Getting Started with Textual Analysis

Dr. Christopher M. Church
University of Nevada, Reno
christopherchurch@unr.edu
www.christophermchurch.com

What will we be learning?

- Textual analysis (aka text mining): the process of distant reading

What are Voyant Tools?

- Free online textual analysis suite put out by hermaneuti.ca (Stéfan Sinclair & Geoffrey Rockwell)
- Fairly easy to use and straightforward (requires no programming knowledge)
- Good for getting started and doing light analysis, but sacrifices fine-grained control

What is our data set?

- A month of transcripts from Sunday talk shows from three different networks (Fox News, CNN, and NBC).
    - FNS = Fox News Sunday on the Fox News Channel
    - SOTU = State of the Union on CNN
    - MTP = Meet the Press on NBC
- Enrichment: Books available from Project Gutenberg

Where to go for help?

- Tools from Voyant with documentation: http://docs.voyant-tools.org/tools/
- TAPoR (Text Analysis Portal for Research): http://www.tapor.ca/

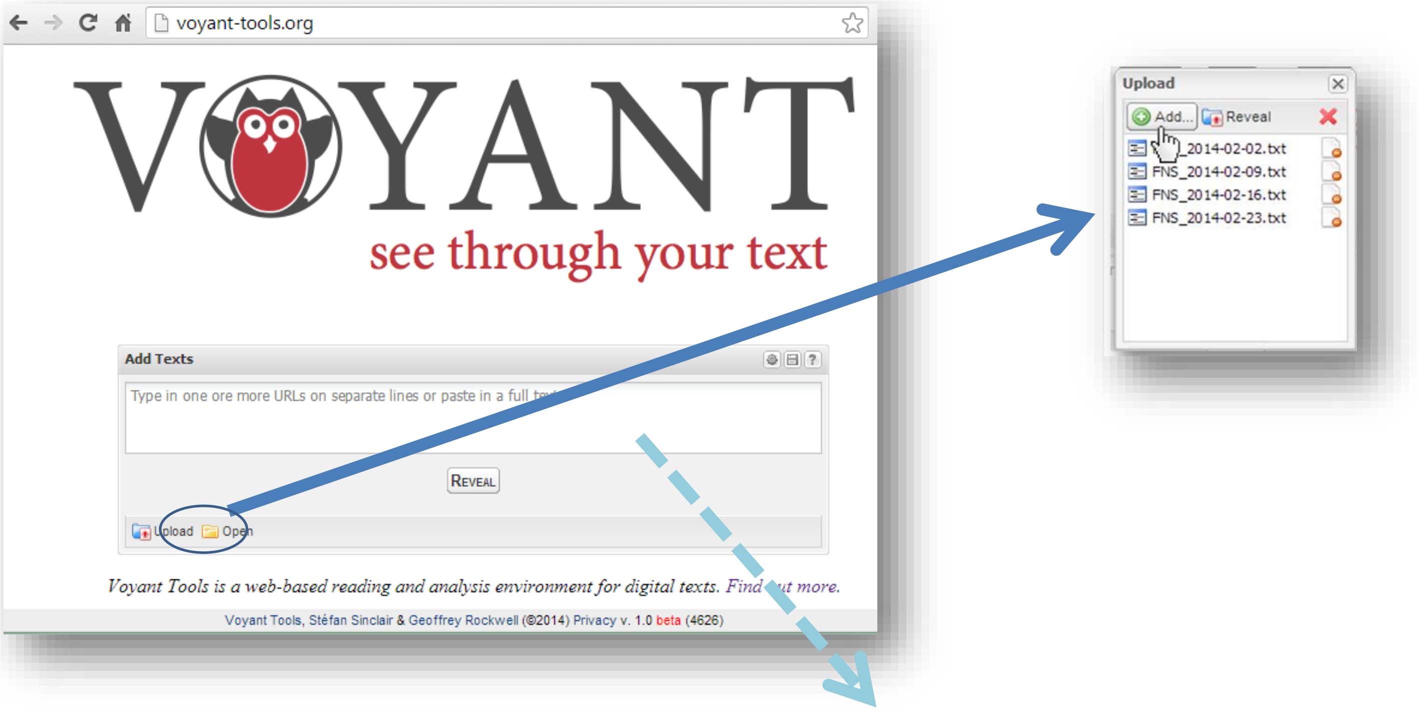
Where to get this packet?

- http://www.christophermchurch.com/uploads/voyant-workshop/docs/

Uploading our corpus

1. Download the Sunday Shows transcript corpus for March 2014 from (http://www.christophermchurch.com/uploads/voyant-workshop/individual/data.zip)
   You should save it to your Desktop for easy access.

2. Now navigate your browser to www.voyant-tools.org and click "Upload." Add each of the text files one at a time in the dialog. Make sure to add all the documents with prefixes FNS, SOTU, and MTP.



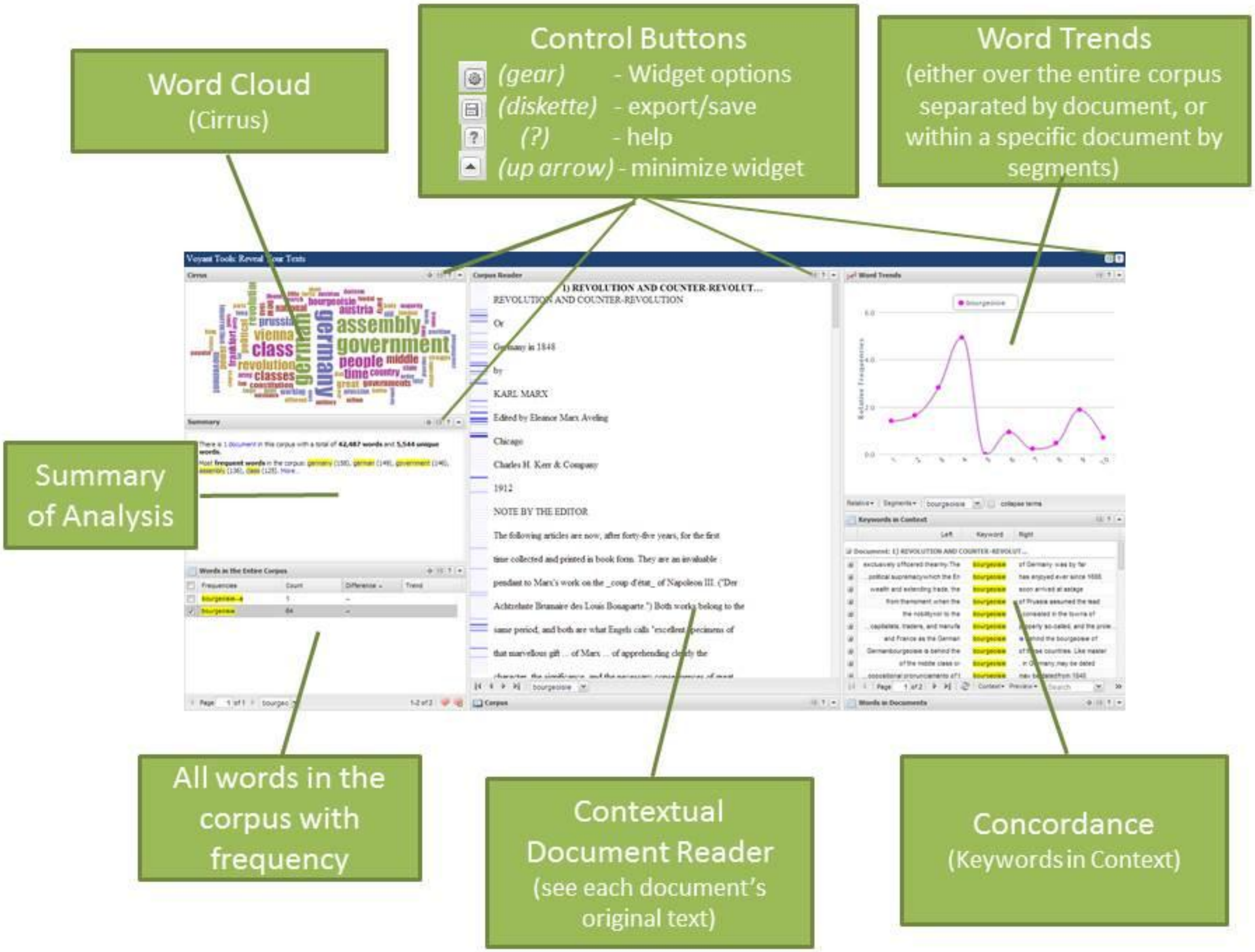
Voyant Tools is a web-based reading and analysis environment for digital texts. Find out more.

Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (©2014) Privacy v. 1.0 beta (4626)

> NOTE: To save time, you can grab the files from my server by putting the following URLS in the "Add Texts" section (one per line). *Copy and paste*. (if the PDF gives you trouble, copy the links from here: http://www.christophermchurch.com/uploads/voyant-workshop/docs/plain-text-list-of-links.txt

```
http://www.christophermchurch.com/uploads/voyant-workshop/individual/FNS_2014-02-02.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/FNS_2014-02-09.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/FNS_2014-02-16.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/FNS_2014-02-23.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/FNS_2014-03-02.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/MTP_2014-02-02.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/MTP_2014-02-09.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/MTP_2014-02-16.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/MTP_2014-02-23.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/MTP_2014-03-02.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/SOTU_2014-02-02.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/SOTU_2014-02-09.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/SOTU_2014-02-16.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/SOTU_2014-02-23.txt
http://www.christophermchurch.com/uploads/voyant-workshop/individual/SOTU_2014-03-02.txt
```

3. Once you have added all the files (15 in all), click Reveal to continue to the analysis screen. Take some time getting acquainted with the tools screen. Click around and see what happens. You can read more about the analysis screen's layout here: http://hermeneuti.ca/voyeur/users



Word Cloud
(Cirrus)

Control Buttons
⚙ *(gear)* - Widget options
💾 *(diskette)* - export/save
? *(?)* - help
▲ *(up arrow)* - minimize widget

Word Trends
(either over the entire corpus separated by document, or within a specific document by segments)

Summary of Analysis

There is 1 document in this corpus with a total of **42,487 words** and **5,544 unique words**.
Most **frequent words** in the corpus: germany (150), german (148), government (140), assembly (136), class (125). More...

1) REVOLUTION AND COUNTER-REVOLUT...
REVOLUTION AND COUNTER-REVOLUTION

Or

Germany in 1848

by

KARL MARX

Edited by Eleanor Marx Aveling

Chicago

Charles H. Kerr & Company

1912

NOTE BY THE EDITOR

The following articles are now, after forty-five years, for the first time collected and printed in book form. They are an invaluable pendant to Marx's work on the _coup d'etat_ of Napoleon III. ("Der Achtzehnte Brumaire des Louis Bonaparte.") Both works belong to the same period, and both are what Engels calls "excellent specimens of that marvellous gift ... of Marx ... of apprehending clearly the

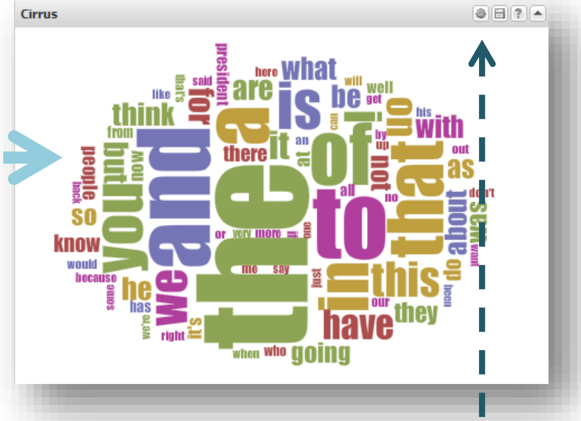
All words in the corpus with frequency

Contextual Document Reader
(see each document's original text)

Concordance
(Keywords in Context)

Voyant Tools Overview

Doing Some Analysis

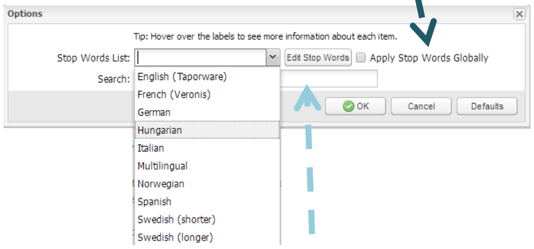


1. OK, at this point, all you really know is that the news anchors and their guests use a lot of **prepositions, articles, conjunctions**, and other ***stop words*** that tell us very little about their topics of conversation.

2. Let's filter out the stop words to get a better picture of what's going on:

   a. Click on the gear

   b. Now we need to filter out the **stop words**. To do so, choose **"English (Taporware)"** and then check **"Apply Stop Words Globally"** and click "Ok."



Tip: Hover over the labels to see more information about each item.

Stop Words List:

☐ Apply Stop Words Globally

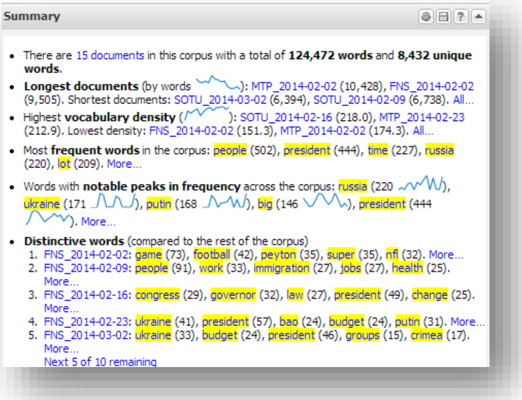
Search:

   c. We should now have a better image of what is in the corpus. Take a look.

   d. We can also customize our stop word list if we want to get rid of other common words that we're not interested in (e.g. common verbs, pronouns). We do this by clicking **"Edit Stop Words."** I've created a more expansive list. (http://www.christophermchurch.com/uploads/voyant-workshop/stopwords.txt).

   e. Please select all, copy and paste (**CTRL-A, CTRL-C, CTRL-P**) replace the current stop word list with the custom one I've created. ***What has changed?***

3. Now that we have the stop words settled, we can see what was covered by each show on each date. Let's check out the **Summary Statistics**.



- There are 15 documents in this corpus with a total of **124,472 words** and **8,432 unique words**.
- **Longest documents** (by words): MTP_2014-02-02 (10,428), FNS_2014-02-02 (9,505). Shortest documents: SOTU_2014-03-02 (6,394), SOTU_2014-02-09 (6,738). All...
- Highest **vocabulary density** ( ): SOTU_2014-02-16 (218.0), MTP_2014-02-23 (212.9). Lowest density: FNS_2014-02-02 (151.3), MTP_2014-02-02 (174.3). All...
- Most **frequent words** in the corpus: people (502), president (444), time (227), russia (220), lot (209). More...
- Words with **notable peaks in frequency** across the corpus: russia (220), ukraine (171), putin (168), big (146), president (444). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. FNS_2014-02-02: game (73), football (42), peyton (35), super (35), nfl (32). More...
  2. FNS_2014-02-09: people (91), work (33), immigration (27), jobs (27), health (25). More...
  3. FNS_2014-02-16: congress (29), governor (32), law (27), president (49), change (25). More...
  4. FNS_2014-02-23: ukraine (41), president (57), bao (24), budget (24), putin (31). More...
  5. FNS_2014-03-02: ukraine (33), budget (24), president (46), groups (15), crimea (17). More...

  Next 5 of 10 remaining

   a. Take a look at the information available.

   b. A few things to note:

      i. **Distinctive Words** give a good sense of what was being discussed on each show. You can check this against the show's title (for FNS and SOTU) by looking at the transcript contents available at http://www.christophermchurch.com/uploads/voyant-workshop/transcript-contents.txt

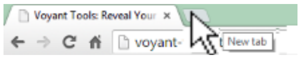
   c. Take some time looking at the other Summary Statistics. Try clicking on things and see what happens (note: most things can be clicked on, and the analysis tool will react by showing you more context).

Aggregate Analysis

1. OK, now what if we wanted to compare the different Sunday Shows against one another for the entire month of February? Well, to do so, we'd need to do some **pre-processing** on the data. In fact, the individual transcripts were already **pre-processed** in order to remove the speaker tags from the text.

    i. n.b. All data analysis requires manipulation of the data prior to analysis, and this varies from a small amount to a great deal, and is always based on the scholar's knowledge of the data set. When doing data analysis, always try to make this transparent (you can see my pre-processing code at https://github.com/cmchurch/NLP-SUNDAY-NEWS_tutorial).

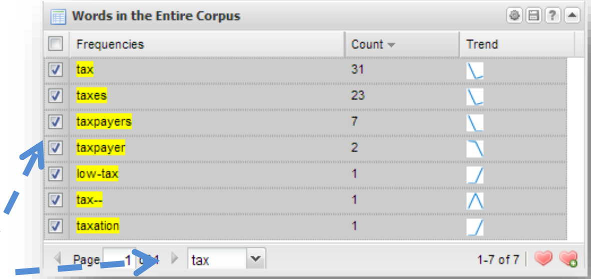
2. So, now we need to start a new Voyant Tools display for our aggregate analysis. Leaving the previous one open, create a new tab, go to voyant-tools.org, and enter the following three URLS:



http://www.christophermchurch.com/uploads/voyant-workshop/aggregate/FNS_all.txt
http://www.christophermchurch.com/uploads/voyant-workshop/aggregate/MTP_all.txt
http://www.christophermchurch.com/uploads/voyant-workshop/aggregate/SOTU_all.txt

NOTE: As mentioned earlier, you can also download the data and use the Upload Dialog (http://www.christophermchurch.com/uploads/voyant-workshop/aggregate/data.zip).

3. As before, make sure you **filter out the stop words** to get a meaningful picture of the corpus' contents.


Words in the Entire Corpus

□	Frequencies	Count	Trend
☑	tax	31	
☑	taxes	23	
☑	taxpayers	7	
☑	taxpayer	2	
☑	low-tax	1	
☑	tax--	1	
☑	taxation	1	

4. Let's take a look at some aggregate comparisons, starting with discussions of "**taxes.**"

    a. Let's first search for all words in the corpus that have "tax" in it.

       Check all the boxes next to words with **tax** in them.

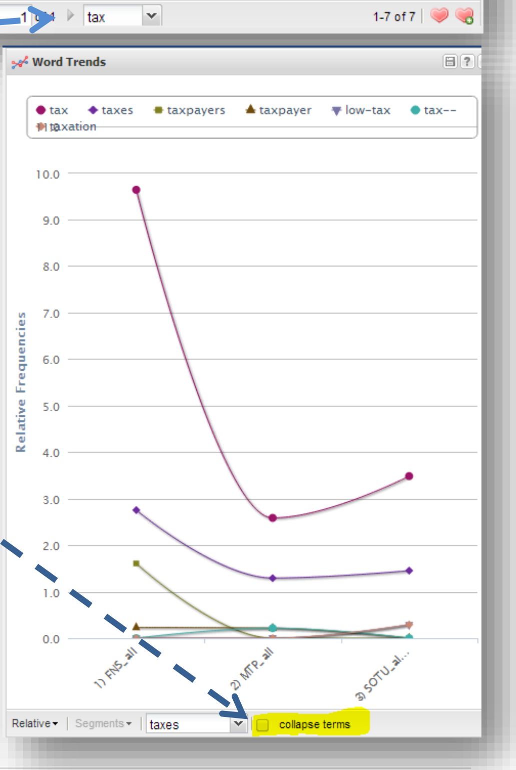
    b. This will plot all the words on our **Word Trends** widget.

    c. To combine the terms to get a sense of the overall discussion of taxes, click the **collapse terms** check box.

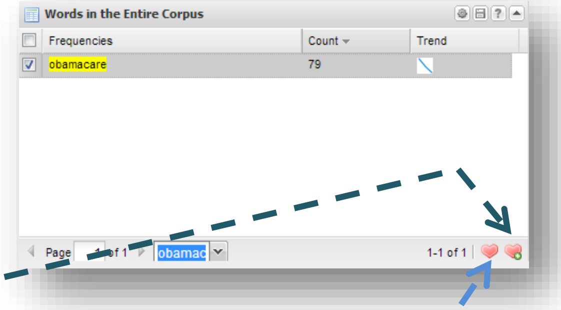
    d. Try this out with some of your own terms to compare the three Sunday Shows (you can use the Cirrus or Summary widgets to think up some terms, or you can use the following suggestions):

        i. democrat(s), republican(s)

        ii. job(s)



Aggregate Analysis (2)



Words in the Entire Corpus

☐ Frequencies	Count	Trend
☑ obamacare	79	

Page 1 of 1 obamac 1-1 of 1

5. To compare different terms, we'll need to use the **Favorites list.**

   a. Let's compare the use of "**obamacare"** with the use of the word "**health care**" across the three shows.

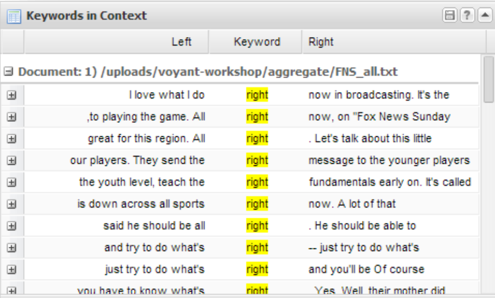
   b. To do so, search for "obamacare," check the box, and then click "Add to favorites." ❤ (heart plus).

   c. Do the same for "healthcare" and then bring up the **Favorites list** by clicking on the heart. ❤
   n.b. ignore the error that pops up (it's from the space character in "health care"

   d. Now we can check both "obamacare" and "heath care" and see them on the **Word Trends** widget. Make sure to **uncheck the "collapse terms"** button to make the comparison.

6. Now, we need to be careful about how we draw conclusions regarding comparisons between the three shows. Many words have a wide variety of meanings (e.g. right vs left) that can complicate what we think we know.

   To get a sense of how the word is being used, you use the **Keywords in Context** widget.



Keywords in Context

Document: 1) /uploads/voyant-workshop/aggregate/FNS_all.txt

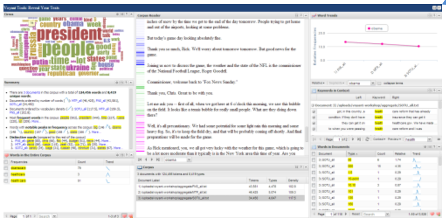
Left	Keyword	Right
I love what I do	right	now in broadcasting. It's the
,to playing the game. All	right	now, on "Fox News Sunday
great for this region. All	right	. Let's talk about this little
our players. They send the	right	message to the younger players
the youth level, teach the	right	fundamentals early on. It's called
is down across all sports	right	now. A lot of that
said he should be all	right	. He should be able to
and try to do what's	right	-- just try to do what's
just try to do what's	right	and you'll be Of course
you have to know what's	right	. Yes. Well, their mother did

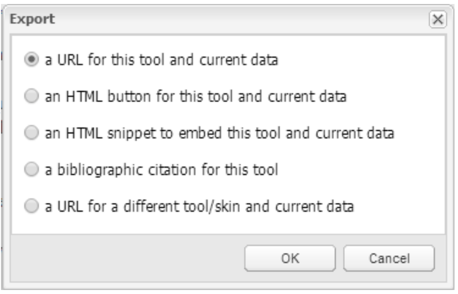
   a. Try it now with the word "**right**."

      i. How is the word "right" being used? What about "left"?

   b. Now, go back and look at your earlier search terms from step 4 in context. Were they being used how you envisioned?

7. Now, let's save the link to our data. We can use the Export Option to save our work, or to use the same data with different tools. Access the Export by clicking the diskette icon in the upper left corner.





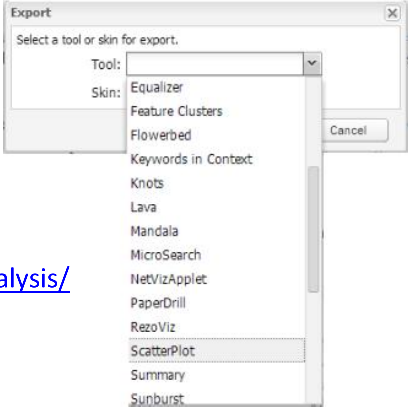
Export

● a URL for this tool and current data
○ an HTML button for this tool and current data
○ an HTML snippet to embed this tool and current data
○ a bibliographic citation for this tool
○ a URL for a different tool/skin and current data

OK Cancel

Other Tools to Try and Enrichment Activities

Collocation and Correspondence


Export

Select a tool or skin for export.

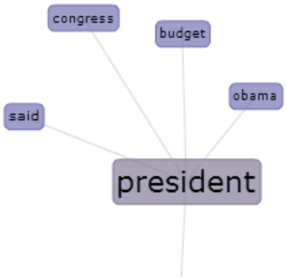
ScatterPlot (Correspondence Analysis)

Excellent Descriptions of Multivariate Correspondence Analysis:

by Lisa Goddard: http://www.tapor.ca/?id=20

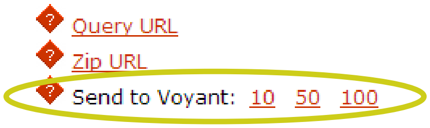
by Stefan Sinclair: http://stefansinclair.name/correspondence-analysis/

Collocation Networks



Full list of additional tools from Voyant with documentation: http://docs.voyant-tools.org/tools/

Old Bailey API (http://www.oldbaileyonline.org/obapi/)



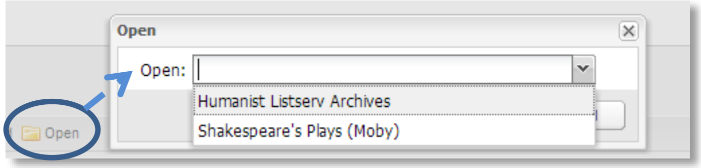
Extra Documents to Check Out

Comparing Corpora in Voyant Tools:
http://www.briancroxall.net/2012/07/18/comparing-corpora-in-voyant-tools/

Voyant Tools used to Analyze Runaway Slave Ads in Mississippi and Arkansas:
http://digitalhistory.blogs.rice.edu/files/2014/02/voyant-presentation.pdf

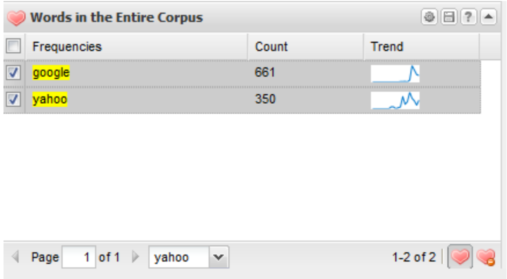
Data Visualization Handout:
http://lib.ua.edu/sites/default/files/digitalhumanities/Abbott%20Data%20Visualization%20Workshop%20Handout.pdf

Humanists' Listserv Data Set (note: very large and slow)



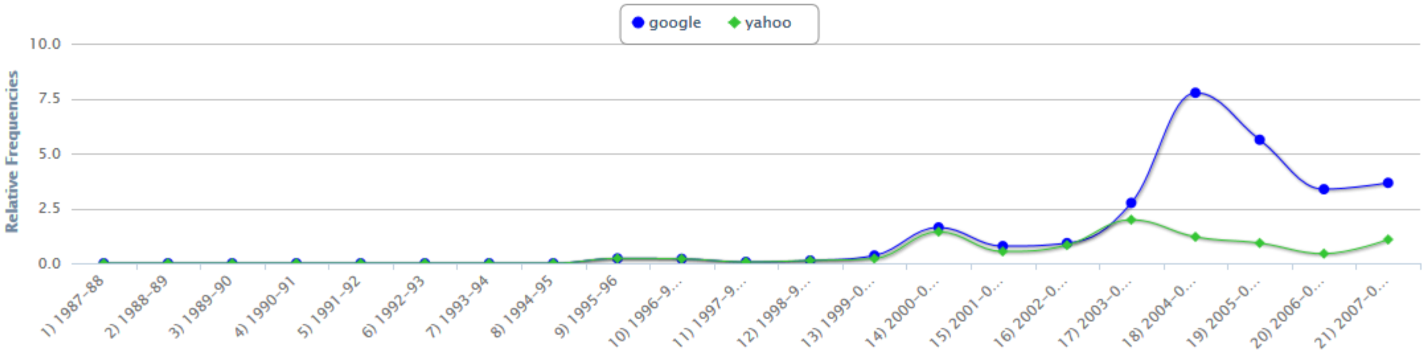
If you want to take a look at a fairly large data set, you can explore the Humanists' Listserv data, which contains all listserv emails among humanists (think H-Net) from 1987 to 2008. You can do so by clicking "Open" on the main voyant-tools screen. This can give you a change-over-time view of what sorts of things scholars were discussing over the past two decades. Below are two examples of the changes you can see in the data. You can replicate these results, or discover your own.

The Triumph of Google

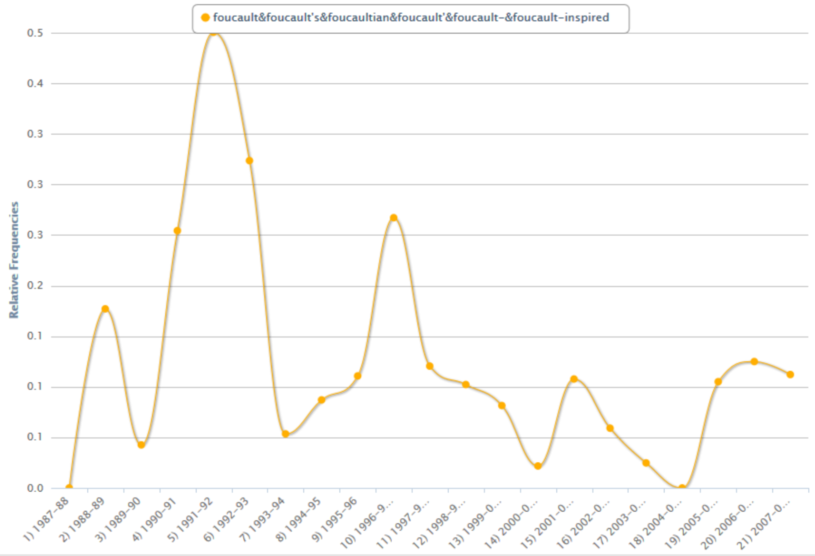


Words in the Entire Corpus

□ Frequencies	Count	Trend
☑ google	661	
☑ yahoo	350	



The decline and resurgence of Foucault


● foucault&foucault's&foucaultian&foucault'&foucault-&foucault-inspired
Relative Frequencies
1) 1987-88
2) 1988-89
3) 1989-90
4) 1990-91
5) 1991-92
6) 1992-93
7) 1993-94
8) 1994-95
9) 1995-96
10) 1996-9...
11) 1997-9...
12) 1998-9...
13) 1999-0...
14) 2000-0...
15) 2001-0...
16) 2002-0...
17) 2003-0...
18) 2004-0...
19) 2005-0...
20) 2006-0...
21) 2007-0...