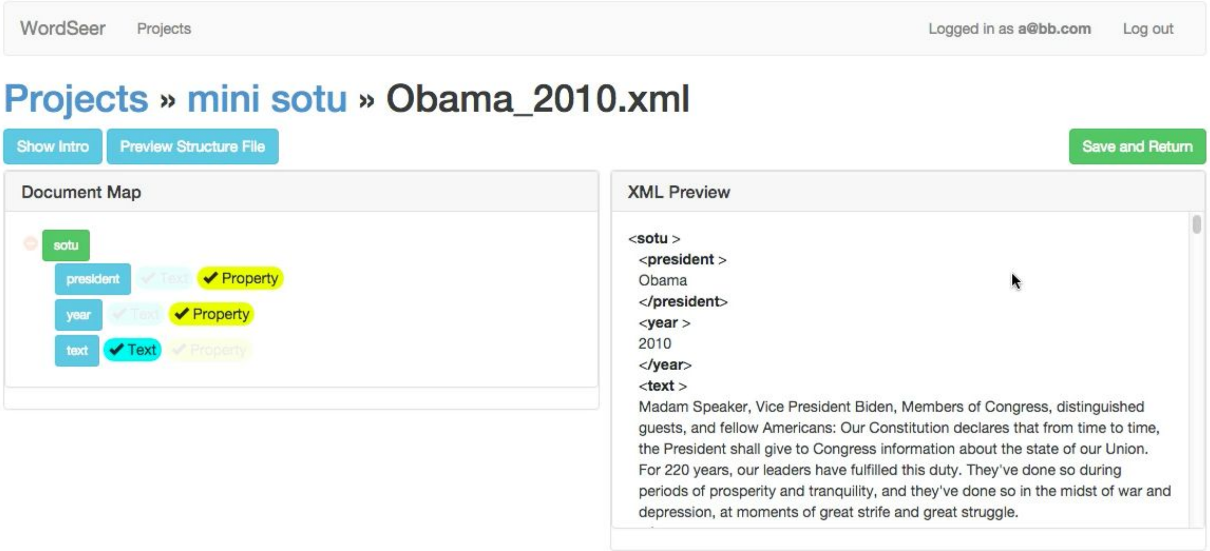
YEAR

1980 (1714) 1900 (595) 2000 (497) 2010 (424) 1830 (392)
1870 (268) 1960 (264) 1950 (232) 1990 (207) 1970 (194)
1930 (182) 1940 (122) 1820 (85) 1920 (74) 1790 (71)
1810 (63) 1800 (41)

Starting page for the WordSeer instance created by the XML selection tool.

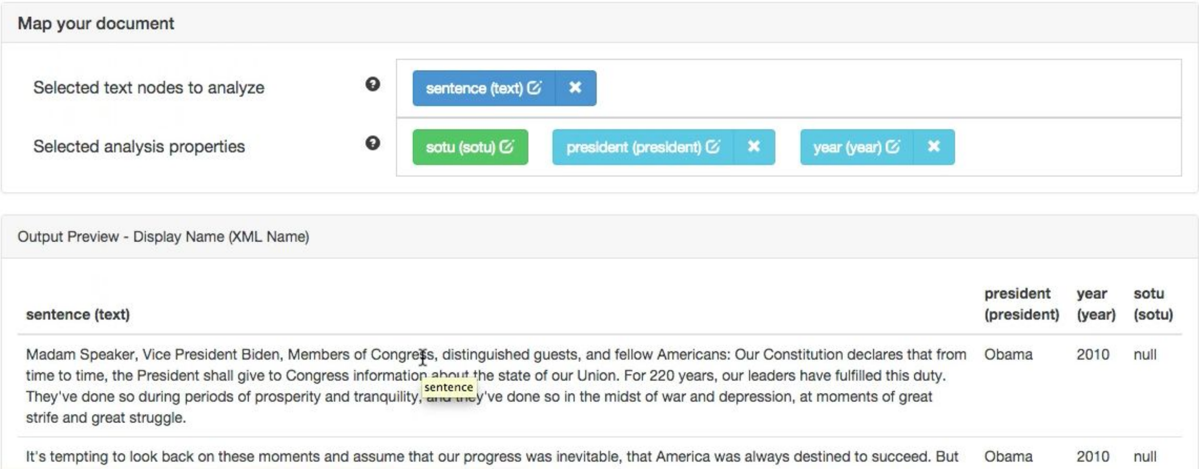
Example: Shakespeare’s Plays

A more complex example uses XML for Shakespeare's plays (we get the XML from Jon Bosak's collection). A few screenshots illustrate the visual interface.



```
<sotu >
  <president >
  Obama
  </president>
  <year >
  2010
  </year>
  <text >
  Madam Speaker, Vice President Biden, Members of Congress, distinguished
  guests, and fellow Americans: Our Constitution declares that from time to time,
  the President shall give to Congress information about the state of our Union.
  For 220 years, our leaders have fulfilled this duty. They've done so during
  periods of prosperity and tranquility, and they've done so in the midst of war and
  depression, at moments of great strife and great struggle.
```

Selecting from the XML for Shakespeare.



Selected text nodes to analyze: sentence (text)

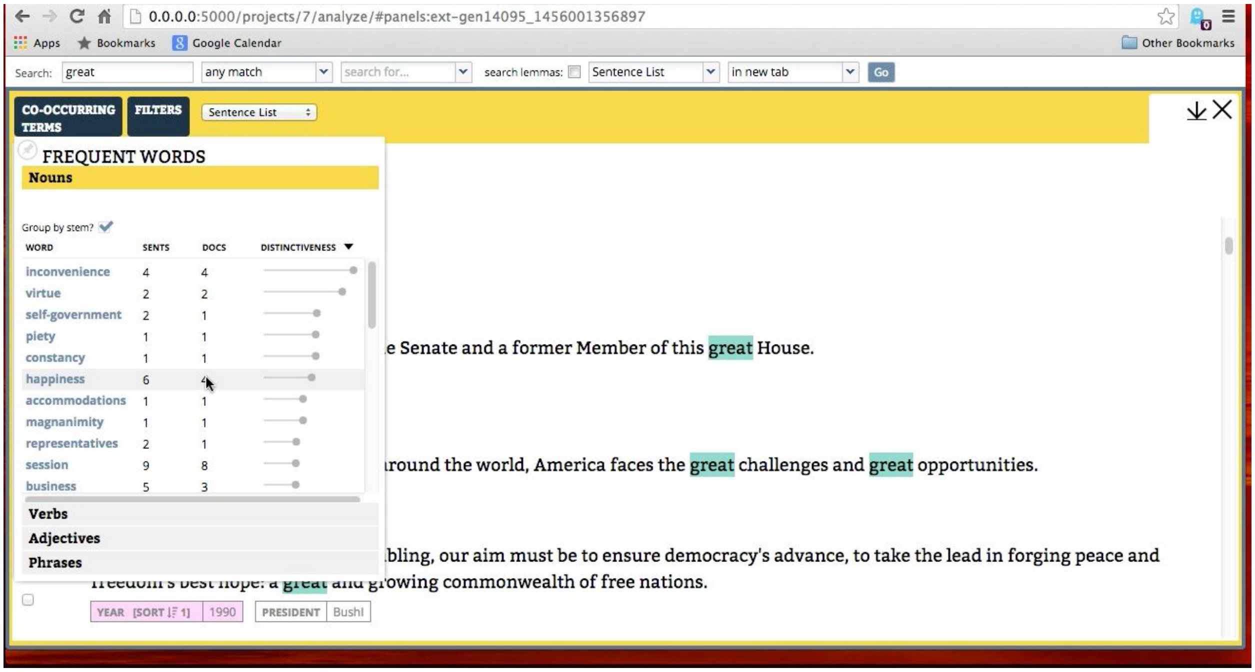
Selected analysis properties: sotu (sotu), president (president), year (year)

sentence (text)	president (president)	year (year)	sotu (sotu)
Madam Speaker, Vice President Biden, Members of Congress, distinguished guests, and fellow Americans: Our Constitution declares that from time to time, the President shall give to Congress information about the state of our Union. For 220 years, our leaders have fulfilled this duty. They've done so during periods of prosperity and tranquility, and they've done so in the midst of war and depression, at moments of great strife and great struggle.	Obama	2010	null
It's tempting to look back on these moments and assume that our progress was inevitable, that America was always destined to succeed. But	Obama	2010	null

Converting XML of Shakespeare's plays to WordSeer's format.

Details of WordSeer 4.0’s New StreamLined Look and Feel

WordSeer 4.0 has a modernized and streamlined user interface that breaks up some of the screens and uses color to better effect to differentiate the functionality. The screenshots below illustrate the functionality, much of which was present in WordSeer 3.0, but using visual hierarchy and color to better differentiate the different operations and functions.



Search: great | any match | search for... | search lemmas: □ | Sentence List | in new tab | Go

Sentence List

FREQUENT WORDS

Nouns

Group by stem? ✓

WORD	SENTS	DOCS	DISTINCTIVENESS ▼
inconvenience	4	4	
virtue	2	2	
self-government	2	1	
piety	1	1	
constancy	1	1	
happiness	6	4	
accommodations	1	1	
magnanimity	1	1	
representatives	2	1	
session	9	8	
business	5	3	

Verbs

Adjectives

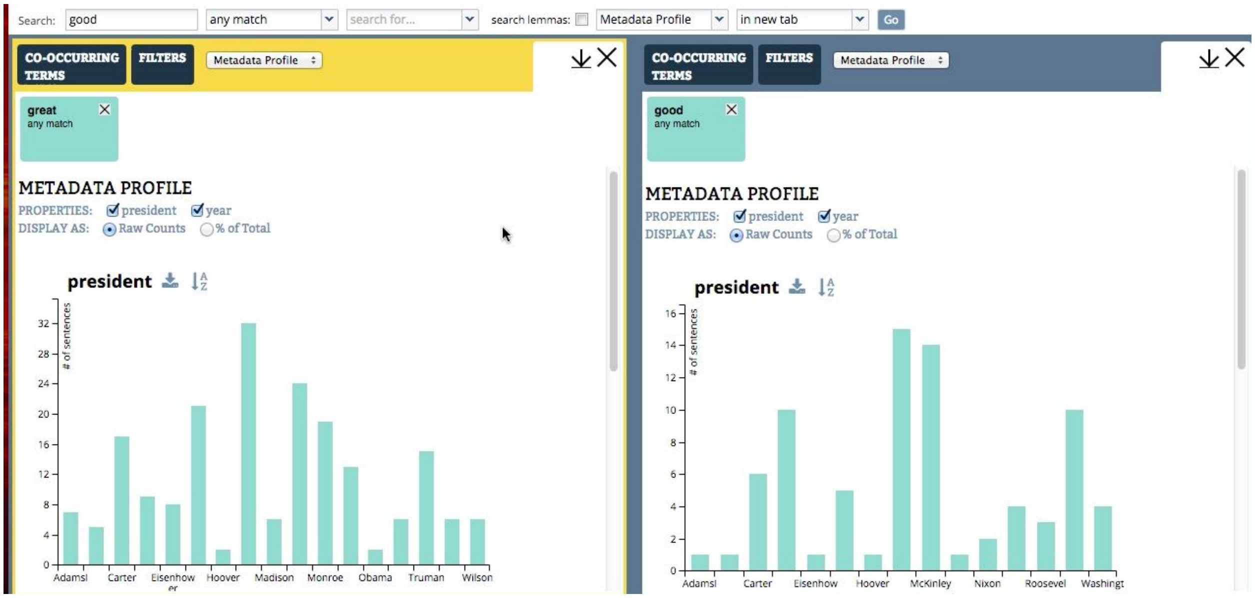
Phrases

e Senate and a former Member of this great House.

round the world, America faces the great challenges and great opportunities.

bling, our aim must be to ensure democracy's advance, to take the lead in forging peace and freedom's best hope: a great and growing commonwealth of free nations.

□ YEAR [SORT ↓ 1] 1990 | PRESIDENT Bushl



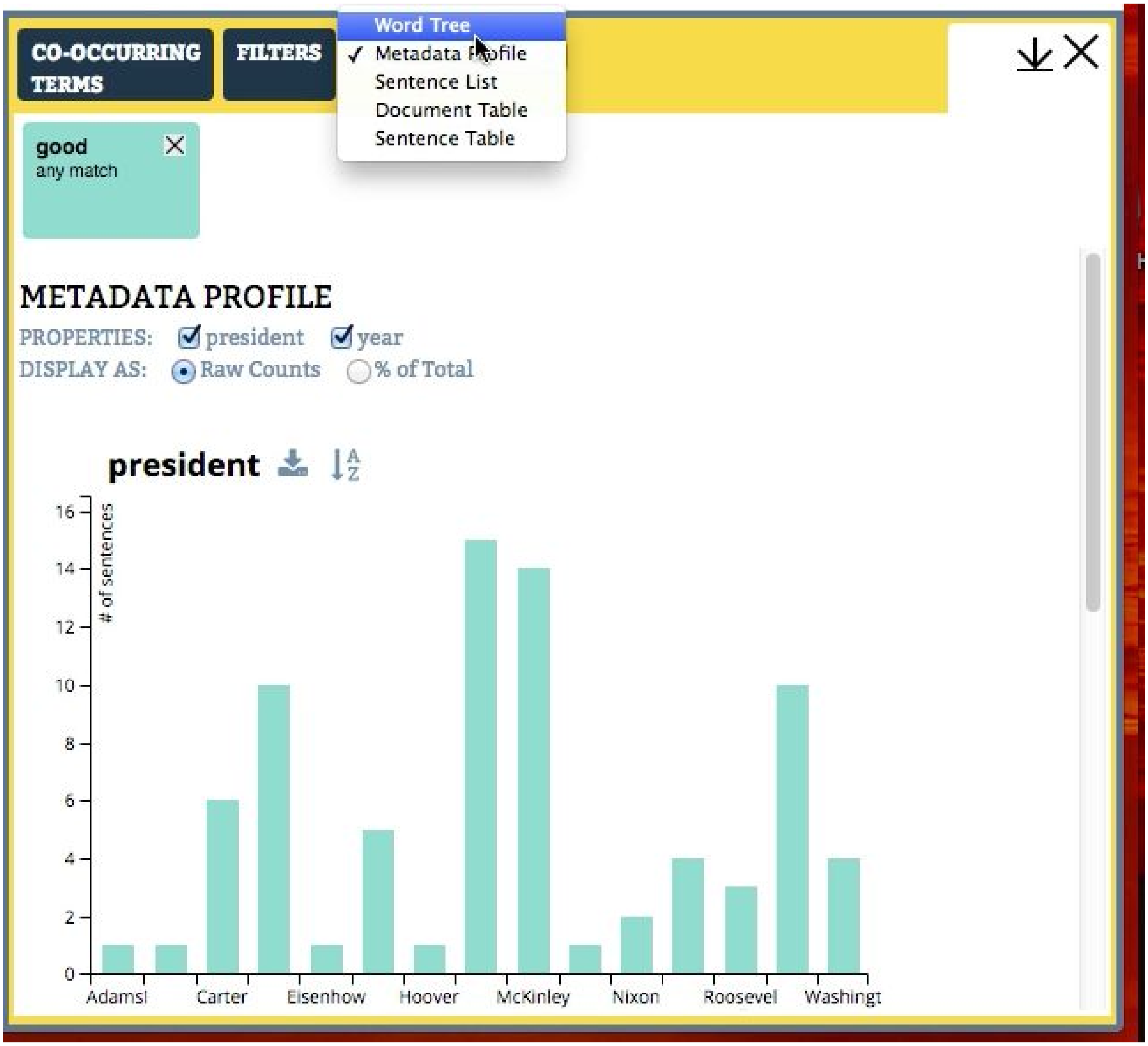
Search: good | any match | search for... | search lemmas: □ | Metadata Profile | in new tab | Go

great
any match

METADATA PROFILE
PROPERTIES: ☑ president ☑ year
DISPLAY AS: ● Raw Counts ○ % of Total

good
any match

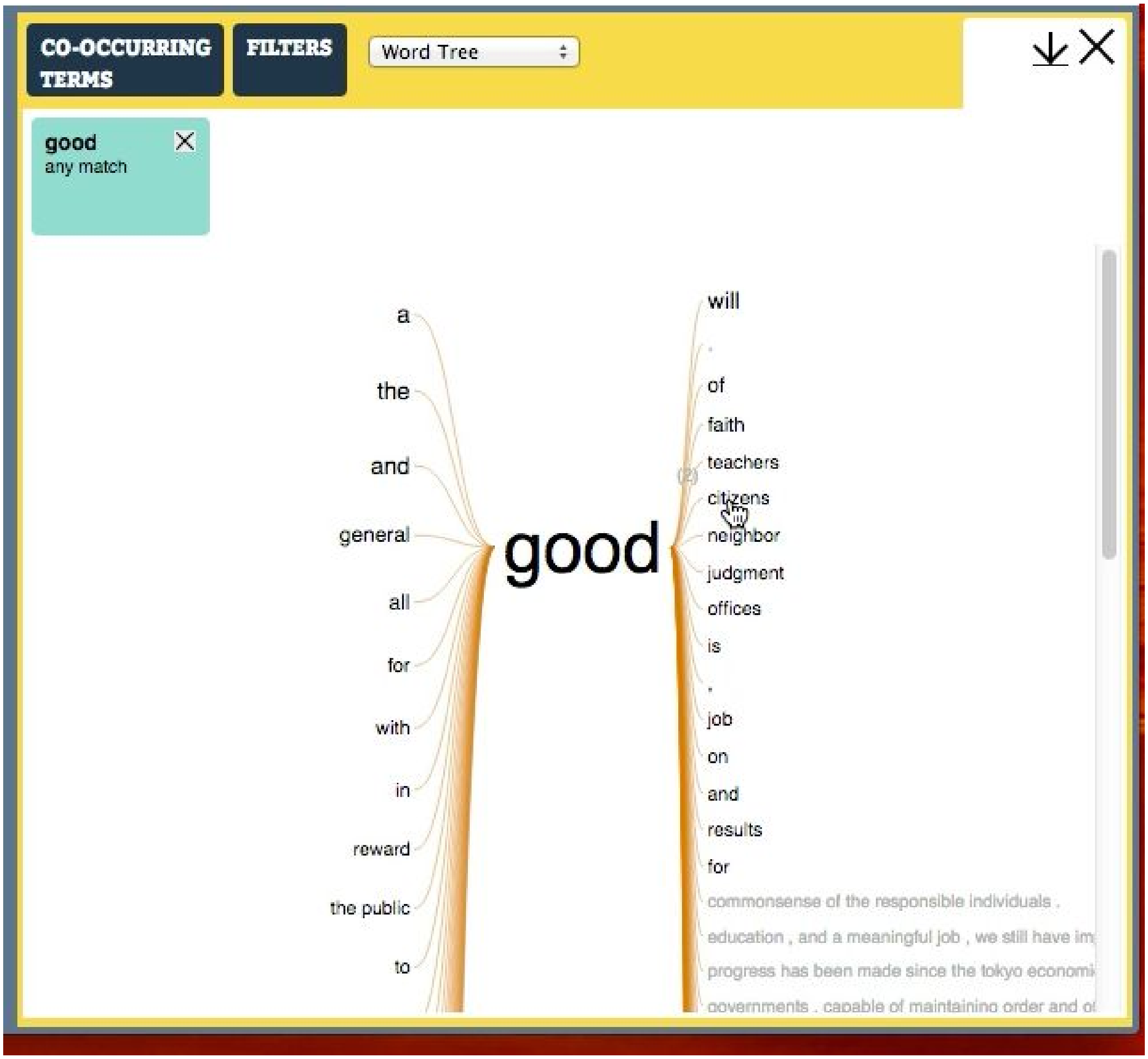
METADATA PROFILE
PROPERTIES: ☑ president ☑ year
DISPLAY AS: ● Raw Counts ○ % of Total

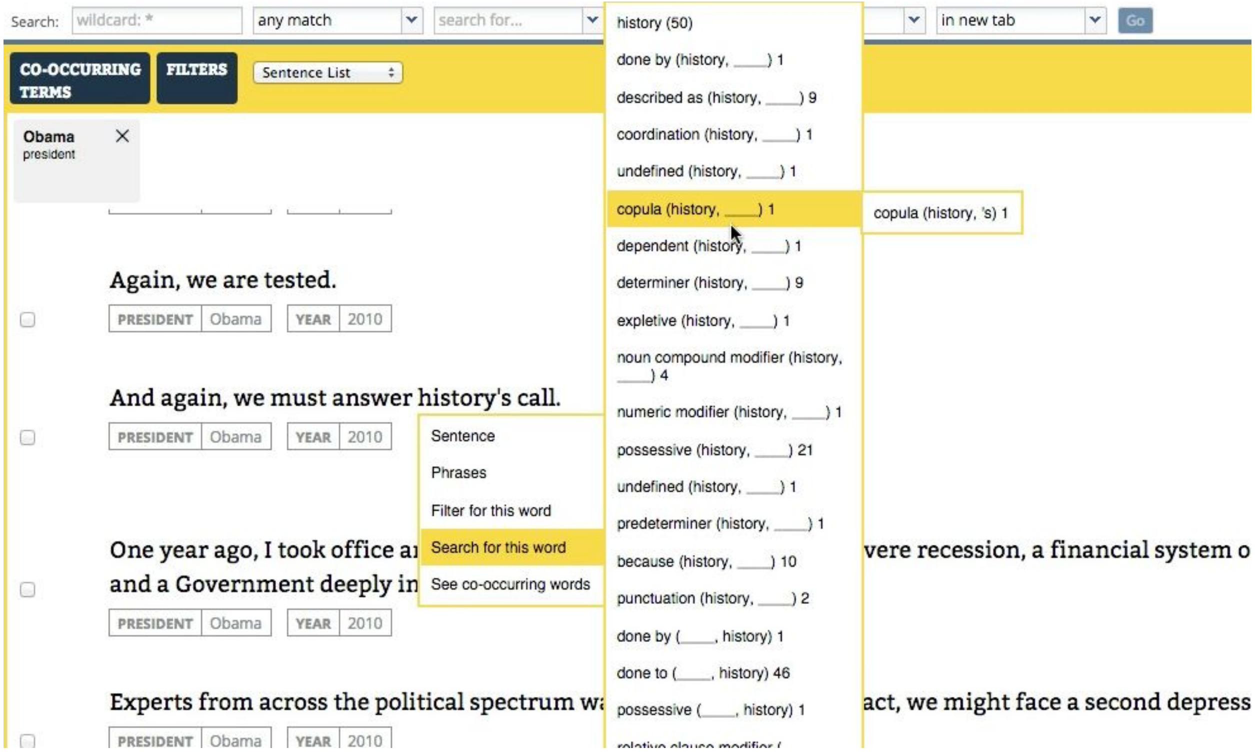


METADATA PROFILE

PROPERTIES: ☑ president ☑ year

DISPLAY AS: ● Raw Counts ○ % of Total


CO-OCCURRING TERMS
FILTERS
Word Tree
good
any match
a
the
and
general
all
for
with
in
reward
the public
to
good
will
.
of
faith
teachers
(2)
citizens
neighbor
judgment
offices
is
,
job
on
and
results
for
commonsense of the responsible individuals .
education , and a meaningful job , we still have im
progress has been made since the tokyo economi
governments , capable of maintaining order and o



Search:
wildcard: *
any match
search for...
in new tab
Go
CO-OCCURRING TERMS
FILTERS
Sentence List
Obama
president
history (50)
done by (history, ____) 1
described as (history, ____) 9
coordination (history, ____) 1
undefined (history, ____) 1
copula (history, ____) 1
copula (history, 's) 1
dependent (history, ____) 1
determiner (history, ____) 9
expletive (history, ____) 1
noun compound modifier (history, ____) 4
numeric modifier (history, ____) 1
possessive (history, ____) 21
undefined (history, ____) 1
predeterminer (history, ____) 1
because (history, ____) 10
punctuation (history, ____) 2
done by (____, history) 1
done to (____, history) 46
possessive (____, history) 1
relative clause modifier (
Sentence
Phrases
Filter for this word
Search for this word
See co-occurring words

□ Again, we are tested.

PRESIDENT Obama YEAR 2010

□ And again, we must answer history's call.

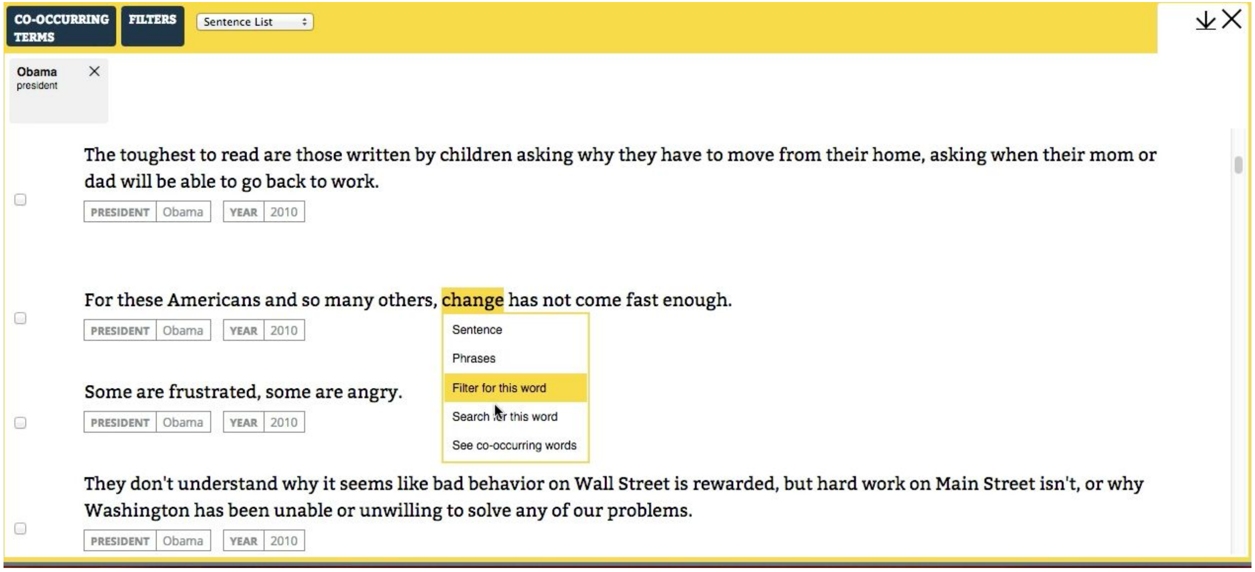
PRESIDENT Obama YEAR 2010

□ One year ago, I took office ar … vere recession, a financial system o and a Government deeply in

PRESIDENT Obama YEAR 2010

□ Experts from across the political spectrum wa … act, we might face a second depress

PRESIDENT Obama YEAR 2010



□ The toughest to read are those written by children asking why they have to move from their home, asking when their mom or dad will be able to go back to work.

PRESIDENT Obama YEAR 2010

□ For these Americans and so many others, change has not come fast enough.

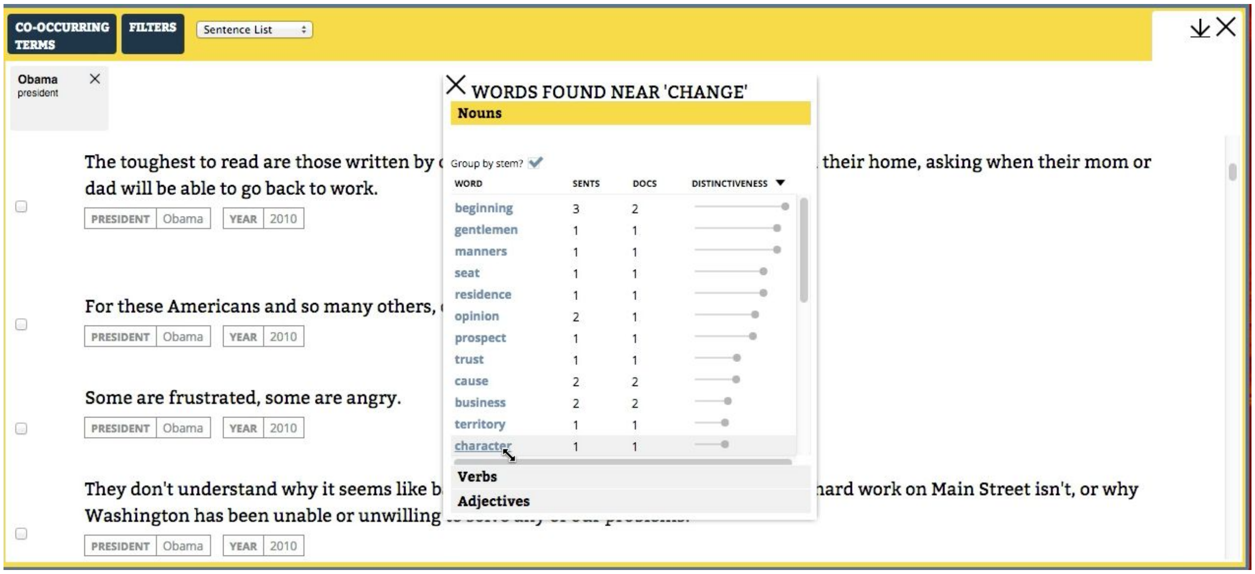
PRESIDENT Obama YEAR 2010

□ Some are frustrated, some are angry.

PRESIDENT Obama YEAR 2010

□ They don't understand why it seems like bad behavior on Wall Street is rewarded, but hard work on Main Street isn't, or why Washington has been unable or unwilling to solve any of our problems.

PRESIDENT Obama YEAR 2010



The toughest to read are those written by c their home, asking when their mom or dad will be able to go back to work.

PRESIDENT Obama YEAR 2010

For these Americans and so many others,

PRESIDENT Obama YEAR 2010

Some are frustrated, some are angry.

PRESIDENT Obama YEAR 2010

They don't understand why it seems like b hard work on Main Street isn't, or why Washington has been unable or unwilling

PRESIDENT Obama YEAR 2010

WORDS FOUND NEAR 'CHANGE'

Nouns

Group by stem? ✓

WORD	SENTS	DOCS	DISTINCTIVENESS ▼
beginning	3	2	
gentlemen	1	1	
manners	1	1	
seat	1	1	
residence	1	1	
opinion	2	1	
prospect	1	1	
trust	1	1	
cause	2	2	
business	2	2	
territory	1	1	
character	1	1	

Verbs

Adjectives

WORDSEER'S FUTURE

WordSeer is now available as open source software.   Digital humanities users are invited to make use of the software, and to discover information about text with the code. We also invite those with software expertise to improve the software's performance and enhance it with new features. We hope that the open source software is taken up by a community of programmers and so improved.

The most commonly suggested new direction that we think is of interest to those researchers in the digital humanities and social sciences is as a tool for coding qualitative data. We think WordSeer provides a compelling software starting point for this application.

Acknowledgements

The WordSeer team is deeply grateful to the NEH for funding this work, and especially to Brett Bobley for his enthusiasm and support. We also thank Jennifer Serventi and the helpful staff at the NEH, especially Jerri Shepherd.