

Computational Text Analysis for Content Analysis

Presented by Colleen Nugent & Chris McNulty

HIST 7370

Professor Dan Cohen

Fall 2022



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Workshop Agenda

- Introduction to definitions and key terms in computational text analysis (What)
- Discussion on its applications and uses in research (Why)
- Demonstration of select text analysis tools (How)
 - Lexos and AntConc

Slides, handouts, and data available at:

<https://bit.ly/fa22-cohen-textanalysis>



What is Computational Text Analysis?



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Computational Text Analysis

Text analysis is **a process to make inferences based on textual data**. Computational text analysis refers to the **array of methods used to “read” texts with a computer**. It is similar to statistical analysis, but the data are texts (words) instead of numbers.

Text analysis:

- Involves a computer drawing out patterns in a text, and a researcher interpreting those patterns.
- Includes methods such as word count frequency, nGrams, and sentiment analysis.
- Is conducted using web-based tools or coding languages like Python and



Why Computational Text Analysis?

Computational text analysis can help us **analyze very large amounts of data, identify keywords, and discover patterns** in a corpus of text(s). Using text analysis, researchers may **find surprising results** that they would not have discovered from close reading or traditional methods alone.

From collections of texts, researchers can **discover keywords that serve as a proxy for major trends in societies, cultures, and policies**. For example, computational tools can reveal patterns on how public officials communicate policies, how the language of policy changes over time, which phrases leaders regularly employ, and much more.



Key Terms

- **Corpus (plural–corpora):** A collection of texts used for analysis and research purposes.
- **Stopwords:** Words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stopwords include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.
- **nGram:** A continuous sequence of n items in a text. A bigram (or 2 continuous words) could be 'United States,' while a trigram (3 words) could be 'yes we can.'
- **Sentiment Analysis:** Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text's overall sentiment.



Discussion: Using Text Analysis for Historical Inquiry

What can digital text analysis reveal about historical sources that traditional close reading cannot?

- Scale
- Patterns (topic modelling)
- Style (stylometry)
- Sentiment (sentiment analysis)



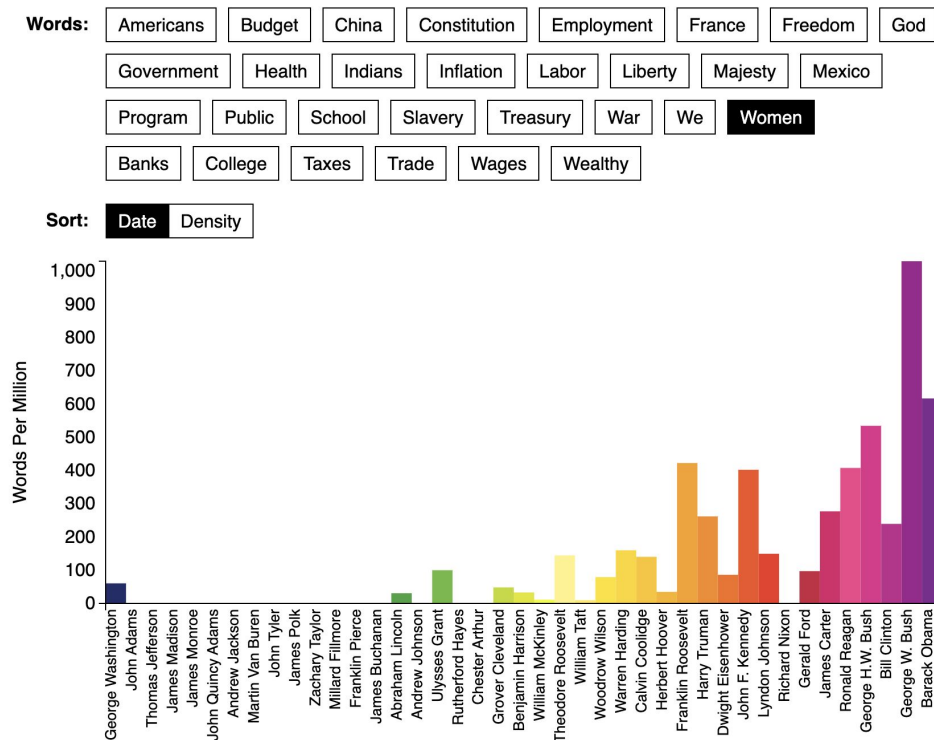
Examples from Practice



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

The Language of the State of the Union



Ben Schmidt's [The Language of the State of the Union](#) contains an interactive visualization that tracks the frequency of certain words across the entirety of Presidential States of the Union in the United States.

The example on the left demonstrates the frequency of the word “women” in State of the Union addresses. It highlights how women were not significantly addressed in these speeches until the turn of the century with Theodore Roosevelt.



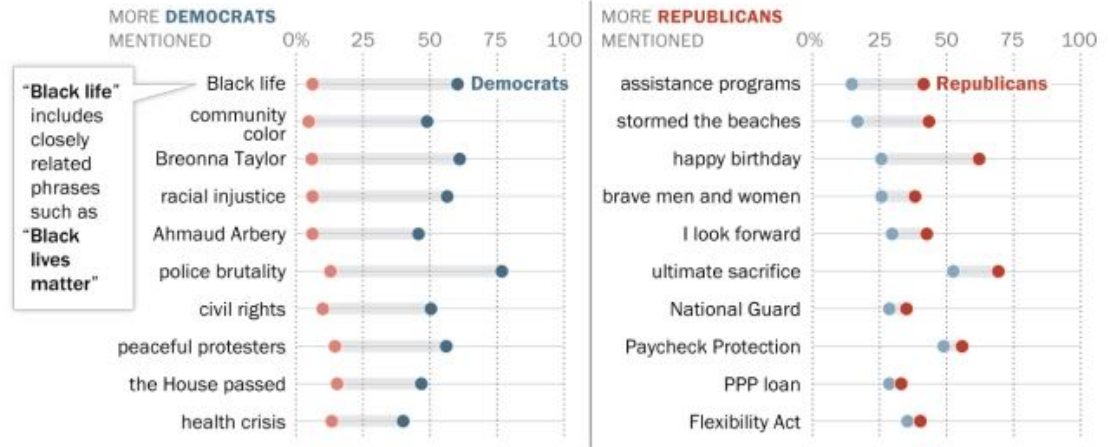
Content warning: police violence, racism

Posts mentioning 'Black lives matter' spiked on lawmakers' social media accounts after the death of George Floyd

- [Pew Research Center July 16, 2020 article](#)
- [Methodology](#)

In weeks following George Floyd killing, Democratic lawmakers' most distinctive language on social media focused on racial justice, police violence

Share of members in each party that mentioned ___ on Twitter or Facebook, May 25-June 14, 2020



Note: Chart shows the top 10 keywords based on how much more likely members of one party were to ever mention a keyword relative to the other party. Terms are displayed in their standardized form (e.g., "Black life" instead of "Black lives") and have been edited slightly in some cases for readability (e.g., "the House passed" instead of "house passed"). Keyword analysis was not case-sensitive. Words from retweets are included in this analysis even if the member who retweeted them did not create the original tweet.

Source: Pew Research Center analysis of congressional social media data from the Twitter API, Facebook Graph API and CrowdTangle, May 25-June 14, 2020.

PEW RESEARCH CENTER



Text Preparation



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Corpus Building

Questions to consider as you begin your research:

- What are my research questions and why am I creating a corpus?
- What am I asking my corpus to do? How can I use texts as a proxy for the historical phenomena I am studying?
- Where can I access relevant texts and how should I organize my corpus to streamline my research processes and save time?
- How representative are my selected texts for the historical phenomena being studied? How balanced are the selected texts, in terms of length, publication date, location, and genre?
- Are there any limitations in the corpus that will impact the questions it can help answer?



Preparing Your Text

- You will need a set of plain text (.txt) files for text analysis. These can often be downloaded directly as .txt files
 - If working from a non-plain text source, copy and paste non-plain text formatted text into a **plain text editor** (on Macs: Text Edit; on Windows: Notepad)
- Your corpus should be a set of plain text files in a folder or a set of folders containing plain text files
 - You will want to consider the order of information in your file names with sorting in mind; think about which pieces of information you will want to be able to sort your files by—and in what order
- *Best practices for naming files:*
 - Make sure to name your files with all the metadata that may be relevant to your analysis (ex: title and year of publication)
 - Avoid spaces and uppercase in your file names
 - Be consistent in your file-naming practices, documenting these as needed



Sample Corpus for Demo and Hands-On

Choose a set of plain text (.txt) files from:

The State of the Union Corpus

You can download directly from the above link. The following .txt files are also available on:

<https://github.com/NULabNortheastern/digitalassignment/showcase/tree/master/text-analysis/fa22-cohen-hist7370-textanalysis/sotu>



Lexos



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

Lexos

Lexos provides a step-by-step guide for text uploading, preparation, and analysis.

- **Upload:** upload your .txt file
- **Manage:** select the files you want to prepare and analyze
- **Prepare:** prepare your text for analysis
- **Visualize:** create visualizations of patterns across your corpus or in single texts
- **Analyze:** analyze your text

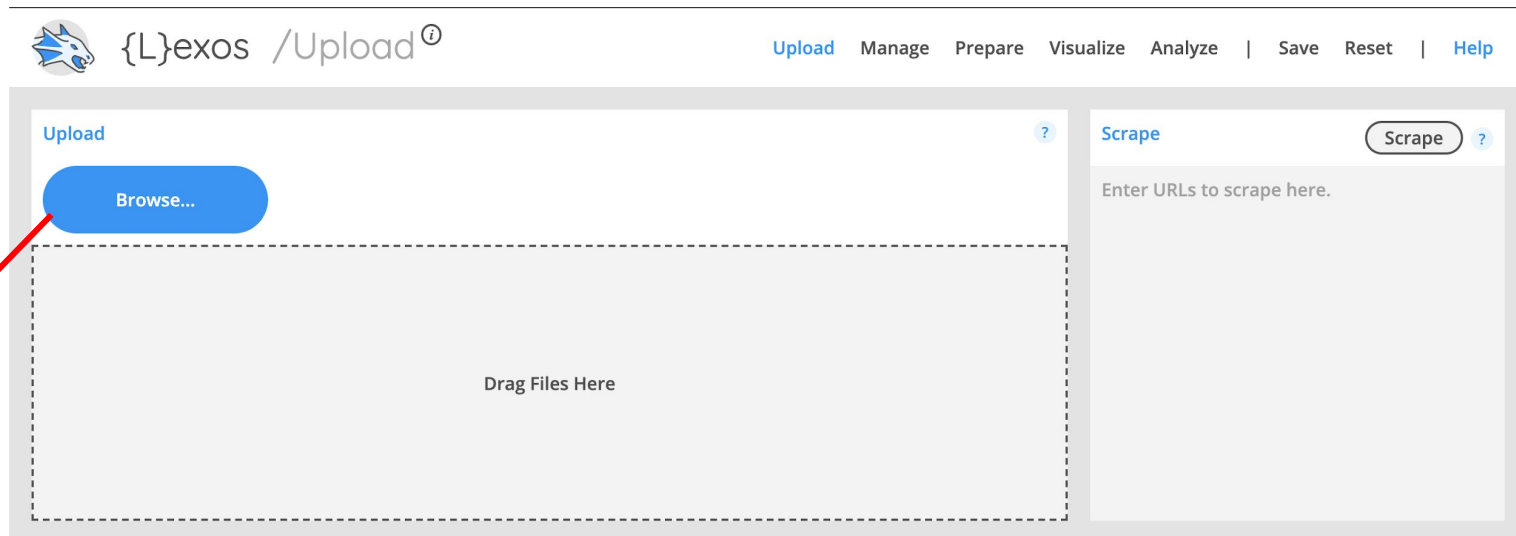
<http://lexos.wheatoncollege.edu/upload>



Lexos: Upload

Click Browse and select your entire text (or drag file into the “Drag Files Here” area)

The notification that each file has uploaded is easy to miss; click “Manage” to double check that the text file is there.



Lexos: Manage

Make sure the document you want to use is selected (blue = selected, gray = not selected)



{L}exos /Manageⁱ

Upload **Manage** Prepare Visualize Analyze | Save Reset | [Help](#)

Active	#	Document	Class	Source	Excerpt	Download	?
	1	Bush_2004		Bush_2004.txt	Mr. Speaker, Vice President Cheney, Members of Congress, distinguished guests, and fellow citizens: America this evening is a n... ..of the years. And in all that is to come, we can know that His purposes are just and true. May God continue to bless America.		
	2	Obama_2009		Obama_2009.txt	Madame Speaker, Mr. Vice President, Members of Congress, and the First Lady of the United States: I've come here tonight not o... ..is very chamber, "something worthy to be remembered." Thank you, God Bless you, and may God Bless the United States of America.		
	3	Trump_2017		Trump_2017.txt	Thank you very much. Mr. Speaker, Mr. Vice President, members of Congress, the first lady of the United Statesand cit... ..selves. Believe in your future. And believe, once more, in America. Thank you, God bless you, and God bless the United States.		



Lexos: Prepare (scrub)

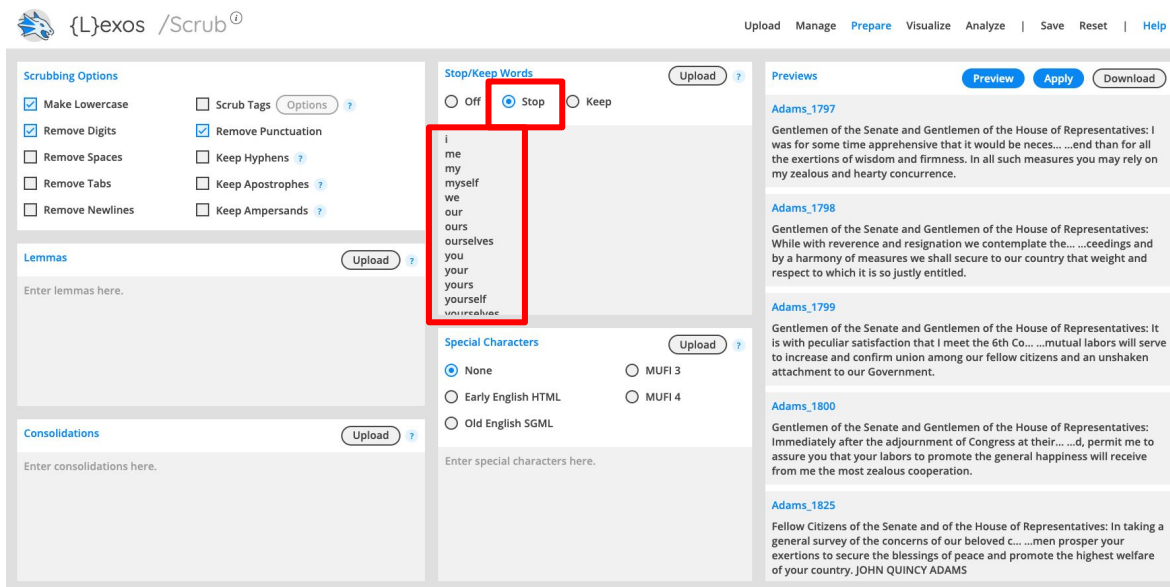
Lexos demonstrates some commonly used options for text preparation. By “scrubbing,” you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **Make Lowercase:** make all your letters lowercase. Even though you know “A” and “a” are the same letter, the computer treats these as two separate characters. Lowercasing removes this distinction.
- **Remove Punctuation:** remove punctuation, which may influence your results.
- **Stop/Keep Words:** remove a list of words. Usually these would be stopwords, or the most common words in a language (English: the, a she, her, it, him, they, etc).
- **Lemmas:** standardize to the *stem* of word. For example, you can stem all forms of talk: talking, talked, talks, etc. to “talk.”



Lexos: Removing Stopwords

Get a list of English stopwords here: <https://gist.github.com/sebleier/554280> (there is also a copy on the GitHub page). Copy and paste the stopwords (or upload the .txt file) into the “Stop/Keep Words” box then select “Stop.”



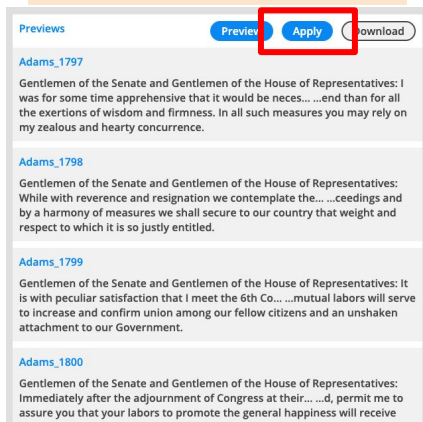
The screenshot shows the Lexos web interface. The top navigation bar includes links for Upload, Manage, Prepare, Visualize, Analyze, Save, Reset, and Help. The main interface is divided into several sections:

- Scrubbing Options:** Includes checkboxes for Make Lowercase, Remove Digits, Remove Spaces, Remove Tabs, Remove Newlines, Scrub Tags, Remove Punctuation, Keep Hyphens, Keep Apostrophes, and Keep Ampersands.
- Lemmas:** A text input field for entering lemmas.
- Consolidations:** A text input field for entering consolidations.
- Stop/Keep Words:** A section with three radio buttons: Off, Stop (selected), and Keep. Below the buttons is a text input field containing a list of stopwords: "i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", and "unintentionally".
- Special Characters:** A section with three radio buttons: None (selected), Early English HTML, and Old English SGML.
- Previews:** A section showing four preview cards for the text "Adams_1797", "Adams_1798", "Adams_1799", and "Adams_1800". Each card displays a snippet of text from the document.



Lexos: Applying your Preparations

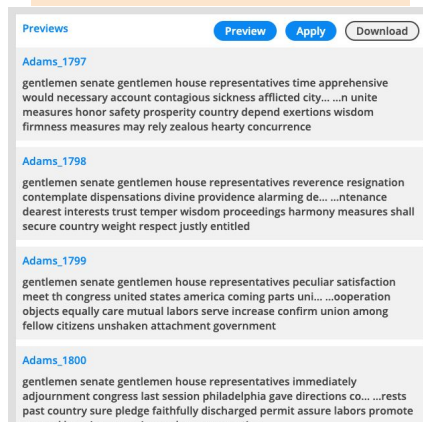
BEFORE PREP



The screenshot shows the Lexos interface with a list of documents. The 'Apply' button is highlighted with a red box. The documents listed are:

- Adams_1797: Gentlemen of the Senate and Gentlemen of the House of Representatives: I was for some time apprehensive that it would be neces... ..end than for all the exertions of wisdom and firmness. In all such measures you may rely on my zealous and hearty concurrence.
- Adams_1798: Gentlemen of the Senate and Gentlemen of the House of Representatives: While with reverence and resignation we contemplate the... ..ceedings and by a harmony of measures we shall secure to our country that weight and respect to which it is so justly entitled.
- Adams_1799: Gentlemen of the Senate and Gentlemen of the House of Representatives: It is with peculiar satisfaction that I meet the 6th Co... ..mutual labors will serve to increase and confirm union among our fellow citizens and an unshaken attachment to our Government.
- Adams_1800: Gentlemen of the Senate and Gentlemen of the House of Representatives: Immediately after the adjournment of Congress at their... ..d, permit me to assure you that your labors to promote the general happiness will receive

AFTER PREP



The screenshot shows the Lexos interface after preparation. The text is now tokenized and lowercased. The documents listed are:

- Adams_1797: gentlemen senate gentlemen house representatives time apprehensive would necessary account contagious sickness afflicted city... ..n unite measures honor safety prosperity country depend exertions wisdom firmness measures may rely zealous hearty concurrence
- Adams_1798: gentlemen senate gentlemen house representatives reverence resignation contemplate dispensations divine providence alarming de... ..ntenance dearest interests trust temper wisdom proceedings harmony measures shall secure country weight respect justly entitled
- Adams_1799: gentlemen senate gentlemen house representatives peculiar satisfaction meet th congress united states america coming parts uni... ..operation objects equally care mutual labors serve increase confirm union among fellow citizens unshaken attachment government
- Adams_1800: gentlemen senate gentlemen house representatives immediately adjournment congress last session philadelphia gave directions co... ..rests past country sure pledge faithfully discharged permit assure labors promote general happiness

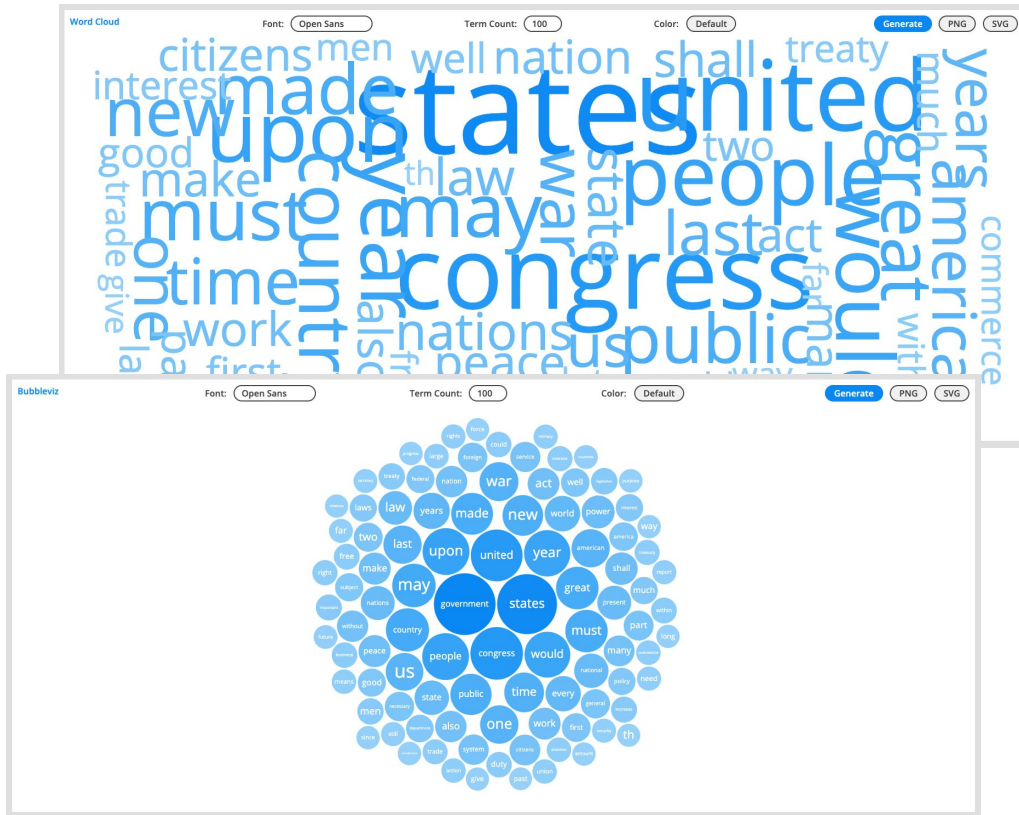
Once you have made decisions about your preparations, click “**Apply**” and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your new corpus.



Lexos: Visualize

Word Cloud: visualize a wordcloud across the entire text/corpus.

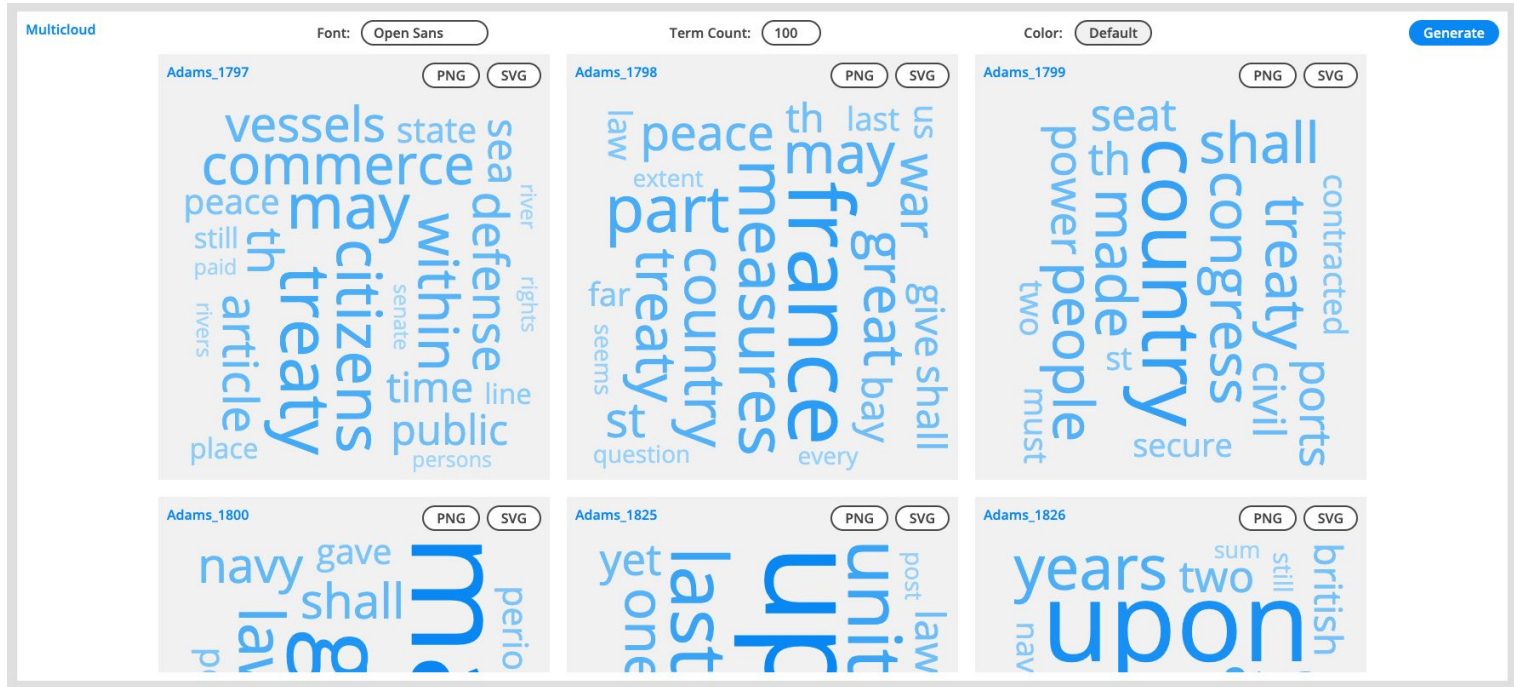
Bubbleviz: visualize word counts through bubbles across the entire text/corpus.



Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Lexos: Visualize > Multicloud



Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Lexos: Rolling Window

Rolling windows allow you to look at word trends across one document. To use a rolling window:

1. Go to “Visualize-> Rolling Window” and type in a search term you want to visualize. You can also search multiple terms by clicking “String” and separating words with a comma (heat, health, flood, storm).
2. Choose a Window size (the number of words each “window” contains). For shorter documents, you can try something between 300–500 words. For larger documents, you may want to make your window larger. Play around with the window size to see how your results change.
3. Click “Generate.”

Calculation Type <input checked="" type="radio"/> Rolling Average ? <input type="radio"/> Rolling Ratio ?	Search Terms ? <input type="text" value="class, information, property"/> <input checked="" type="radio"/> Words <input type="radio"/> Strings ? <input type="radio"/> Regex ?	Window <input ?<br="" type="text" value="200"/> <input checked="" type="radio"/> Words <input type="radio"/> Characters <input type="radio"/> Lines	Display <input type="checkbox"/> Milestone <input ?<br="" type="text" value=""/> <input type="checkbox"/> Show Individual Points <input type="checkbox"/> Black and White
--	--	---	--



Lexos: Analyze > Dendrogram

The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare.

Dendrograms are able to show the hierarchy between objects.

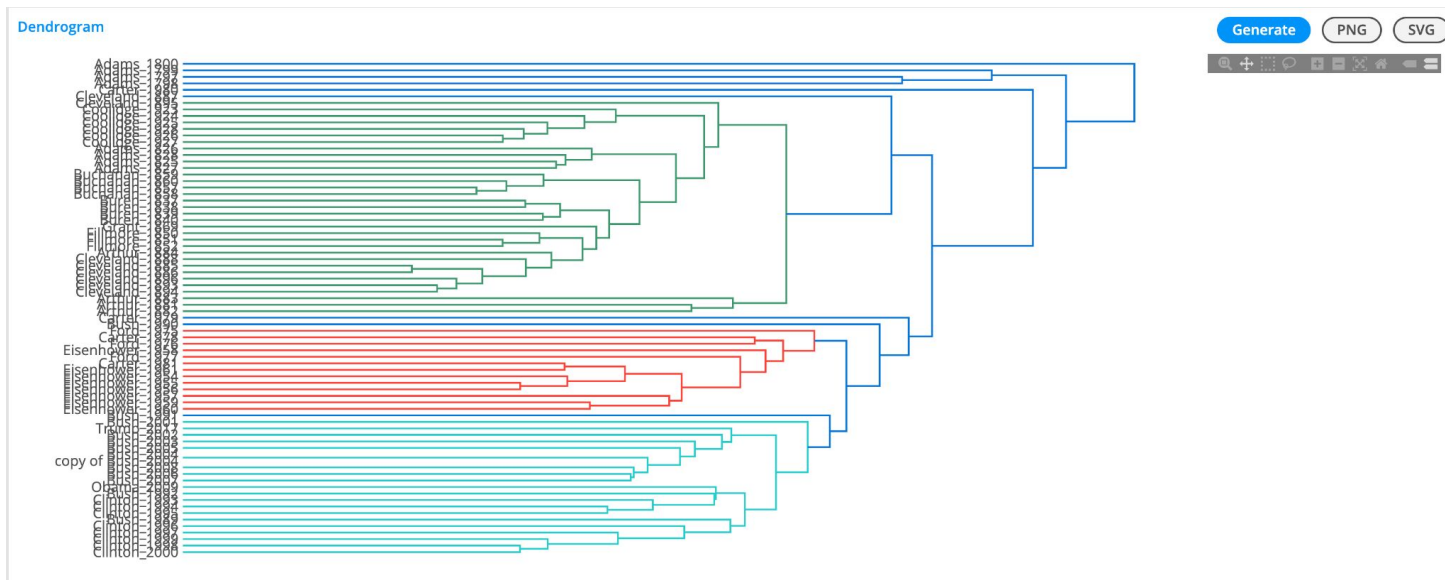
Dendrograms show:

- Similarities between texts
 - The greater the distance between texts, the less similar they are
 - The smaller the distance between texts, the more similar they are



Lexos: Dendrogram

The dendrogram demonstrates similarity between the different documents. The below example highlights how the dendrogram can identify the specific voice of each president, but the big picture view makes it less useful.



Lexos: Dendrogram

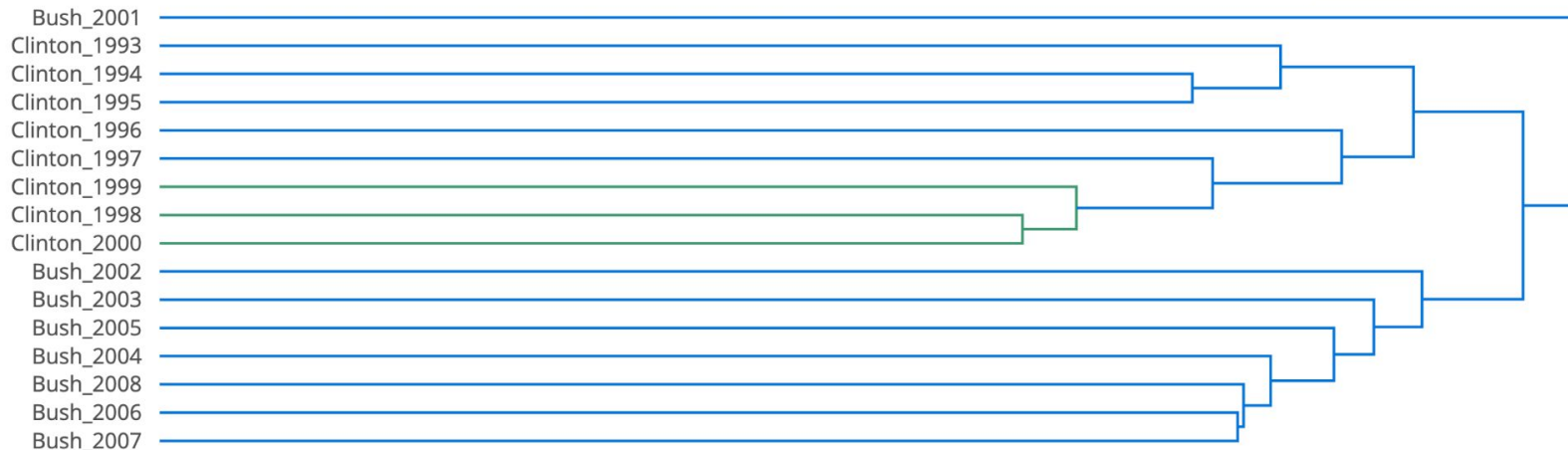
This example only compares the State of the Union speeches from the Clinton administration (1993-2000) and the Bush administration (2001-2008). The results are more easily interpretable when the tool is used on a smaller subset of the corpus.

Dendrogram

Generate

PNG

SVG



Lexos: Save or Reset Your Results

Lexos allows you to **save** your results as a Lexos file. If you do this, you can re-upload the Lexos file any time to access your cleaned-up corpus as well as the different analyses you've done. You can also download modified text files from the “Manage” page, to use these with other tools.

You can also save individual visualizations as images (PNGs).

Finally, if you want to start over, you can “Reset” your Lexos dashboard.



AntConc



Northeastern University
NULab for Texts, Maps, and Networks

*Feel free to ask questions at any point
during the presentation!*

What is AntConc?

A freeware corpus analysis toolkit for concordancing and text analysis. Allows for searching across your entire corpus.

AntConc performs better with many small files, rather than one or two large ones—there is no limit on how many words you can analyze, but larger corpora will take longer to work with.



Downloading AntConc

<https://www.laurenceanthony.net/software/antconc/>

For PC users: download **Windows Installer**

For Mac users: download **MacOS**

For Linux users: download **Linux (Portable)**



AntConc

A freeware corpus analysis toolkit for concordancing and text analysis.

[\[AntConc Homepage\]](#) [\[Screenshots\]](#) [\[Help\]](#) [\[License\]](#)

Downloads:

Official releases

- [Windows \(Installer\) \(4.1.2\)](#) [Recommended]
- [Windows \(Portable\) \(4.1.2\)](#)
- [MacOS 10/11 \(4.1.2\)](#)
- [Linux \(Portable\) \(4.1.2\)](#)

AntConc 3x series

- [Windows \(3.5.9\)](#)
- [MacOS 10 \(3.5.9\)](#)
- [Linux \(3.5.9\)](#)

Older versions

- [Complete history of all versions released.](#)

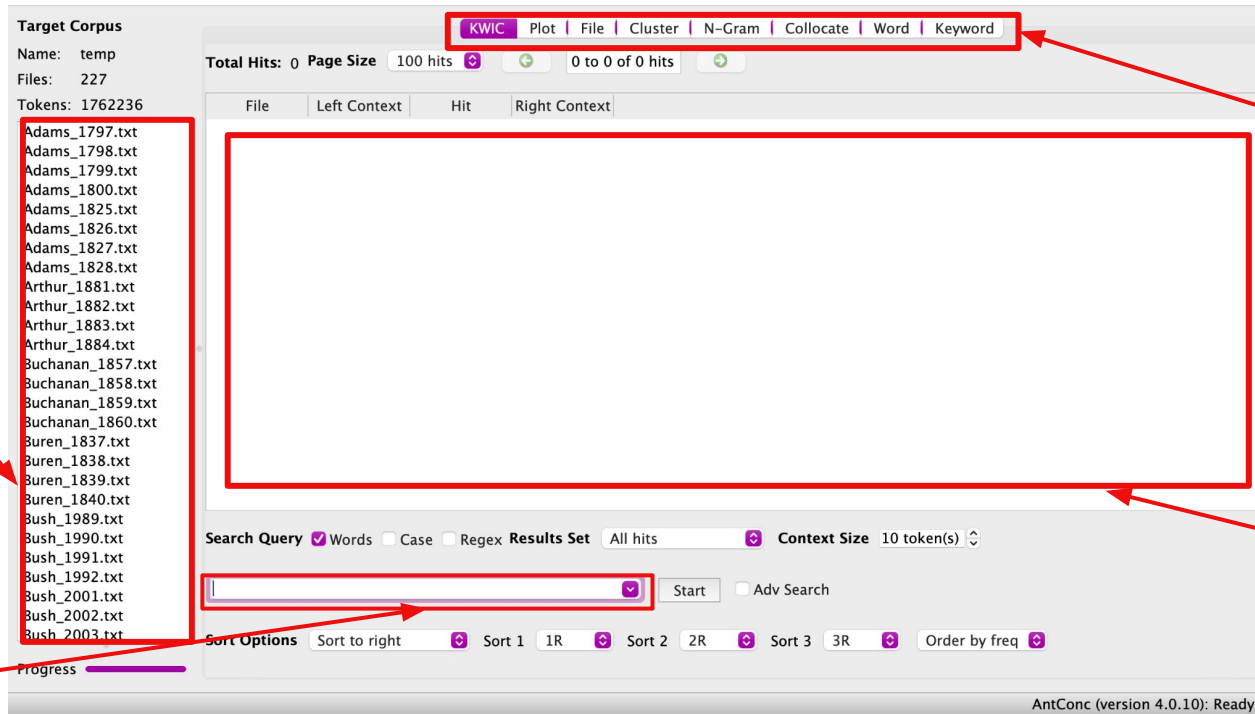


Preparing Corpus for AntConc

- AntConc can only process plain text (.txt) files.
- These plain text files should be prepared prior to uploading into AntConc (i.e. removing stopwords, regularizing capitalization).
- You can use Lexos to prepare your texts for analysis within both Lexos and AntConc.



Anatomy of AntConc: Mac



Corpus Window

Text Analysis Options

Results Window

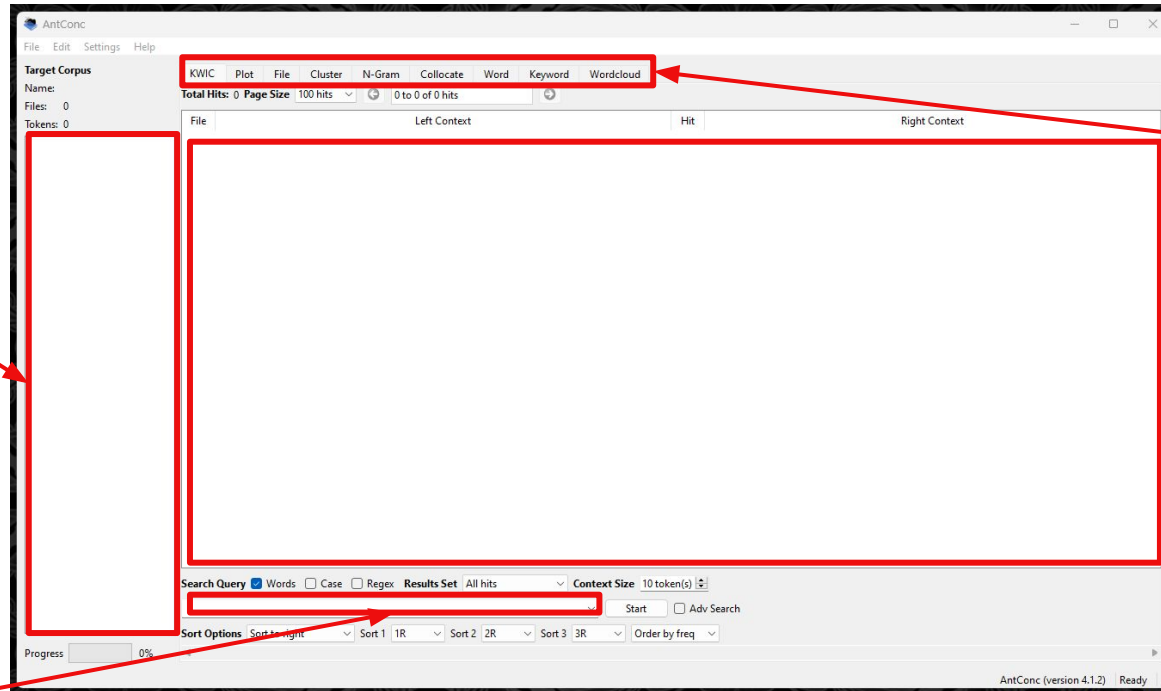
Search Bar



Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Anatomy of AntConc: PC

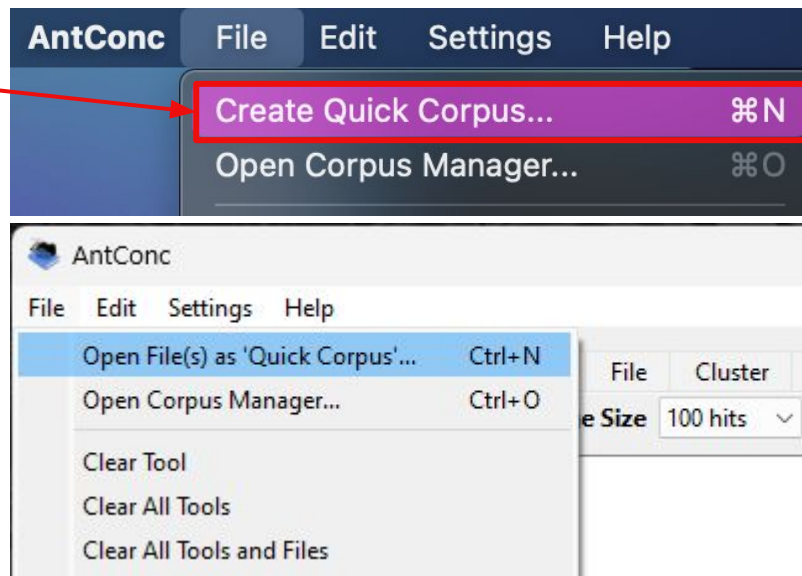


Uploading Your Corpus

After opening AntConc, go to “File” and then “Create Quick Corpus” or “Open File(s) as ‘Quick Corpus’”

Then select all of the .txt files you would like to use in your analysis and upload into AntConc!

You will see the names of all the .txt files in the Corpus Window.



Concordance Tool (KWIC)

	File	Left Context	Hit	Right Context
1	Johnson_1868.txt	free state, the right of the people to keep and	bear arms	shall not be infringed." It is believed that
2	Wilson_1915.txt	mandated that "the right of the people to keep and	bear arms	shall not be infringed," and our confidence has
3	Lincoln_1863.txt	y service, about one-half of which number actually	bear arms	in the ranks, thus giving the double advantage
4	Clinton_1995.txt	do anything to infringe on the right to keep and	bear arms	to hunt and to engage in other appropriate
5	Johnson_1867.txt	nse; a large proportion even of the persons able to	bear arms	were forced into rebellion against their will, and
6	Madison_1813.txt	d States being there under certain circumstances to	bear arms,	whilst of the native emigrants from the United

Search Query ☒ Words ☐ Case ☐ Regex

Results Set All hits Context Size 10 token(s)

Start ☐ Adv Search

This tool shows search results in a 'KWIC' (KeyWord In Context) format. This allows you to see how words and phrases are commonly used in a corpus of texts.

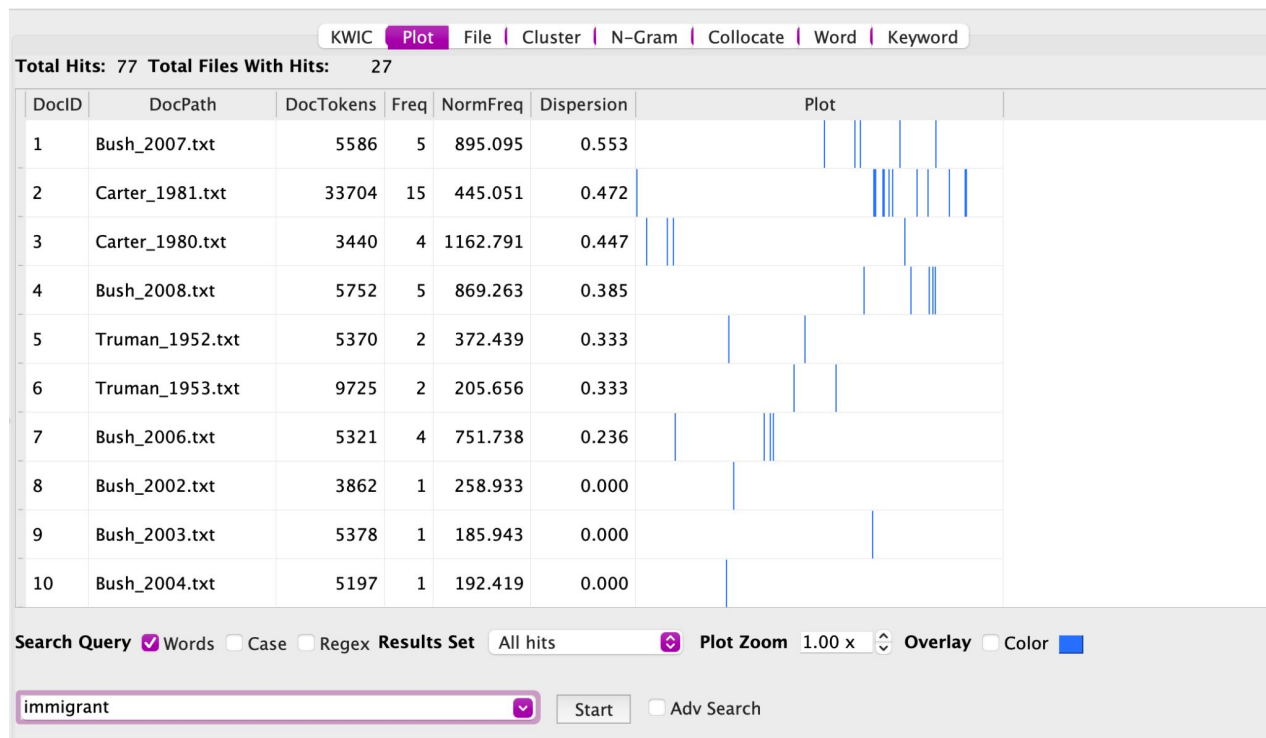
Type a term or phrase into the search box. AntConc allows you to search for both single words and full phrases.



Concordance Plot Tool

This tool shows search results plotted in a 'barcode' format. This allows you to see the position where search results appear in target texts.

Type a word or phrase into the search box to utilize this tool.



N-Gram Tool

	Type	Rank	Freq	Range
1	military academy at west	1	8	6
2	military force on the	2	7	6
3	military and naval commanders	3	6	4
4	military and naval establishments	4	5	4
5	military and naval forces	4	5	3
6	military force of the	4	5	5
7	military and economic aid	7	4	4
8	military establishment including river	7	4	4
9	military forces of the	7	4	4
10	military governor of the	7	4	1
11	military service of the	7	4	4
12	military and naval officers	12	3	3
13	military and naval operations	12	3	3
14	military strength of the	12	3	3
15	military and civilian pay	15	2	2
16	military and civilian personnel	15	2	2

Search Query ☒ Words ☐ Case ☐ Regex

N-Gram Size 4

Open Slots 0

Min. Freq 1

Min. Range 1

military

Start

☐ Adv Search

The N-Grams Tool scans the entire corpus for 'N' (e.g. 1 word, 2 words, ...) length clusters. This allows you to find common expressions in a corpus.

You can adjust the length of the clusters using the “N-Gram Size” option near the search bar.



Collocate Tool

This tool allows you to investigate non-sequential patterns in language.

This tool provides the frequency of words appearing to the left or the right of the search query. The value measures how 'related' the search term and the collocate are.

You can try different sorting options, including increasing the window span, or number of words near the search query.



Northeastern University
NULab for Texts, Maps, and Networks

	Collocate	Rank	FreqLR	FreqL	FreqR	Range	Likelihood	Effect
1	force	1	80	4	76	53	273.093	3.791
2	establishment	2	62	10	52	47	261.177	4.401
3	naval	3	52	10	42	31	201.347	4.144
4	forces	4	54	8	46	33	189.064	3.860
5	our	5	270	204	66	100	185.823	1.376
6	strength	6	48	5	43	27	171.964	3.923
7	academy	7	23	0	23	17	156.051	6.283
8	operations	8	31	1	30	21	103.958	3.747
9	commanders	9	17	2	15	13	99.277	5.604
10	posts	10	18	0	18	16	91.143	5.037
11	civil	11	32	23	9	24	83.274	3.153
12	service	12	51	8	43	36	83.184	2.328
13	economic	13	37	16	21	20	77.821	2.740
14	power	14	49	5	44	32	64.910	2.046
15	defense	15	29	13	16	21	60.999	2.741
16	personnel	16	12	1	11	8	57.728	4.851

Search Query ☒ Words ☐ Case ☐ Regex Window Span From 5L To 5R Min. Freq 1 Min. Range 1

military ☒ Start ☐ Adv Search

If you double click on a word, it shows you that term in the Concordance (KWIC) tool to show the word in its original contexts.

Feel free to ask questions at any point during the presentation!

Your Turn!

Use the sample text or texts of your choice and begin practicing text analysis with **Lexos** and **AntConc**. Try new words with the features we showed, or explore new options for analyzing texts.

Discussion Prompts

- What kind of research can these tools enable?
- What are the limitations of these tools?
- What sort of theoretical assumptions are these tools making about text as a source or a subject of study?



Thank you!

If you have any questions, contact us at nulab.info@gmail.com

Developed by Colleen Nugent and Chris McNulty

DITI Research Fellows

Digital Integration Teaching Initiative

- Slides, handouts, and data available at:
<https://bit.ly/fa22-cohen-textanalysis>
- We'd love your feedback! Please fill out a short survey here:
<https://bit.ly/diti-feedback>
- Schedule an appointment with us! <https://calendly.com/diti-nu>

