

COMPUTER-ASSISTED TEXTUAL ANALYSIS with CLiC, Voyant Tools & Antconc

Najla Jarkas
Mellon Fellow for DLA Programs
The AMICAL Consortium
Nov. 16, 2022



OVERVIEW



❖ Part 1

- Definitions
- Choosing the right digital tools
- Textual analysis
- Distant VS Close reading
- Ngrams
- EEBO Ngram Browser
- CLiC

❖ Part 2

- Compiling a corpus for distant reading
- Voyant Tools

❖ Part 3

- Preparing a corpus for Antconc
- Antconc
- General discussion
- Optional assignment for our next session

+

•

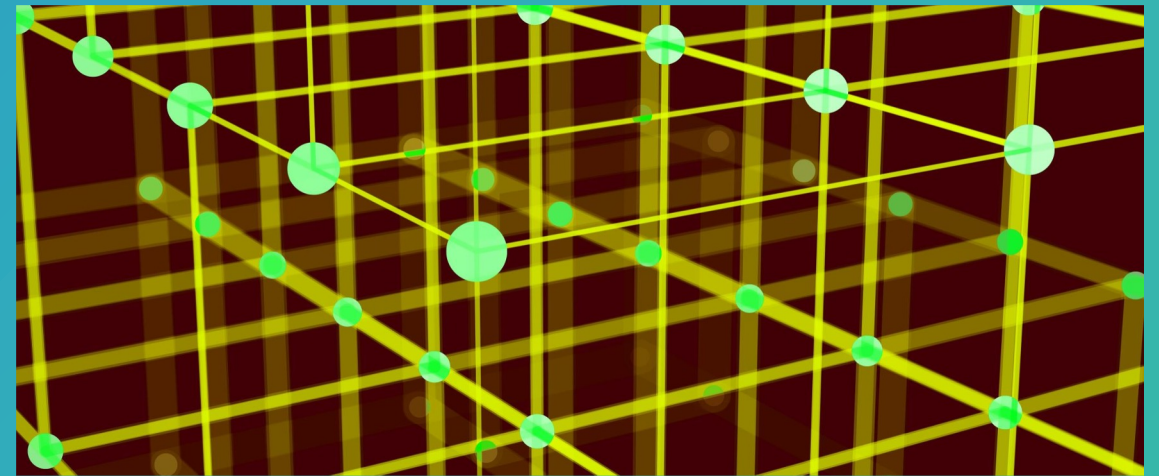
○

Part 1

DEFINITIONS

"Using computational methods and approaches to augment and enhance traditional humanities disciplines."

The use of digital methods and tools to "re-examine", "compel" and "read" literary texts



What Is Digital Humanities?

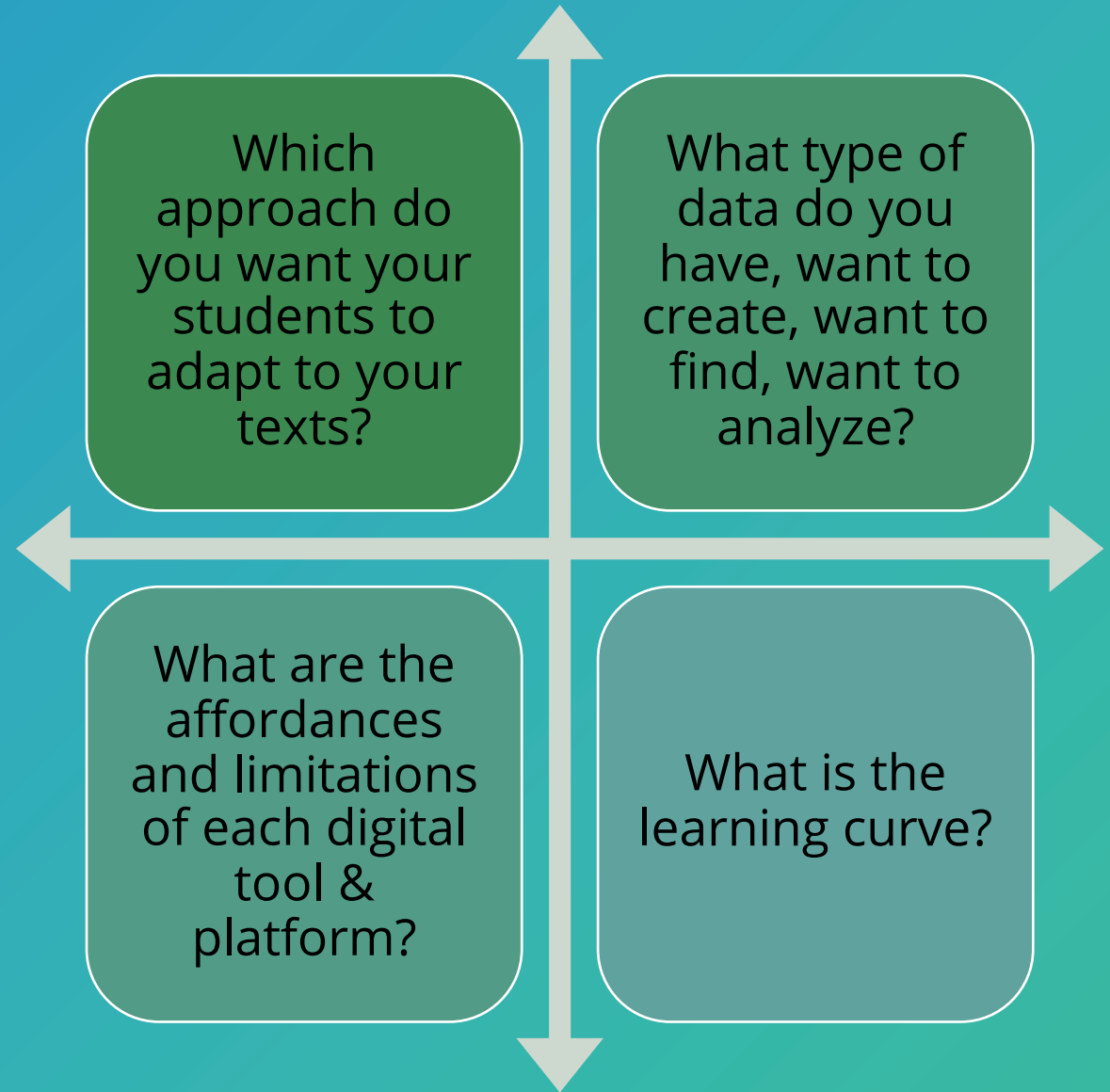
Digital humanities is the use of computational tools or approaches to find patterns in humanistic production, when those patterns are then used for interpretive purpose. *Brian Croxall*

NB: Refresh the page to get a new definition. Quotes were pulled from participants from the [Day of DH](#) between 2009-2014. As of January 2015, the database contains 817 rows and randomly selects a quote each time the page is loaded. If you want to do something cool with the data, I am providing the code and data [here](#).

Made by Jason Heppner. Problems? Questions? I'm @jaheppner on Twitter.

Whatisdigitalhumanities.com

Choosing your digital methods and tools



+

-
- # Textual Analysis & Close Reading

Textual analysis is not new to college students

- It involves close reading, which is one of the most widespread scholarly methods in literary criticism
- It is an indispensable tool for academic readers/students in many courses and fields
- By staying close to a text, the reader studies the language carefully, interprets its deeper meanings, and believes there is value in the text itself

+

•

○

Textual Analysis & Close Reading

Example of a close reading exercise of a poem:

- ❖ Title: What strikes you about the title? What do you expect the poem to be about?
- ❖ Keywords: Try to circle the keywords in the poems? What do you notice? What do you notice about nouns, adjectives, and verbs?
- ❖ Word Choice: What do you notice about the selected words in the poem? Are they pertinent to certain fields?
- ❖ Punctuation: What do you notice about punctuation?
- ❖ Figurative Language: Are there any interesting figures of speech or metaphors in the poem? What is their significance?
- ❖ Form: What do you notice about the form of the poems? How many lines do they consist of?
- ❖ Illustrations: How are illustrations helpful in interpreting the poems?
- ❖ How can you relate all these formal elements to the meaning of the poem?

+

•

○

Distant Reading

Computer-assisted textual analysis, means using a computer to help with the above plus another type of reading called **Distant Reading** which can:

- Survey hundreds and thousands of texts
- Identify patterns across centuries and boundaries
- Draw on natural and social sciences
- Utilize visualizations like graphs, maps, word clouds, scatter plots, etc.

+

•

○

Distant Reading

- uses computers to help process texts, explore research questions, and assist with sense making ("distant reading")
- finds patterns of recurring words, collocates, in singular or multiple texts across centuries and boundaries
- works with larger number of texts, using a computer to identify features of those texts, working with surface level phenomena to look for patterns, relies on visualization to represent such lexical patterns
- assists, complements, and facilitates the reading of large sets of texts (corpus)
- applies a *quantitative* and *qualitative* approach to reading texts
- does not replace human reading

Distant Reading

Quantitative text analysis (1850s)

Counting of words goes back in time:

Comparing word counts

Viewing words in context

Calculating frequencies / statistical
relevant

“Like science experiments, there is value in
reproducibility.”

Cf. Rockwell/Sinclair, “Thinking-
Through”

[El Khatib, Randa and Najla Jarkas. “Digital Humanities in the
Classroom.” Winter Institute in Digital Humanities. 19-22
January 2020.]

Distant Reading

Why do we need Computer-Assisted "reading" applications?

- "Computers are good at mechanical processes, such as searching, counting, and sorting into alphabetical or numerical order. They perform these tasks not only much faster than a human being, but also very much more accurately." (Hockey, 2000, p. 3-4)

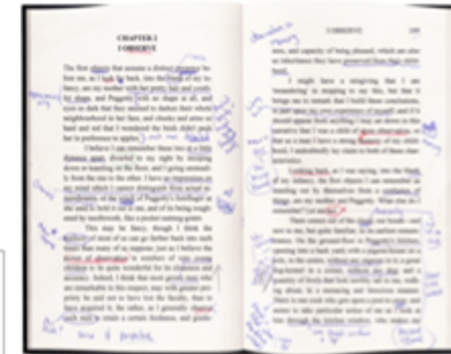
But ...

- interrogation of texts can't be done fast
- Researchers working with visualization tools need time and a lot of patience to reach meaningful results

Distant and Close Reading

Distant & Close reading

- alternate between distant and close reading
 - **Step 1:** Do a distant reading of the text and identify interesting themes, keywords, and formulate a research question
 - **Step 2:** Close read the text to examine the significance of passages
 - **Step 3:** Go back to distant reading... to revisit and refine your research question or formulate new ones



Main figures in Distant Reading

- [Franco Moretti](#) (1950-) Italian literary historian and theorist, called people to abandon their close reading practices and adopt distant reading.
- “[W]e know how to read texts, now let’s learn how *not* to read them.” (Moretti, 2000, p. 57)

Examples of Distant Reading

Moretti's *Graphs, Maps, Trees: Abstract Models for Literary History* (2005)

FIGURE 12: Authorship of new novels, Britain 1800–1829: gender breakdown (percentage)



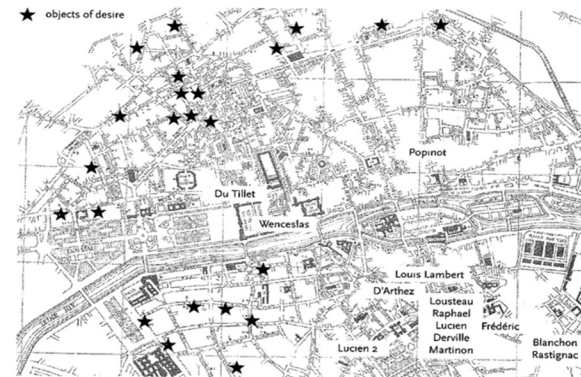
Source: Garside, Raven and Schöwerling, eds, *The English Novel 1770–1819*.

Examples of Distant Reading

from Moretti's *Graphs, Maps, Trees: Abstract Models for Literary History* (2005)

Quantitative approaches to the literature of space, place and landscape allow for new readings and approaches to literature

FIGURE 22: Protagonists of Parisian novels, and objects of their desire

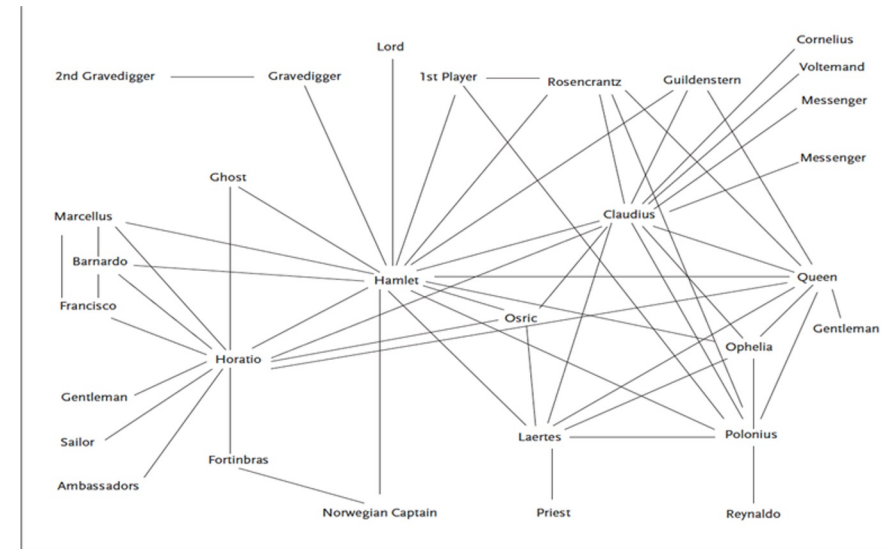


Protagonists of Parisian novels and objects of their desire
(Moretti 2005)

Examples of Distant Reading

The *Hamlet* Network (Moretti, 2011)

Network visualizations of *Hamlet* show another example of computer-assisted applications "reading" a text for us



+

•

○

Textual analysis

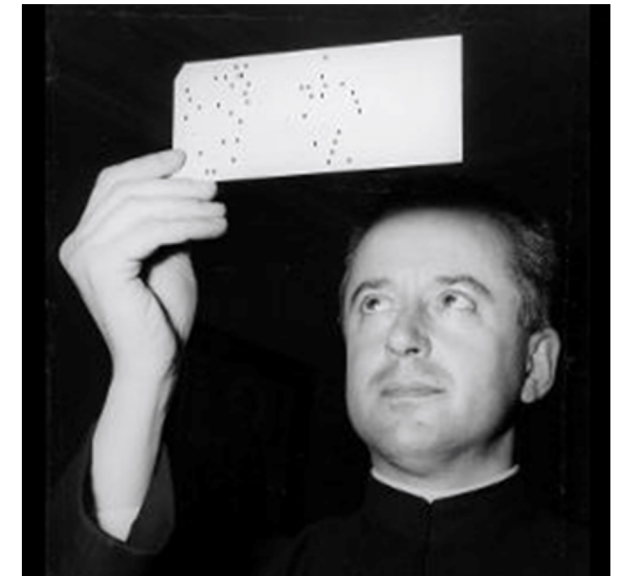
Precursors of distant reading

Josephine Miles (1911-1985): adopted quantitative and computational approaches to analyze poetry in the 1930s and 1940s

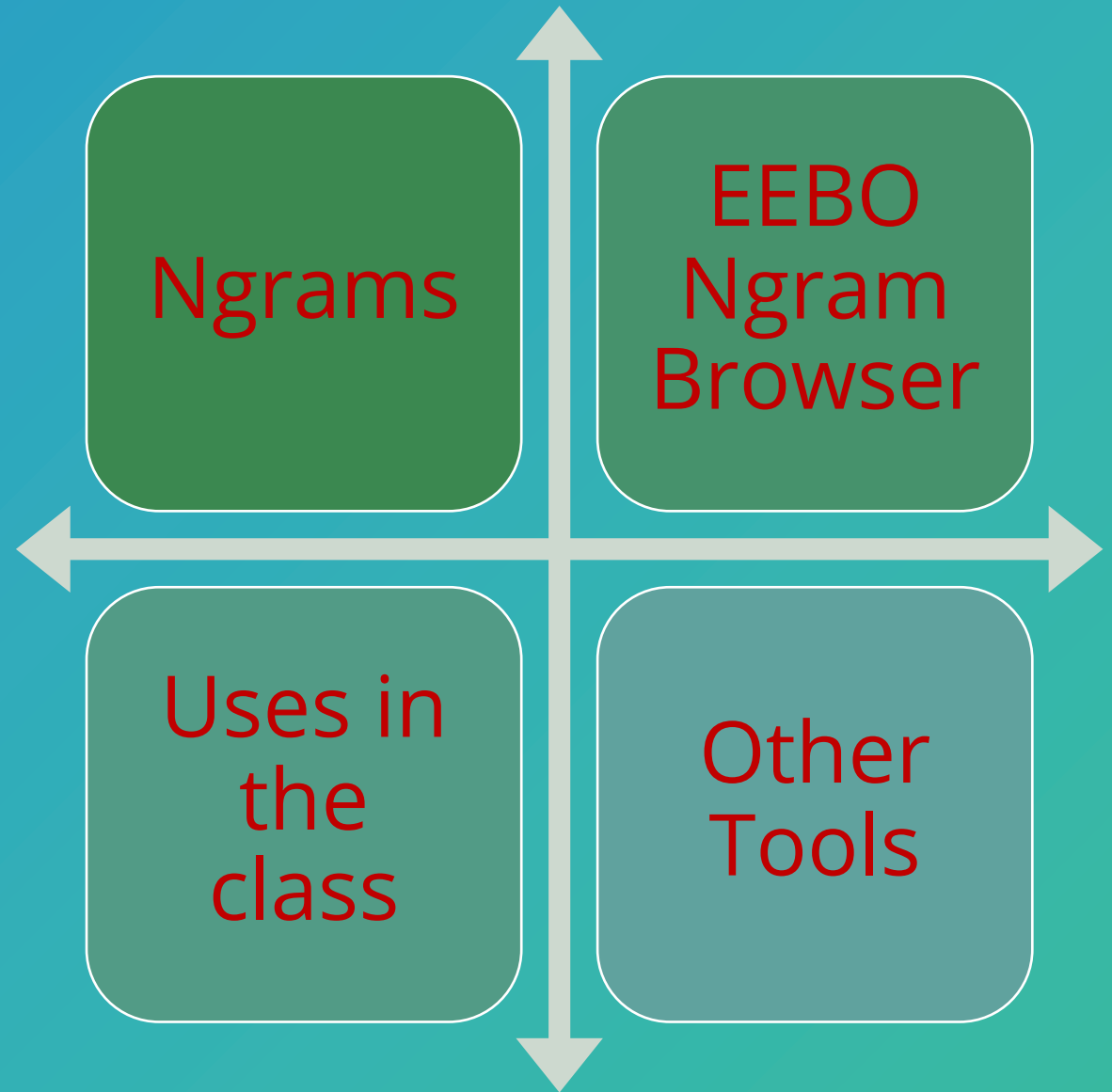
Examined adjectives in Poems from [Wyatt to Auden](#) (1946)



Father Roberto Busa (1913-2011): created the *Index Thomisticus* marking the beginning of DH for linguistic and literary analysis



Distant Reading in practice: Ngrams and EEBO Ngram Browser



Ngrams

- <https://books.google.com/ngrams/>
- **Google Books** (previously known as **Google Book Search** and **Google Print** and by its codename **Project Ocean**)^[1] is a service from [Google Inc.](#) that searches the full text of books and magazines that Google has scanned, converted to text using [optical character recognition](#) (OCR), and stored in its digital database.^[2] Books are provided either by publishers and authors, through the Google Books Partner Program, or by Google's library partners, through the Library Project.^[3] Additionally, Google has partnered with a number of magazine publishers to digitize their archives
- Search for the following on the NGram Viewer:
 1. Different themes in literature: Love, Death, Mercy.
 2. A word from a poem and two related/similar words.
 3. The title of a poem/book(in the Google Ngram Viewer, you need to abbreviate it to five words) and the titles of two other modernist poems.
 4. Means of transportation: trains, horse, metro, subway, plane, etc

Ngrams

- Through Google's language-analysis tool you can search for verbs, nouns of the same word by using tags and wildcards:
- <https://www.lifewire.com/google-books-ngram-viewer-1616701>
- Another example, is to search for the vocabulary of the author to spot word trends:
- Emily Dickinson for examples: search for her vocabulary in one of her poems and compare to another poet's poem on the same theme
- Usage of offensive words: nigger, negro, slave, black
- Historians can “plot the trajectories of words and phrases over time based on an enormous corpus of data extracted from the Google Books digitization project”
<https://www.theatlantic.com/technology/archive/2013/10/googles-ngram-viewer-goes-wild/280601/>

+

•

○

EEBO Ngram Browser

<https://earlyprint.wustl.edu>

Click on EEBO Ngram Browser and search for the following terms: press “Draw Graph”

1. “love” and “honor”
2. “king” and “queen”
3. “man” and “woman”

What do you notice?

Other Distant Reading tools

- **WordCake:** <http://www.carrieroy.com/wordcake.html>
- **WordArt:** <https://wordart.com>
- **TreeCloud:** <http://treecloud.univ-mlv.fr/index.php>
- HathiTrust Bookworm: <https://bookworm.htrc.illinois.edu/develop/>
- Distant Reader: <https://distantreader.org/>
- Learn more: <https://libguides.sdsu.edu/digitalhumanities/create>



CLiC

CLiC (Corpus Linguistics in Context)

- Free web app
- Developed by the University of Birmingham and the University of Nottingham
- Builds on Corpus Linguistics / Corpus Stylistics
- Used to analyze literary texts and fictional characters
- Includes 19th Century corpora and more
- Link: <http://clic.bham.ac.uk/>

+

•

○

CLiC

CLiC consists of four 19th Century corpora:

- DNov - Dickens's Novels
- 19C - 19th Century Reference Corpus
- ChiLit - 19th Century Children's Literature Corpus
- ArTs - Additional Requested Texts

+

•

○

CLiC Functions: Keywords

Keywords are by definition words whose occurrence in a text is statistically significantly higher than their occurrence in another reference corpus.

“... analysing the usage of keywords can serve as a starting point for the literary interpretation of data” (Fischer-Starcke, 2010, p. 65).

A “reference corpus” is a hazy notion

+

•

○

CLiC Functions: Key Clusters

Key clusters are very similar to key but they focus on groups of words. They could refer to body parts, time and place, etc. (see Mahlberg, 2013 for a categorization)

Michaela Mahlberg, *Corpus stylistics and Dickens's fiction*. London: Routledge, 2013. ISBN 780-0-415-80014-3.

+

•

○

CLiC Functions: Concordances

A **concordance** analysis reveals the context in which a particular word occurs

+ CLiC in practice: Keywords in *Wuthering Heights*

- **Go to Keywords**
 - Select *Wuthering Heights* in the Target Corpora
 - Select “All Text”... within subset
 - Select “19th Century Reference Corpus”
 - Select “All Text”... within subset
 - n-gram: 1-gram
- What conclusions can you make about the novel?
- Who are the main characters in the work?
- Now select n-gram: 3-grams.
- What new conclusions can you now make?

+ CLiC in practice: “Kitchens” in *Wuthering Heights*

Go to Clusters

1. Search the corpora: *Wuthering Heights*
2. Only in subsets: All text
3. n-gram: 3-gram
4. Filter rows: **kitchen**

What can you tell about kitchens in the novel?

What other spaces/words stand out?

+ CLiC in practice: “Eating” in *Wuthering* *Heights*

Go to Concordance

1. Search the corpora: *Wuthering Heights*
2. Only in subsets: All text
3. Search for the term: **eat**

What do you notice about food consumption in the novel?

What other words/activities stand out?

+ • CLiC in practice: “The fireplace pose” in Dickens’s works

Go to Concordance

1. Search the corpora: *Dickens’s Novels*
2. Only in subsets: All text
3. Search for the phrase: **back to the fire** or **the fireplace**
4. What do you notice?

Michaela Mahlberg (2013) analysis of the gendered “fireplace pose” in Dickens

“Corpus Linguistics in Action: The Fireplace Pose in 19th Century Fiction”

<https://programminghistorian.org/posts/corpus-linguistics-in-action>

+

•

○

Why do we need Distant Reading?

“Computers are good at mechanical processes, such as searching, counting, and sorting into alphabetical or numerical order. They perform these tasks not only much faster than a human being, but also very much more accurately.” (Hockey, 2000, p. 3-4)




Reflections on CLiC

- ❖ What can you do in CLiC?
- ❖ What can't you do?
- ❖ When and where will you be using it?



Part 2



Compiling a corpus for Distant Reading?

- **So what's a corpus--again?**
- “The word *corpus* is Latin for body – a corpus is therefore a ‘body’ of [digitized] texts.” —Paul Baker (2010)
- A corpus is a collection of (small or large) naturally occurring texts that have been digitized for analysis.
- **Why do I need a corpus?**
- Distant reading (and Digital Humanities in general) requires that you examine texts in digitized format.

Compiling a corpus for Distant Reading?

- Today, numerous texts have been digitized and are available freely on numerous portals:
 - Project Gutenberg: <https://www.gutenberg.org>
 - Internet Archive: <https://archive.org>
 - Open Library: <https://openlibrary.org>
 - Many Books: <https://manybooks.net>
 - BNF Gallica: <https://gallica.bnf.fr/>
 - More can be found in online repositories

You may need to assemble your own corpus and, in extreme cases, you may need to scan and OCR texts.

Dialogica

To save yourself time today, download texts from Gutenberg or the Internet Archive

Classroom activity:

- ❓ Together with your team, try to access these websites. What kinds of texts do they include? What language are these texts in? Does it include recent or ancient books?
- ❓ Can you find works by your selected author on these platforms? In which format are they available?
- ❓ Try to download all the texts you can find by your selected author.

+ • The Public Domain

Starting at midnight on January 1, tens of thousands of books (as well as movies, songs, and cartoons) entered the public domain, meaning that people can download, share, or repurpose these works for free and without retribution under US copyright law.



Per the [Sonny Bono Copyright Term Extension Act](#) of 1998, "corporate" creations (like Mickey Mouse) can be restricted under copyright law for 120 years. But per an amendment to the act, works published between [1923 and 1977](#) can enter the public domain 95 years after their creation. This means that this is the [first year since 1998](#) that a large number of works have entered the public domain.

- Every year, thousands of books are added to the public domain on **January 1**.
- This year, works created in 1925 were added to the public domain.
- Here's the link to the list of books added this year: <https://web.law.duke.edu/cspd/publicdomainday/2022/>

HathiTrust Retweeted



HathiTrust @hathitrust · Jan 2

#PublicDomainDay numbers: Yesterday HathiTrust opened 43,104 volumes published in U.S. with expired copyright, on top of 10,305 volumes from 1923 already opened. Includes 8,806 separate titles, including serials and gov docs. See entire 1923 collection at is.gd/HathiTrust1923



138



189





The Public Domain

Activity:

With your team members, visit:

<https://www.press.uchicago.edu/Misc/Chicago/copyright.html>

As you go over the document, try to answer the following questions:

- What works are in the public domain?
- For works published after 1978, how many years after the death of the author does it take the work become part of the public domain?
- What should you do if your text is published after 1925 and is not yet in the public domain?

- +
-
-

Textual Analysis with Voyant Tools

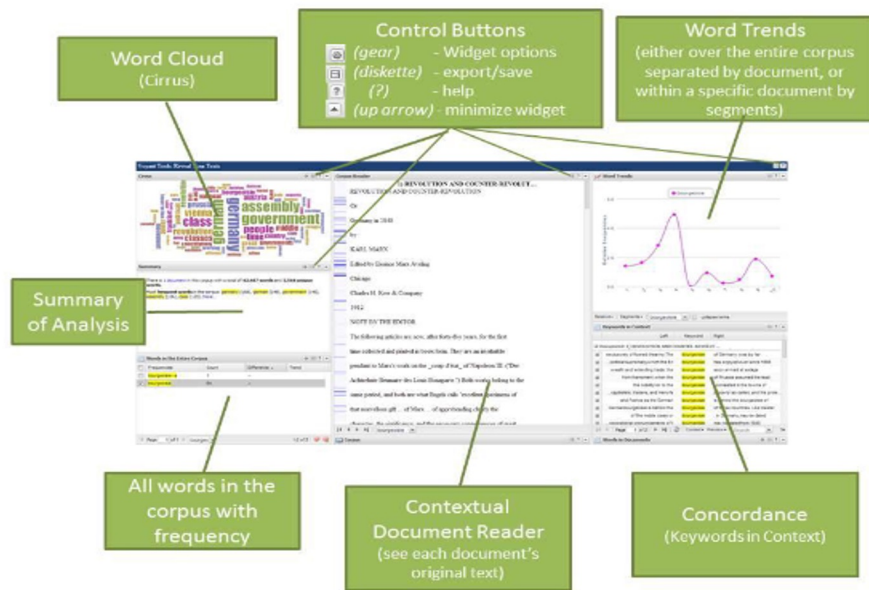
Voyant Tools is a free web-based reading and text analysis environment for large digital collections.

Stefan Sinclair and Geoffrey Rockwell of the University of Alberta developed the project

Support for Voyant came from the Canadian Institute for Research Computing in the Arts, the Social Science and Humanities Research Council, and the Canada Foundation for Innovation; (therefore, it was specifically designed for the digital humanities)

Voyant Tools

- VT offers digital text-mining tools, such as relative word frequency and distribution, collocation of words
- With a minimum of training and a small body of texts any student or researcher can use VT
- VT “computationally processes large amounts of text (**corpus/corpora**) to produce a series of customizable visualizations for text analysis”
- VT holds two internal corpora, the [Shakespeare's Plays](#) and [Jane Austen's Novels](#) and recently Mary Wollstonecraft Shelley's *Frankenstein* was added
- users can use any one of these author's texts or the entire corpus of their writing
- Example: Shakespeare
 - visualizations e.g. word clouds, bubblelines, graphs of word
 - word trends in single/multiple texts
 - context of specific words
 - Reading at different scales



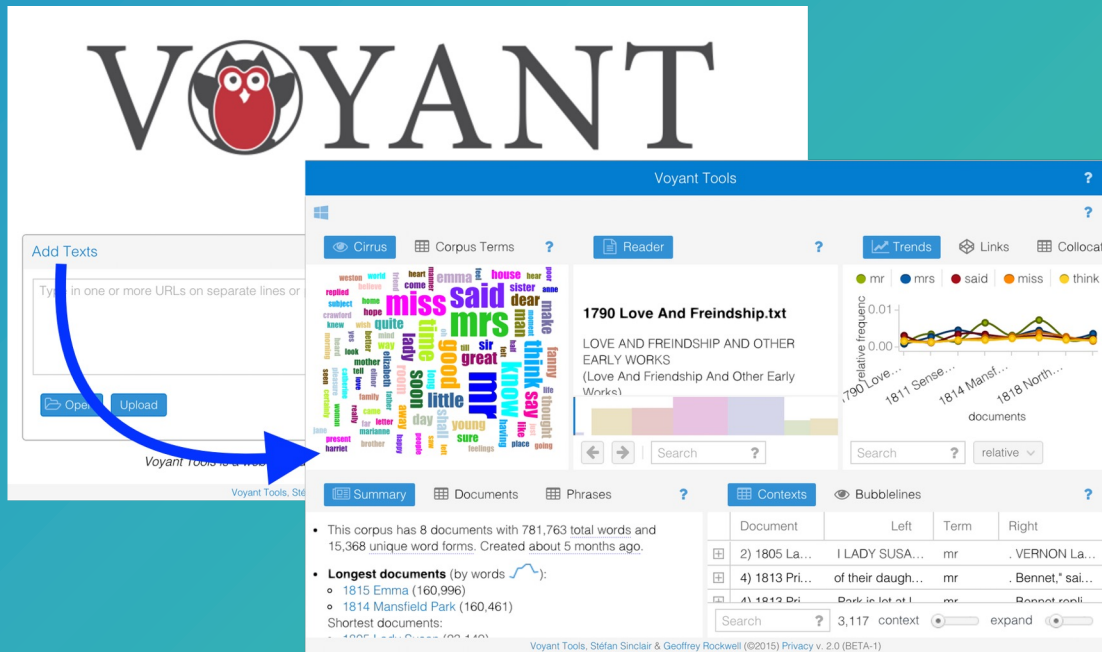
Voyant Tools Overview

Voyant Tools

Voyant with the Works of Jane Austen. You can also [use Voyant with your own corpus](#).

- Cirrus
- Reader
- Trends
- Summary
- Contexts

Voyant Tools



<https://voyant-tools.org/docs/#!/guide/start>

What do you need to make VT work?

- Electronic text(s) - “plain” text(s), unstructured or structured
 - Project Gutenberg (1971)
 - [HathiTrust](#) Digital Library
 - [Digital Public Library of America](#)
 - [Gale Digital Scholar Lab](#)
 - [Collections as Data](#)
- Internet connection, laptop with browser (not responsive)
- An [offline version of VT](#) can be downloaded on a local server

+

•

○

Textual analysis *Voyant Tools*

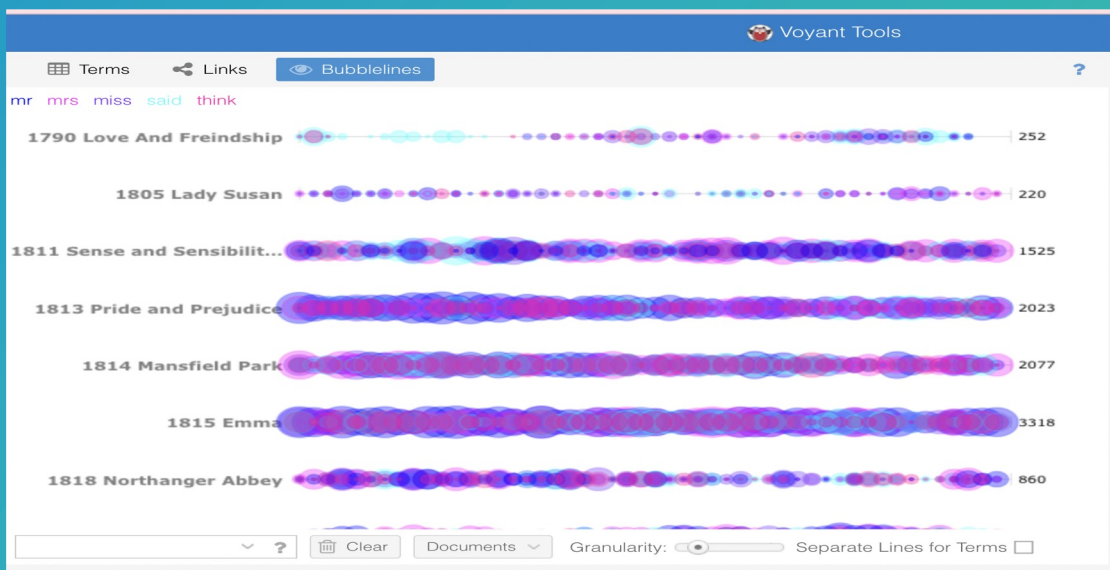
- VT is extremely easy to get started; however, it works best in highly specific conditions
- Like most text mining tools, it requires a large collection of data
- It is possible to input a single URL or few PDFs, but the results would not reveal many patterns, nor draw any large lexical comparisons
- VT does not work with tabbed data

+

•

○

Voyant Tools



Sinclair, S. & G. Rockwell. (2022). Bubblelines. *Voyant Tools*. Retrieved May 29, 2022, from <https://voyant-tools.org/?query=mr&query=mrs&query=said&query=miss&query=think&docId=344c515bb464b5a>

Researchers can:

- Create visualizations e.g. word clouds, bubblelines, graphs of word frequencies and embed them into their pages for future exploration
- Analyze vast amounts of textual data within and between individual texts
- See word trends in single/multiple texts
- Find the context of specific works
- Trace shifts of emphasis over time that were not previously possible
 - Where was a word/phrase said, when?
 - Did words/phrases change over time?
- Apply close and distant reading simultaneously through the Reader and visualizations tools

+

•

○

Questions to ask of Voyant Tools

- What are the most frequently used words across an entire text? (Cirrus)
- How are certain words used in different parts of texts? (Bubblelines) (Trends)
- What are the most frequently used words in their context? (Context)
- What are the most frequent collocations of words? (Phrases) (Links)
- If you build a corpus, you could
 - compare most frequently used words of politicians
 - compare most frequently used words of an author over time/genre
 - can you think of more?

+

•

○

Experiments with Voyant Tools

- The visualizations allowed for new questions to be investigated in ways that would have been much more difficult to do so fast and accurately
- I downloaded a plain text version of *Memoirs of a Midget from* to the Gutenberg online library. exported Cirrus and Bubblelines
- I found out that the word “eyes” occurs 291 times in *Memoirs of the Midget* and searched for it in other works. By visualizing the texts and tracing the context of each occurrence, things began to sound interesting to someone used to close reading practices
- I continued this line of thought and enquiry by searching for words that occur in the corpus of Jane Austen’s novels (already installed in VT).
- I found an interesting occurrence of the theme of “time” in her corpus and can begin to link it to the same theme which is predominant in de la Mare’s works

I’m sure other new possibilities will be offered by going back to de la Mare’s corpus and searching again for other words or terms

+

•

○

Experiments with Voyant Tools

Paper in progress: Punishment and entrapment in the works of Walter de la Mare

- Using the **contexts** tools and the Reader, I will see in which **context words associated with these** themes appear
- **where** and **when** each occurrence appears in each text, throughout his writing career, across genres, or in a corpus of the entirety of his works.
- At the press of a button, I can produce multiple visualizations based on quantitative approaches, alongside qualitative evidence, to support my claims (which was not possible before)
- More importantly, visualizations will help gear my research questions into new directions
- I will continue to use a close reading approach with the aid of a software that is quantitative and nested in distant reading methodologies.

+

•

○

Searching and reporting results Voyant Tools

Discussion:

How do you search and report findings?

If you don't get results from small amount of data,
keep testing tinkering until it takes you to something
you hadn't seen before

+

•

○

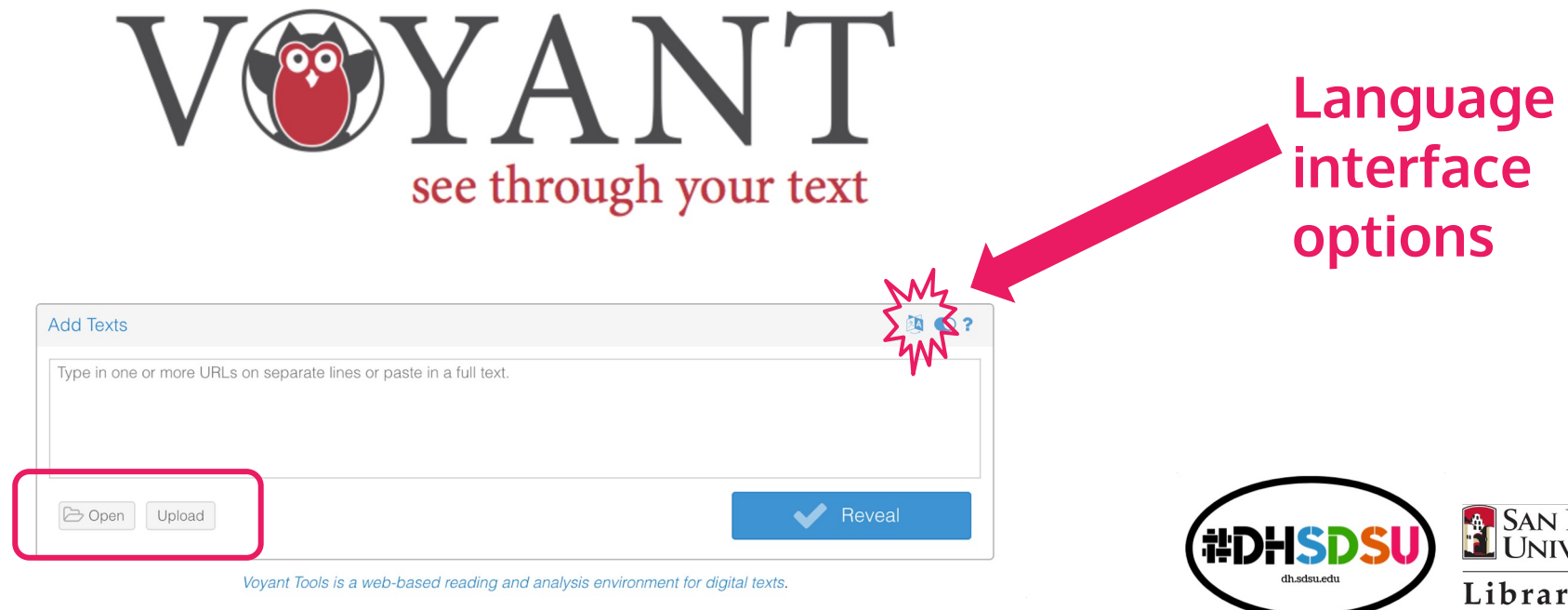
Hands-on Voyant Tools

<https://voyant-tools.org/docs/>

<https://voyant-tools.org/docs/#!/guide/start>

loading text into voyant

1. Paste URLs for online texts not behind a firewall
2. Open existing corpus (Shakespeare or Austen)
3. Upload .txt file(s) or zip files (preferred)



quick tour of initial views + tools ("skins")

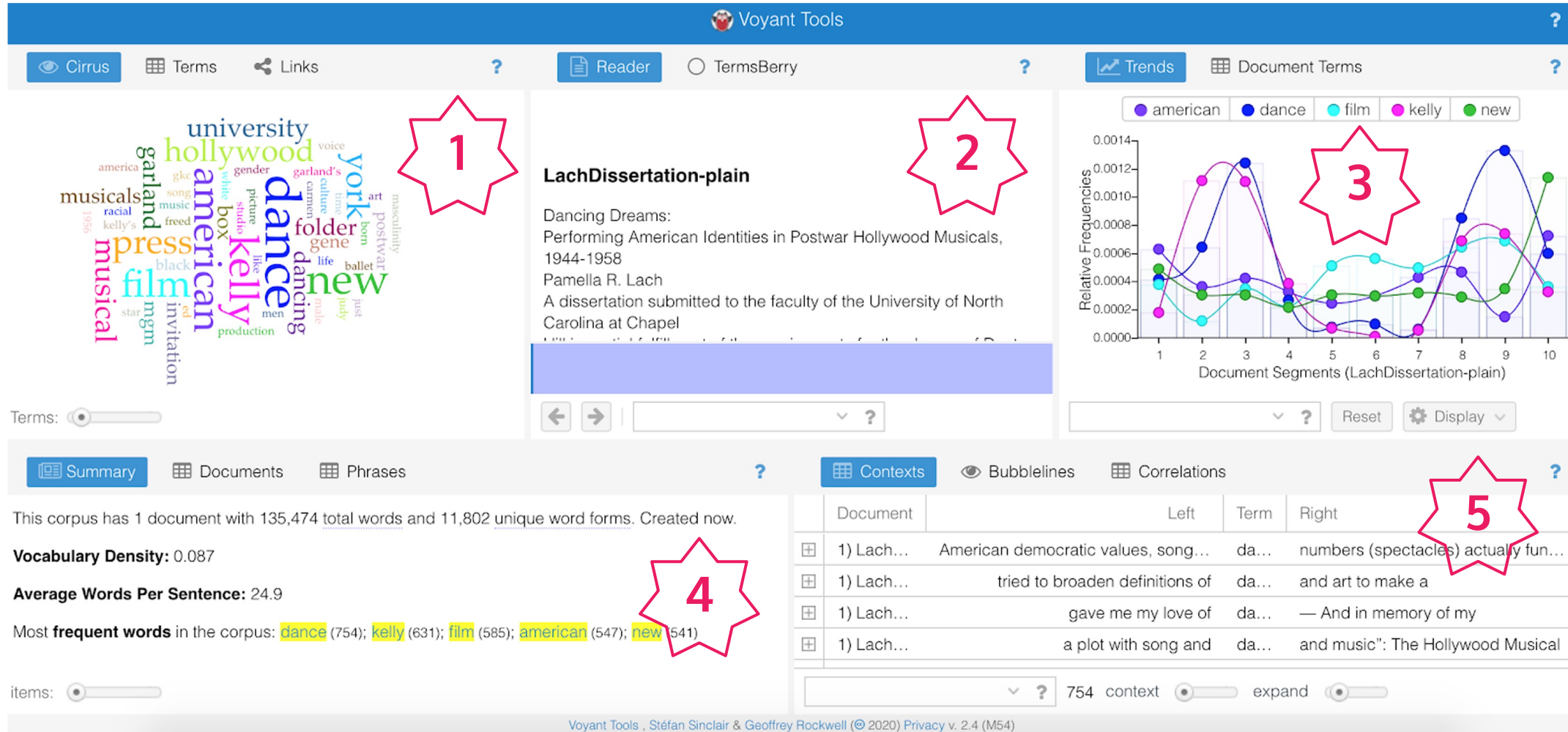
1. Cirrus
word cloud

2. Reader

3. Trends
(TF)

4. Summary

5. Contexts



customizing + refining results

Results can be customized/refined in a number of ways:

- Adjust # of terms/items you're seeing (except Reader)
- [Stopwords](#) to filter out noisy words
- [Stemming](#) + other wildcard queries to zoom in on specific terms/patterns
 - * to truncate or control for variance
 - ^ for exact match (if you only type a word, ^ is implied)
 - term01 | term02 for exact pairs

Learn more [here](#)



stopwords

- Click the options in the panel to access
- Edit list to add corpus-specific terms
- Apply globally doesn't always work :-)



Options

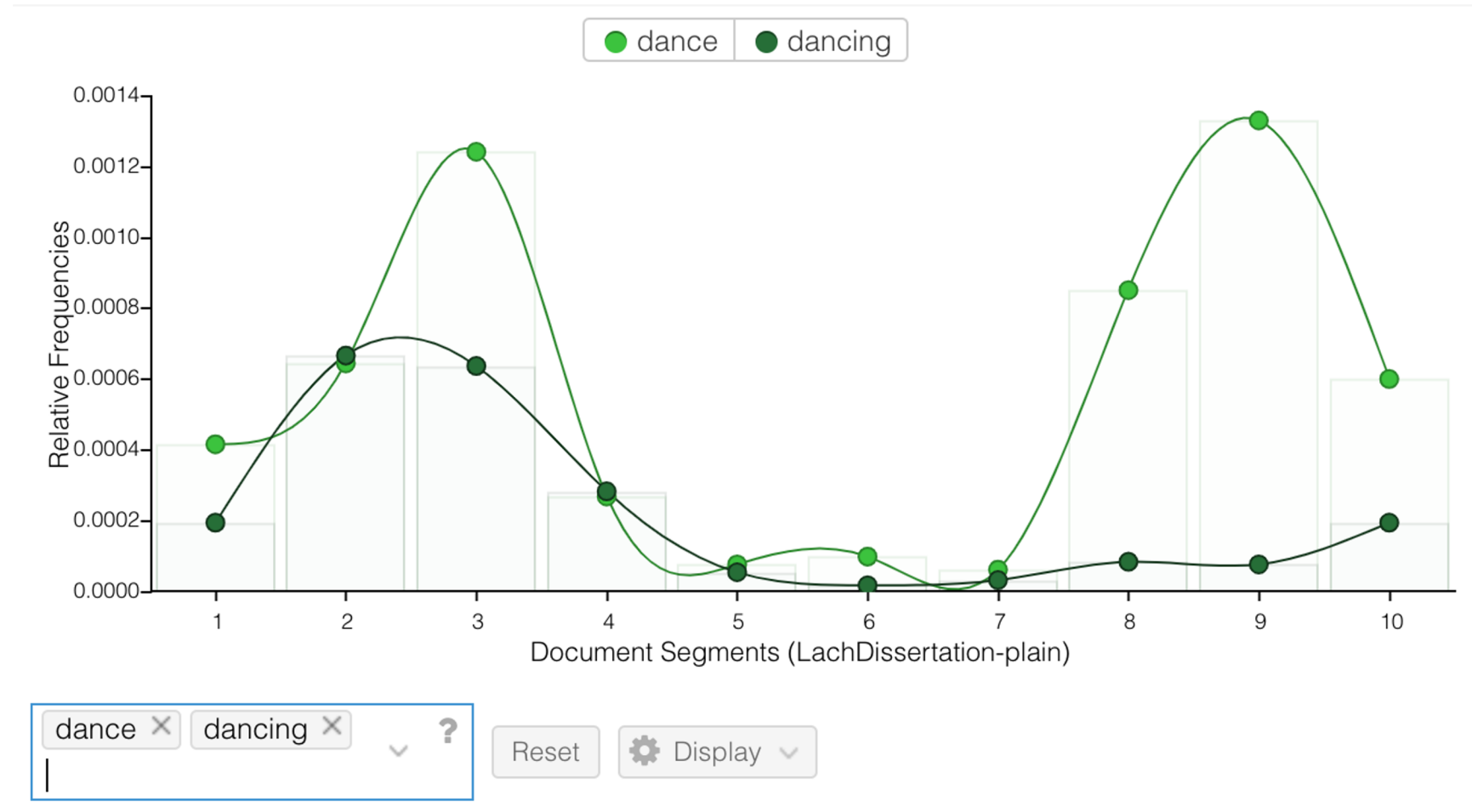
Stopwords:	Auto-detect	Edit List	<input checked="" type="checkbox"/> apply globally
White List:		Edit List	
Categories:	auto	Edit	
Font family:	Palatino		
Palette:	default	Edit List	

Reset Cancel Confirm

filtering for specific words

term
occurrences

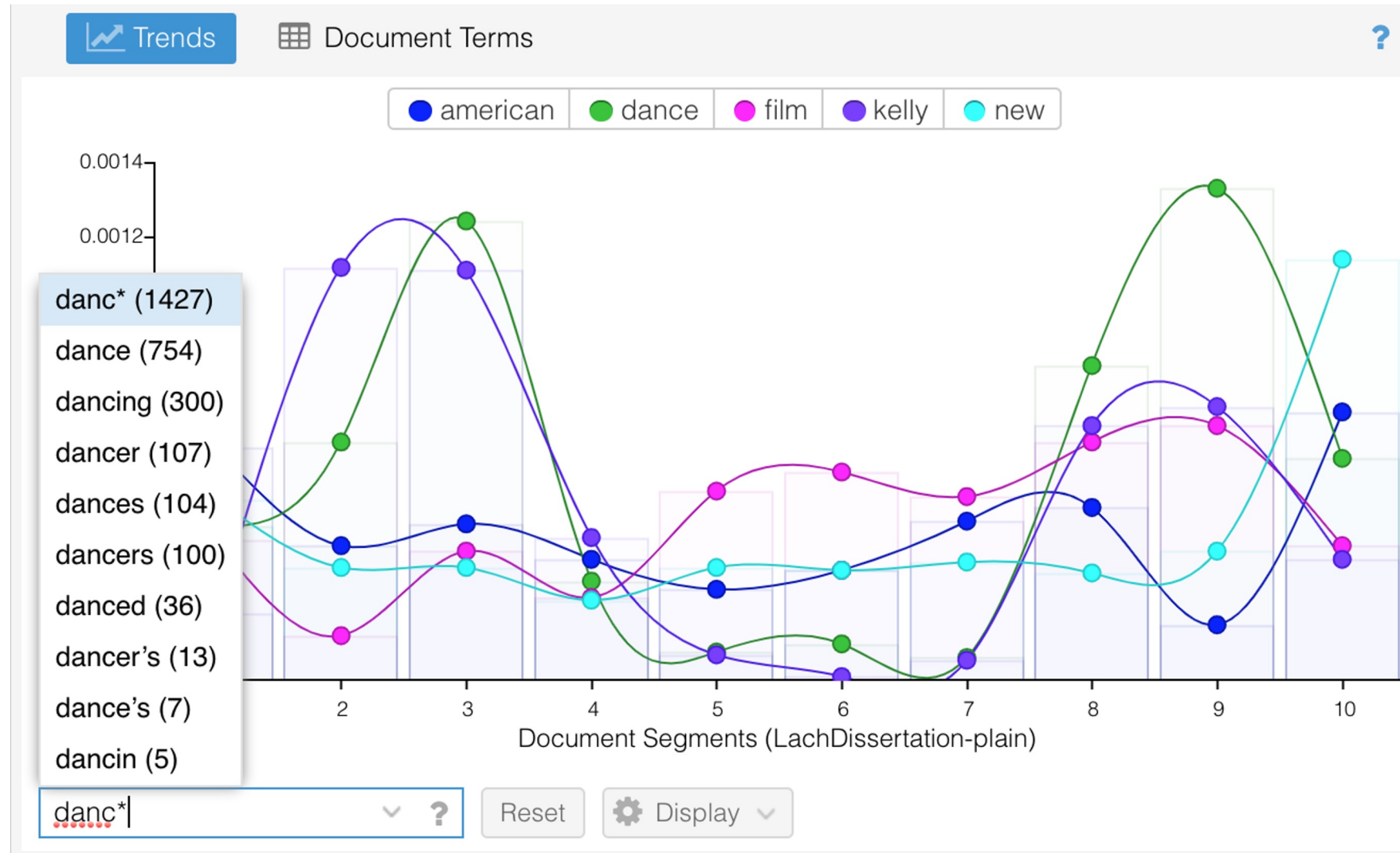
dance = 754
dancing = 300



stemming

danc* captures variations → more results = saves you from typing all the possibilities

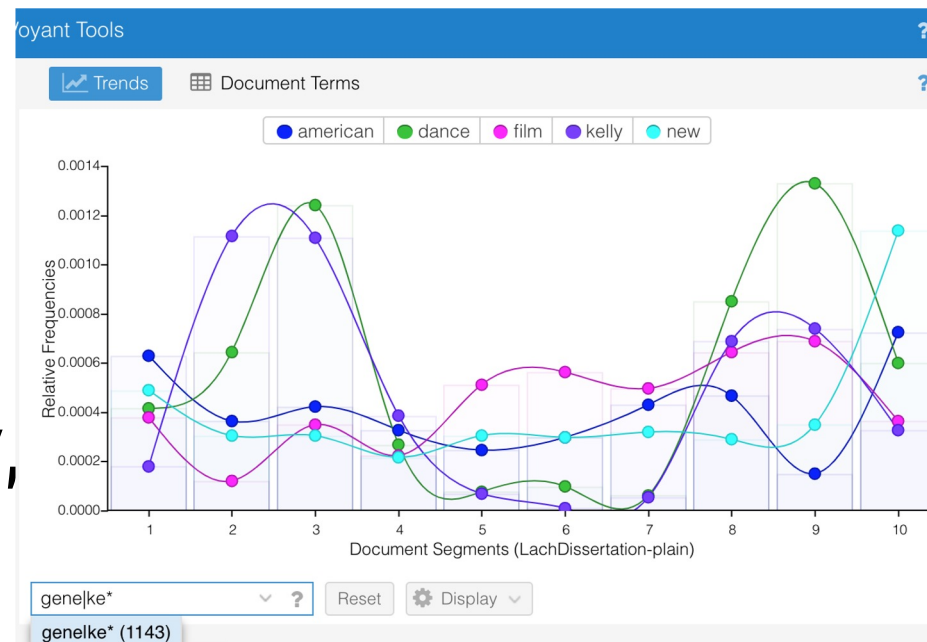
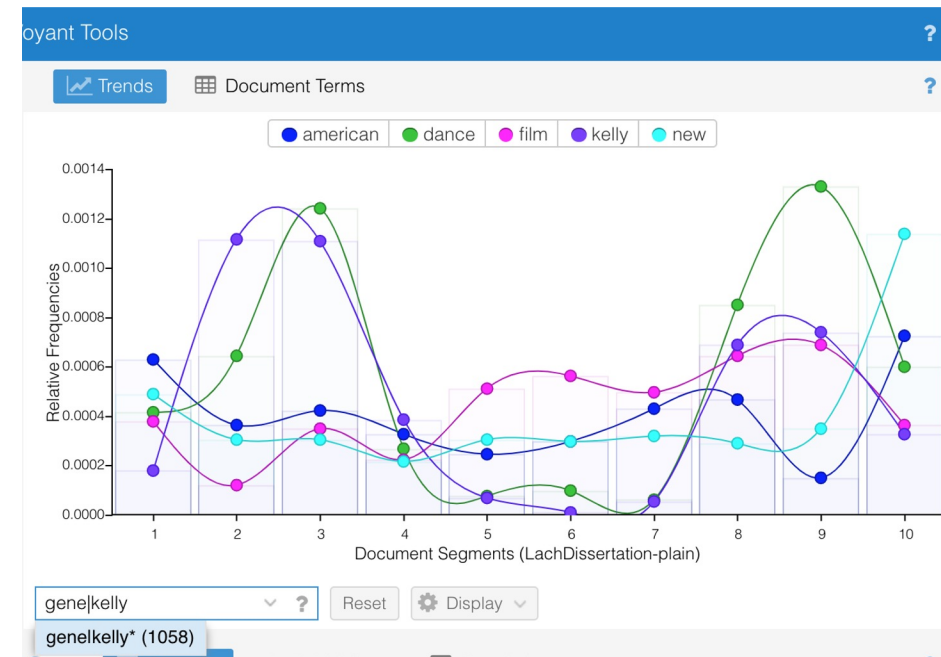
(not always what you want)



wildcard queries

Top: Using the | to match first and last names (Gene Kelly)

Bottom: Using the | plus stemming for variations (Gene Kelly, Gene Kelly's)

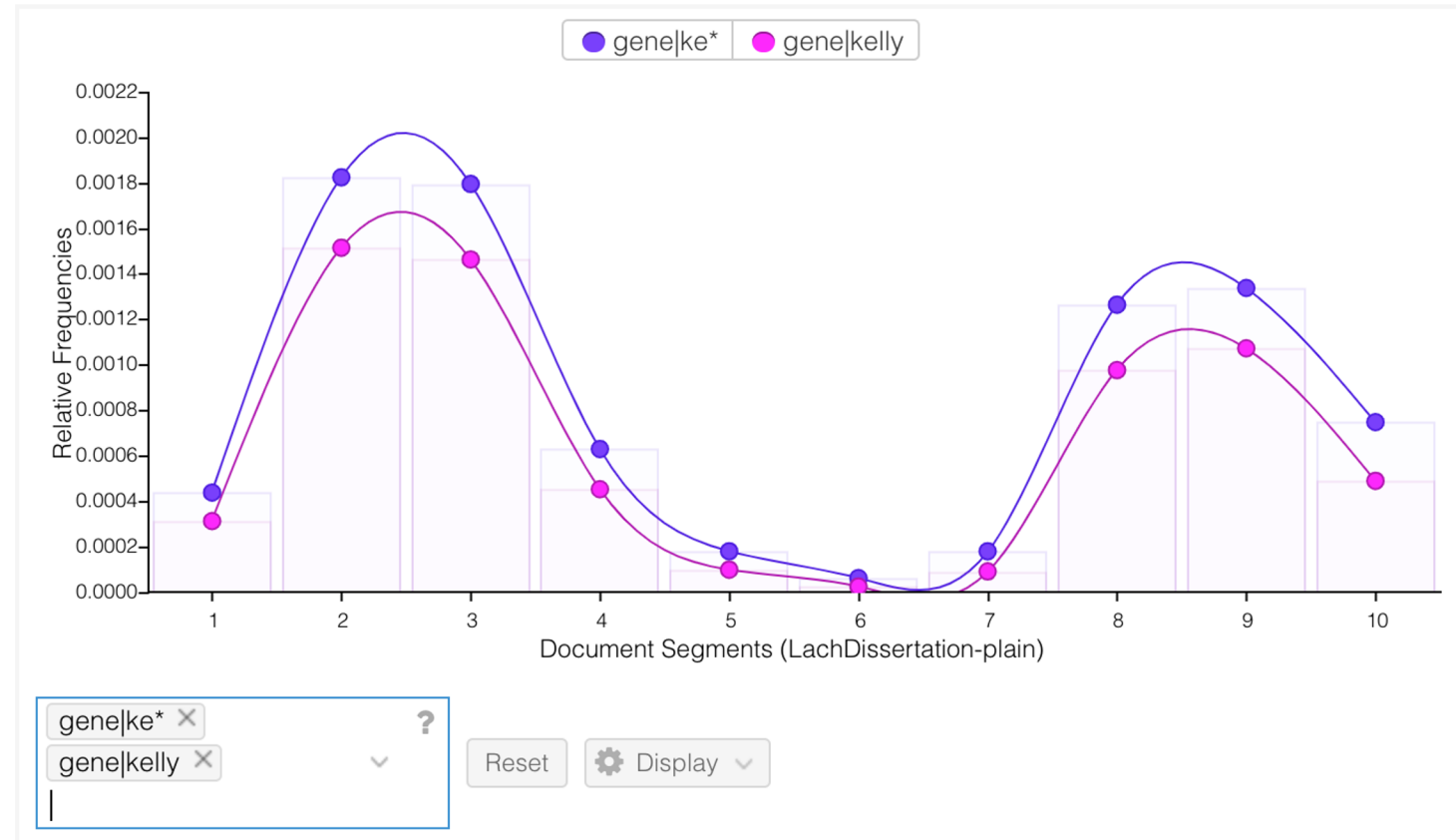


wildcard queries

Comparing results of
different wildcard
queries

gene|kelly = 1058

gene|ke* = 1143



changing the vis tools ("skins")

All of the panels/tools are resizable + interchangeable

- Tools ("skins") have baked-in options to switch tools
- But you can also change to other tools
- Options are categorized (but overlapping) - some are beta

- Corpus Tools
- Document Tools
- Visualization Tools

- Grid Tools
- Other Tools

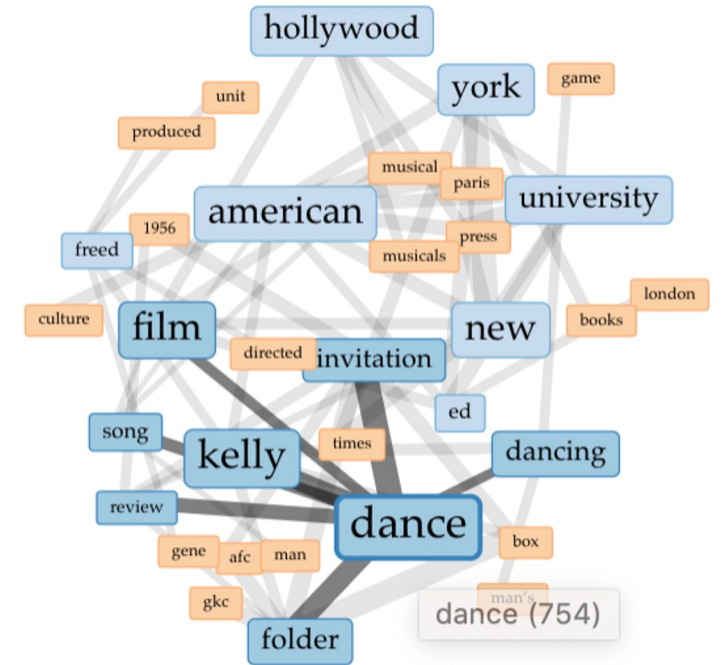


links

shows the
collocations
(links) between
words

Collocates graph shows a network graph of higher frequency terms that appear in proximity. Keywords are shown in blue and collocates (words in proximity) are showing in orange. Features include:

- hovering over keywords shows their frequency in the corpus
- hovering over collocates shows their frequency in proximity (not their total frequency)
- double-clicking on any word fetches more results
- a search box for queries (hover over the magnifying icon for help with the syntax)

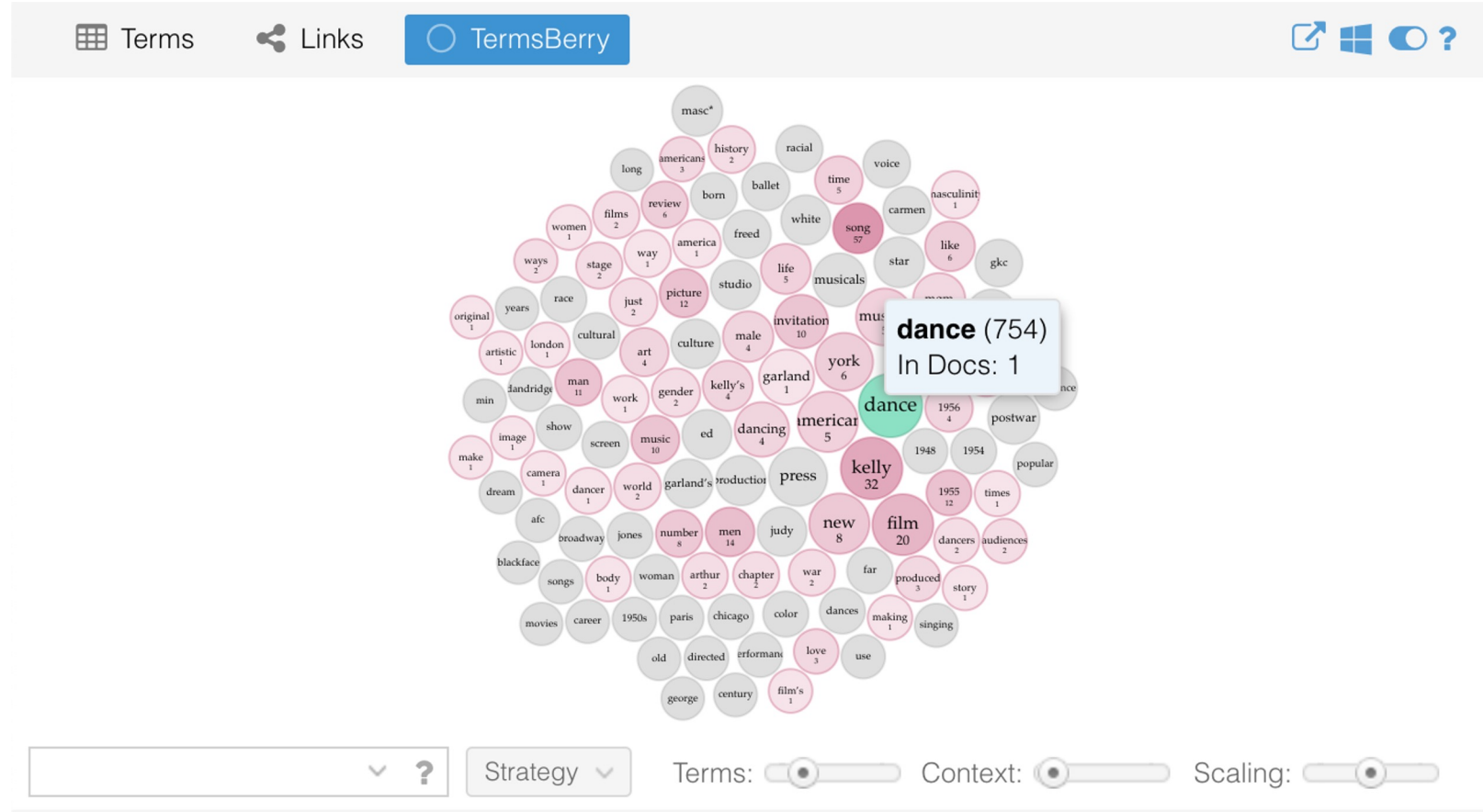


Context

terms berry

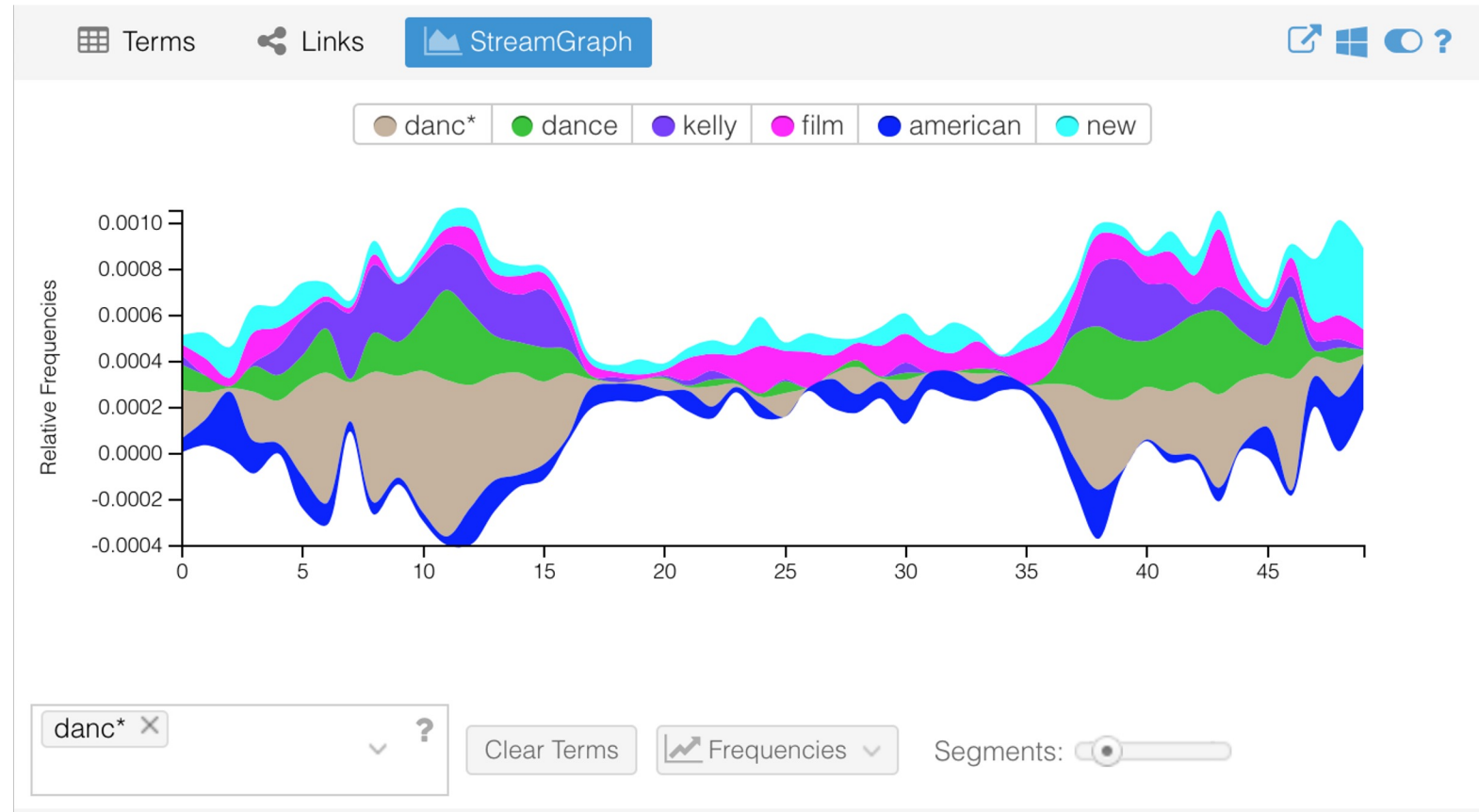
explore high
frequency terms
and their collocates
(words that occur in
proximity)

See also Tina Lumbis, "[A Creative Reading of Web 2.0 and Literary Criticism Using Voyant's TermsBerry](#)," *Post45: Web 2.0 and Literary Criticism*, 9 September 2019.



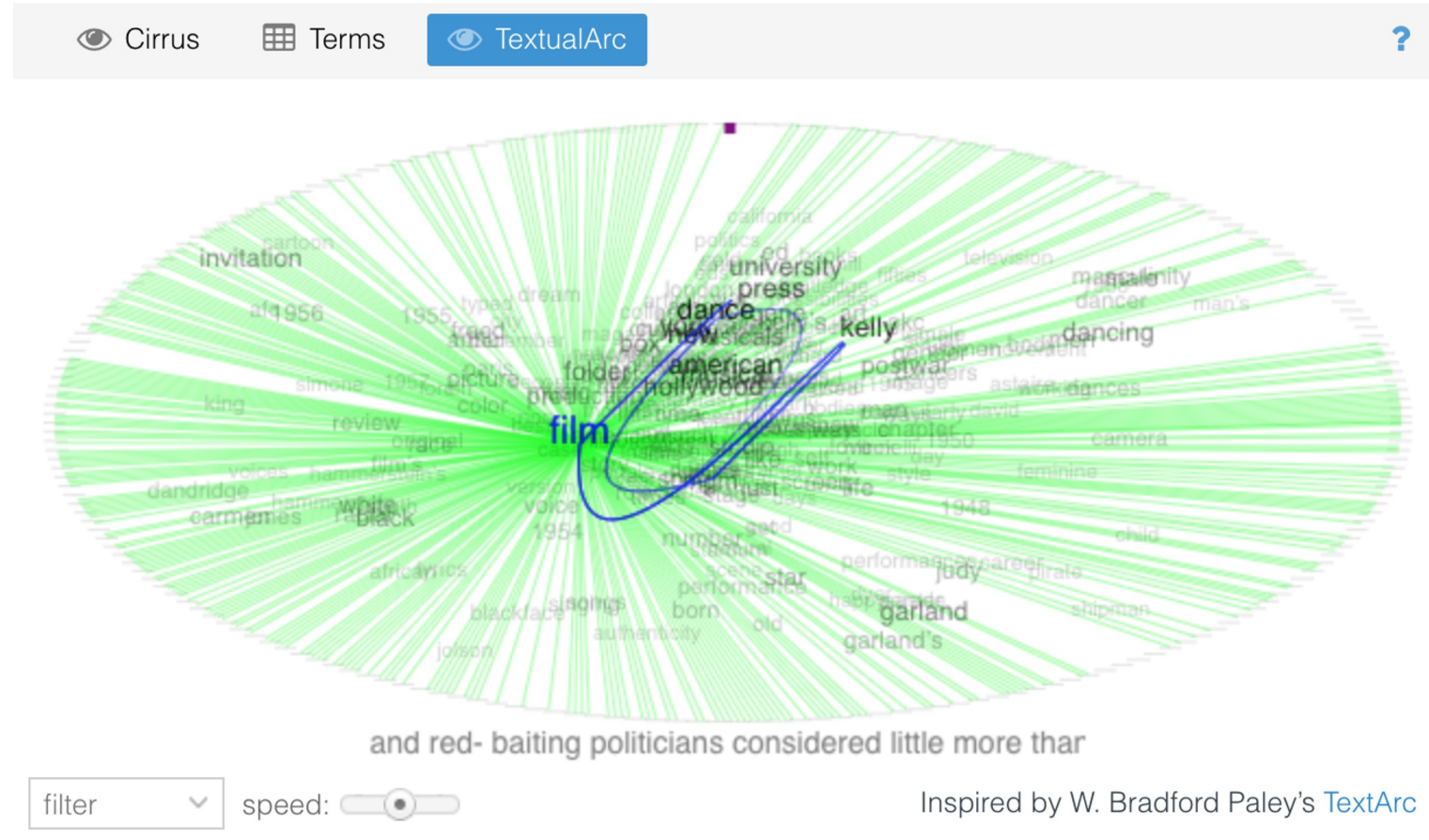
steam graph

depicts the change of word frequency across the corpus; typically used when the order of the documents in a corpus is meaningful



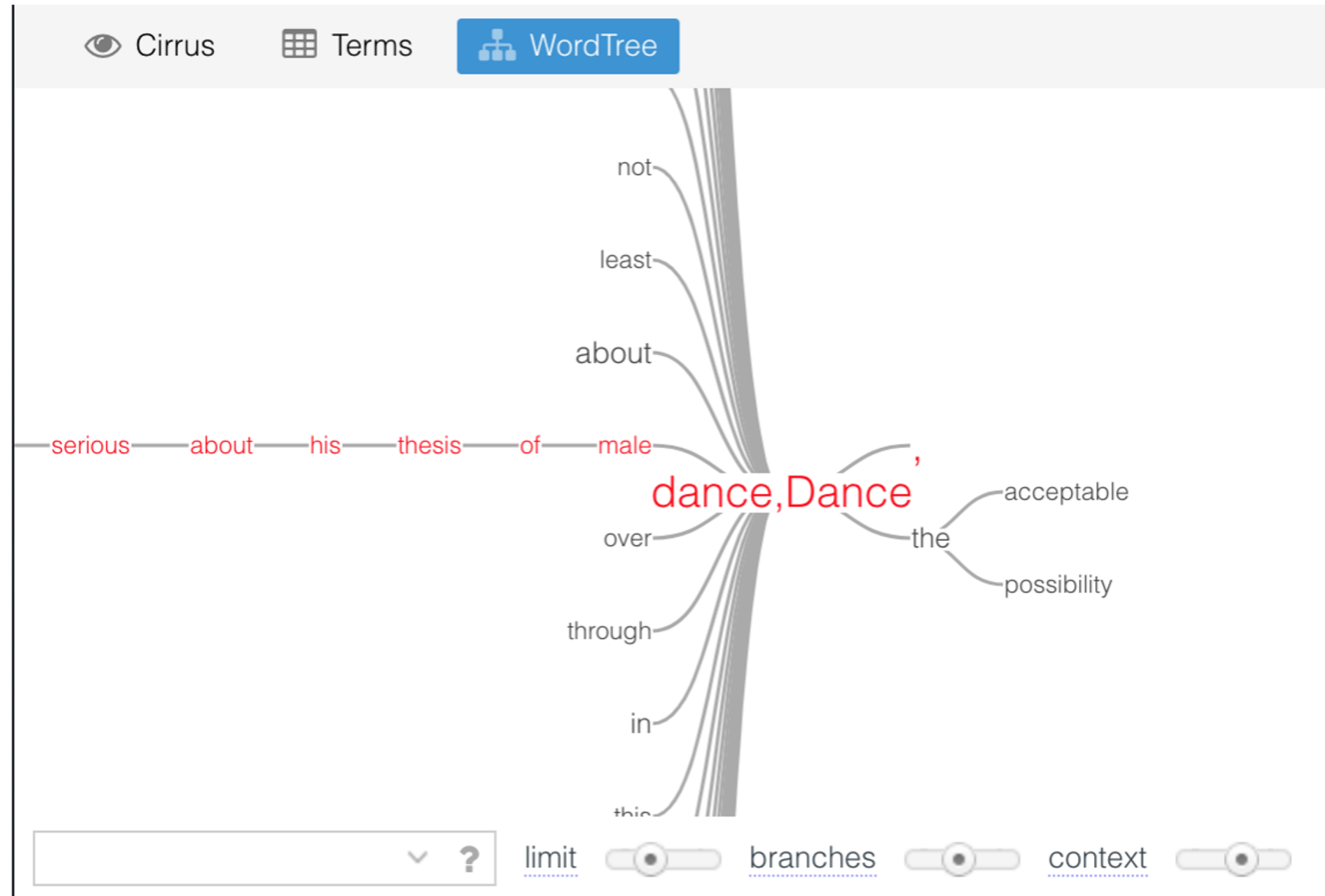
textual arc

dynamic (animated)
visualization of the
terms w/ weighted
centroid of terms and
an arc that follows the
terms in document
order



word tree

allows you to
explore how
keywords are
used in different
phrases in the
corpus



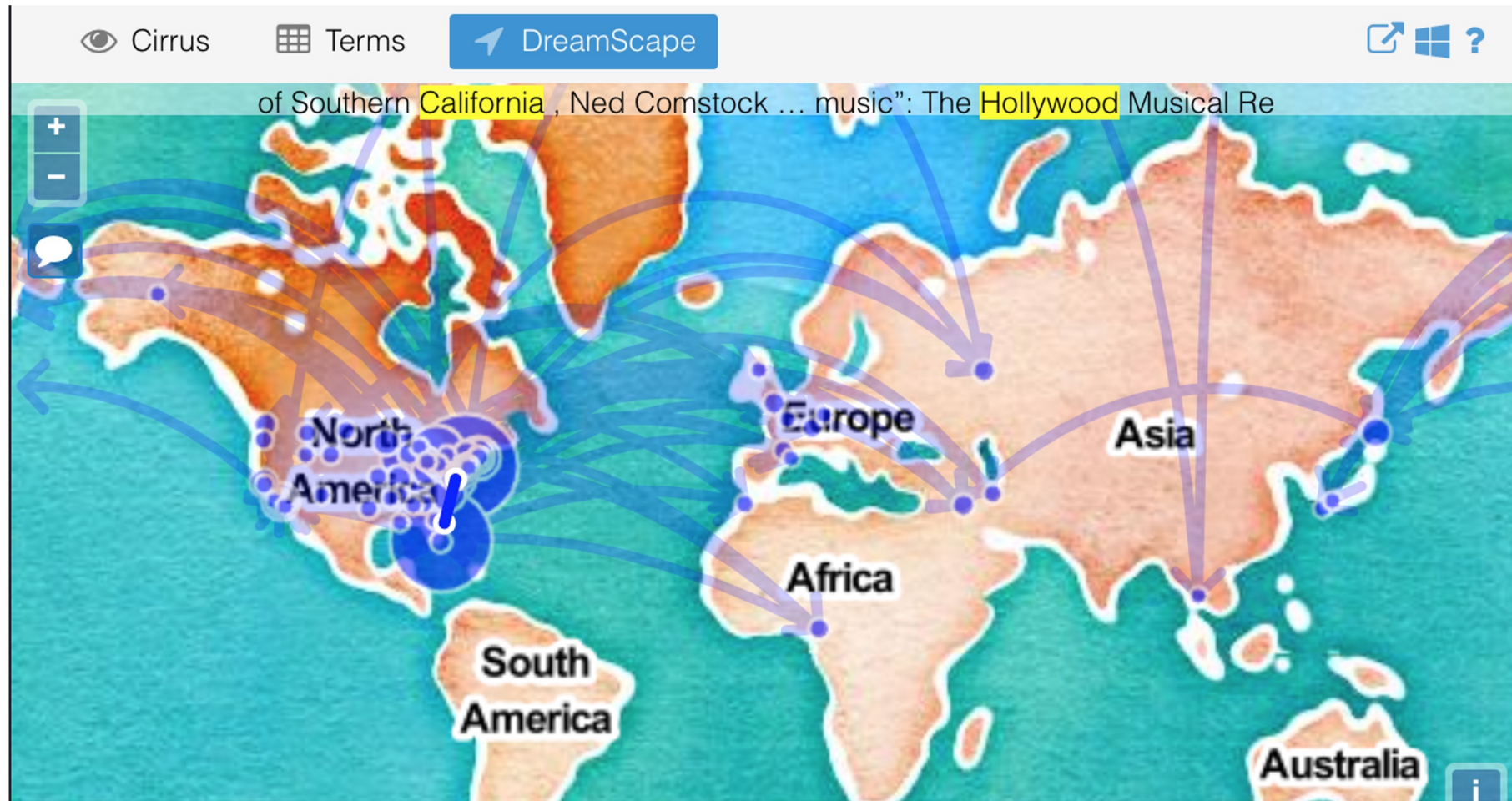
dreamscape (beta)

preliminary attempt to explore how texts might be represented geo-spatially. The tool tries to identify locations (especially city names) mentioned in texts, and suggests patterns of recurring connections between locations, patterns that might help identify travel of people, ideas, goods, or anything else

*my example is very noisy because it's pulling city names from my bibliography



dreamscape (beta)



corpus collocates (grid tool)

table view of
which terms
appear more
frequently in
proximity to
keywords
across the
entire corpus

	Cirrus	Terms	Collocates	?
	Term	Collocate	Count (context)	
<input type="checkbox"/>	dance	invitation	265	
<input type="checkbox"/>	new	press	88	
<input type="checkbox"/>	press	new	88	
<input type="checkbox"/>	dance	kelly	81	
<input type="checkbox"/>	kelly	dance	78	
<input type="checkbox"/>	american	paris	77	
<input type="checkbox"/>	hollywood	musical	74	
<input type="checkbox"/>	new	times	66	
<input type="checkbox"/>	american	culture	64	
<input type="checkbox"/>	dance	review	59	
<input type="checkbox"/>	american	musical	58	
				10,715 context

	Cirrus	Terms	Collocates	?
	Term	Collocate	Count (context)	
<input type="checkbox"/>	danc*	invitation	270	
<input type="checkbox"/>	danc*	kelly	152	
<input type="checkbox"/>	danc*	dance	96	
<input type="checkbox"/>	danc*	gene	74	
<input type="checkbox"/>	danc*	song	60	
<input type="checkbox"/>	danc*	review	60	
<input type="checkbox"/>	danc*	film	50	
<input type="checkbox"/>	danc*	men	46	
<input type="checkbox"/>	danc*	dancing	44	
<input type="checkbox"/>	danc*	new	42	
<input type="checkbox"/>	danc*	man's	41	
				2,432 context



how to pick a vis tool ("skin")?

1. Figure out what you're trying to do and why
 - a. Does context matter?
 - b. Collocation? Relationship between terms?
2. Determine the type of tool (corpus, document, visualization, grid)
 - a. Do you need to see something visualized?
3. Read about each tool and its intended uses to determine if the use fits your need: voyant-tools.org/docs/#!/guide/skins

exporting

- You can export a URL, an [embeddable](#) tool (interact data, or a bibliographic reference
- Applies to the entire Voyant project or a single particular tool ("skin")
- You can also export a static .png of visualizations (a screenshot may have better image quality)





Reflections on Voyant Tools

- ❖ What can you do with Voyant Tools?
- ❖ What can't you do?
- ❖ When and where will you be using it?

+

•

○

Part 3



Preparing a corpus for Antconc

Preparing your corpus for analysis

- How to save files and prepare a corpus relies on research goals
- Save files as plain texts in ways
 - With author's surname _ short title of work _ pub year (AntConc): Dickens_Taleoftwocities_1859
 - With pub year _ author's surname _ short title (Voyant): 1859_Dickens_Taleoftwocities
- Save files in two folders/corpora to compare perspectives:
 - Corpus/Folder A: texts representing one perspective
 - Corpus/Folder B: texts representing the opposite perspective
- To compare keywords:
 - Corpus/Folder C: texts for analysis of special keywords
 - A Reference Corpus/Folder D: texts for comparison
- More options appear as you revise your research questions and develop your analysis

+

•

○

Preparing a corpus for Antconc

- **So, how do I compile my corpus?**
- Today, numerous texts have been digitized and are available freely on numerous portals:
- Project Gutenberg: <https://www.gutenberg.org>
- Internet Archives: <https://archive.org>
- Open Library: <https://openlibrary.org>
- Many Books: <https://manybooks.net>
- BNF Gallica: <https://gallica.bnf.fr/>
- [GitHub](#)
- Consult a **library guide** like the [Library of Congress: Finding E-books: A Guide](#).
- Your Librarians could also suggest more places to find available datasets and corpora



Preparing a corpus for Antconc

- **Start small and work up to larger data sets**
- **"Text analysis projects are iterative.** You should always start small and work up to larger and larger data sets. The data set you curate we will call a **corpus**. It could be one or more texts unified by some criterion relevant to your research. You may treat one text as many or many as one. You may slice it in different ways. The corpus selected, captured, and generated will constrain and enable questions so you should have some idea of what you want to study when you start gathering a corpus. As you ask questions you will realize you need to add or remove from the corpus. It will change as your thinking does. You are in a conversation with text."

[Geoffrey Rockwell and Stéfan Sinclair. Dialogica (<http://dialogi.ca>)]

+

•

○

Preparing your corpus for Antconc

What do you want to study?

- A process for developing a corpus could take this form:
- Identify **what it is you want to study**. For example, you may want to study Mary Shelly's novel *Frankenstein; or the Modern Prometheus*. [
- Try to frame some **questions** and develop **conjectures** about that which you want to study.
- Identify what **evidence** would allow you to answer the questions or test the conjectures. For example, would the full text of the novel help with the questions you have or are your questions about the reception of the text?
- **Gather** the evidence in whatever form. It could be a novel or a set of web pages on a subject. Make sure you keep track of where you got the evidence as you gather it. *Dialogica*

+

•

○

Preparing a corpus for Antconc

- **Convert it to data (datafy)** that can be analyzed. This may involve entering data or digitizing a text if you can't find an electronic copy already digitized.
- **Clean** and **enrich** the data if needed. Depending on where you got the data you may need to clean it up and you may want to add information. For example, texts from the Gutenberg Project often have long notices from Gutenberg at the beginning and end that need to be removed before analysis / [boilerplate language] *(my addition)*
- **Document** where you got the evidence and what you did to it so you can report accurately later. Likewise, you should be prepared to document the various results you will produce.
- Try to identify a novel or single text that you would like to study more closely. **Choose something you know. Save it as a text file on your computer and clean the text if necessary.** *Dialogica*

+

•

○

Computer-Assisted Textual analysis with Antconc

- Enter into a dialogue with the texts
- One way to think about how you can study a text is to engage in a conversation with the text through the computer-assisted textual analysis applications.
- Think about questions you might ask about the text like:
- What is this about? What words would I expect to see as describing the text?
- What does this text say about something that matters to me like “friendship”?
- What words in texts make sense, and what words are anomalous?
- How does the language of the text change over the span of the text/corpus? Are some words more important in the beginning and some at the end? Could there be framing words?
- Text analysis tools like Antconc and Voyant grew out a tradition of developing concordances for important texts like the bible or Shakespeare’s plays. Concordances were printed tools like indexes that allowed a preacher or scholar look up a word like “friendship” and see all the instances across the bible in one place. She could then think through friendship and prepare a sermon or paper discussing the theme. With Antconc and Voyant you can now study any electronic text in a similar fashion. (adapted from *Dialogica*)

+

•

○

Computer-Assisted Textual analysis with Antconc

- Identifying themes
- Identify words that interest you. See if there are sets of words that might form a theme. Ask yourself why certain words stand out.
- Try thinking in terms of themes that you would want to explore in a text much as you might using an index to a book or a paper concordance, if you had one. You can explore the high frequency words.
Dialogica

+

•

○

Computer-Assisted Textual analysis with Antconc

Search for words that interest you and follow themes/trends

- For example the word “life” might be interesting in *Frankenstein* as the novel is about giving life. What other words might you look at related to “life”? How about “death”?
- Now explore a set of words that might make up a theme. Ask yourself if the theme tells you something interesting about the text.
- Once you know how to explore one word, then you can start exploring related words. If you explore “life” you might also explore “death”. If you explore “father” you might look for other words having to do with family relations like “mother”, “brother”, and “sister.”
- So, before your search take some time to think about words that are related to your project’s goals

+

•

○

Computer-Assisted Textual analysis with Antconc

- Reporting on a Theme
- Once you get interesting results it is important to find a way to save the results for any paper or report that you want to write
- Export/screen shot example quotes that show how the word you are interested in is used.
- Learn how to use the Export/Cloning features in the applications to get examples in support of your results and analyses in your paper. *(Adopted from Dialogica)*



Computer-Assisted Textual analysis with Antconc

- Exporting quotes and trend graphs
- It is always a good idea to put a couple of quotes from the text being studied that exemplify what you see as a pattern across the text. The easiest way to do that is to use the **Reader** in **VT** or the **Concordance** in **Antconc** to simply **select** and **copy** the text and then **paste** into your Word Processor or wherever you are keeping notes. One disadvantage of some of these applications such as VT is that it will not give you the page number for the quote unless that is in the full text uploaded and even then you will need to scroll up to find the page.
- Export trend graphs and create an Infographic Summary of your Texts
(Adopted from *Dialogica*)

+

•

○

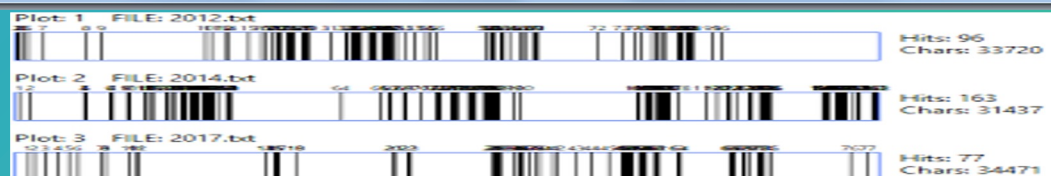
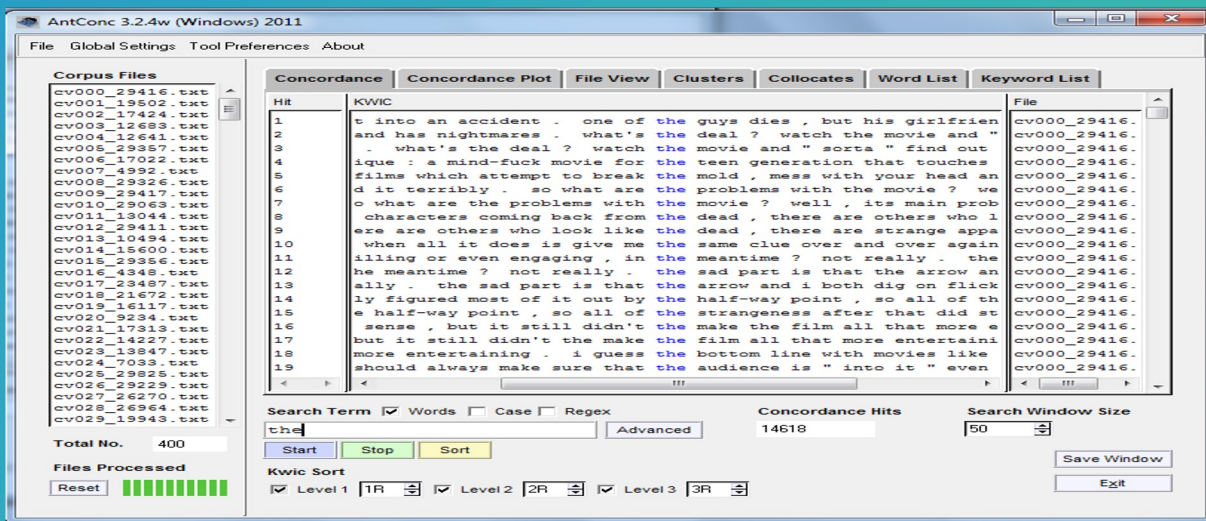
Antconc

(Lawrence Anthony)



- a general-purpose corpus analysis tool kit.
 - used by corpus linguists, translators, and English teachers.
 - a “freeware corpus analysis toolkit for concordancing and text analysis.”
 - Plenty of tutorials
 - Antconc takes plain texts as well as tagged and annotated texts
-
- Corpora for analysis:
<https://drive.google.com/drive/folders/12FBn-Pnneu6ssd5lvkgcaRoyzgdXqABS?usp=sharing>
 - A longer presentation on Antconc:
https://docs.google.com/presentation/d/1G4n5Qpew4RtLeYVAdPZdNijpXysThj2_/edit#slide=id.p23

Antconc



- Concordance: is where you search for a word and see how it's used in context
- Concordance plot: is where you see a concordance search results plotted in a 'barcode' format, with the length of the text normalized to the width of the bar and each hit shown as a vertical line within the bar. This allows you to see the position where search results appear in target texts
- File view: is where you see your files
- Clusters/N-grams: This allows you to search for a word or pattern and group (cluster) the results together with the words immediately to the left or right of the search term. The clusters can be ordered by frequency, the start or end of the word, the range of the cluster (number of files in which the cluster appears), or the probability of the first word in the cluster preceding the remaining words.
<https://antconc-manual.readthedocs.io/en/latest/clusters.html>
- Collocates: This tool allows you to search for collocates of a search term. The collocates can be ordered either by total frequency, frequency on the left or right of the search term, or the start or end of the word. They can also be ordered by the value of a statistical measure between the search term and the collocate.
[\[https://antconc-manual.readthedocs.io/en/latest/collocates.html\]](https://antconc-manual.readthedocs.io/en/latest/collocates.html)
- Word List: is where you see your wordlists. This tool counts all the words in the corpus and presents them in an ordered list. This allows you to quickly find which words are the most frequent in a corpus.
- Keyword List: This tool shows the which words are unusually frequent (or infrequent) in the corpus in comparison with the words in a reference corpus. This allows you to identify characteristic words in the corpus
- KWIC: KeyWord In Context

+

•

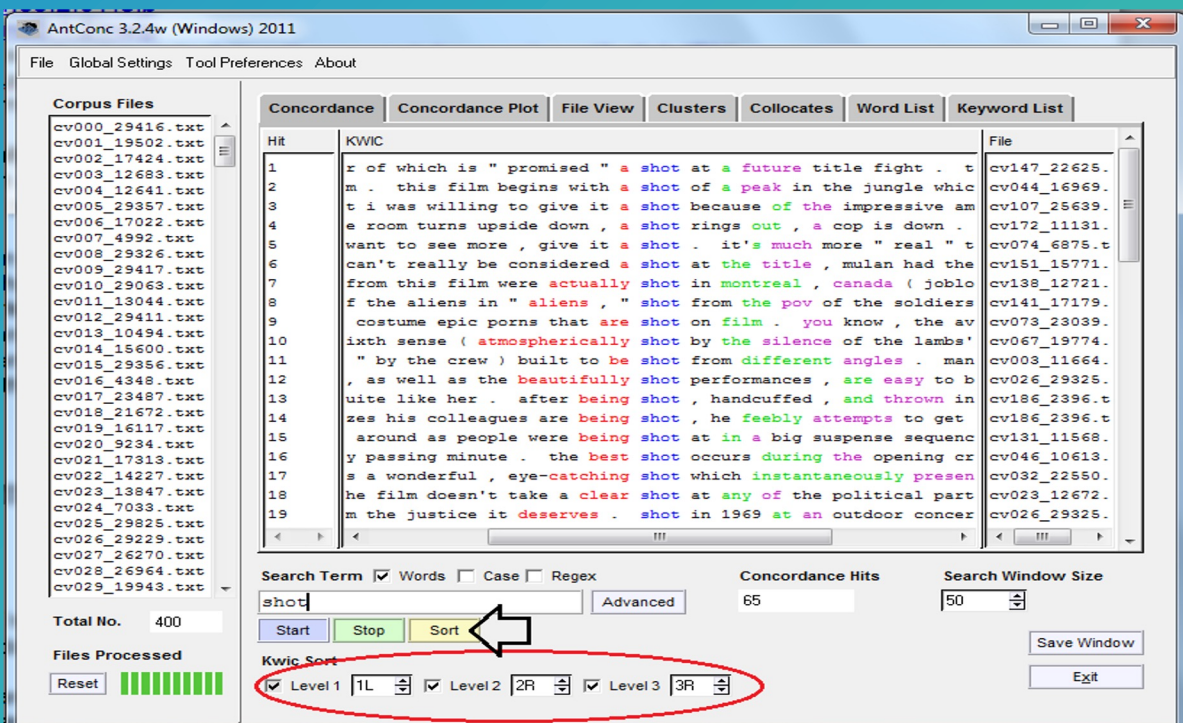
○

Textual Analysis *with Antconc*

- What do we look for in AntConc?
- We look for word frequencies to determine topics and themes
- We look for verbs, adverbs and adjectives
- We look for negative or positive words before and after search words
- We look for patterns, words appearing before and after recurring words
- We look for clusters and collocates
- We compare keywords and word lists

Textual Analysis with Antconc

- If you want to see patterns, you will need to examine words to the **L** or to the **R** of your search word
- Go to **KWIC sort** at the bottom. There are three levels of sort **1 to R**, **2 to R & 3 to R**
- Click on that and you will see a new pattern appearing in the **Concordance**
- Each word to the **R** has a color and is arranged in the alphabetical order of the first **R** word and then to the **2 R** and then to the **3 R**. You can **sort** to the **L**
- If you want to see where your search word appears in the corpus, hover your cursor on the word and click on it. It will take you to the **file** and show you where it appears exactly and enable you to see more context
- Type different forms of the word and see how they occur in the corpus: love, loving, loved. You can also search for phrases: loved by, to love and see how they are used in the corpus



Textual Analysis with *Antconc*

- A Wildcard such as * = **zero** or more characters following the previous letter
- love * = love, loves, loved, lovers, loving, lovely
- You can sort these results on **zero** word, center word, the **1 R** and **2 R**
- Other **Wildcards** can be seen in the **Global Settings** tab at the top to the right of **File**
- See tutorial on **The Programming Historian:**
<https://programminghistorian.org/en/lessons/corpus-analysis-with-antconc>

Textual Analysis with *Antconc*

How do we interpret and write up our results with AntConc?

- You need to test your findings before you interpret your results by clicking on the **Concordance** and closely reading/checking your texts/files
- You also need to compare results, using the **Clone Results** button. This will allow you to examine all your results in different windows & mark differences
- Some tools results can't be cloned (**e.g., Concordance Plot**), so you'll need to either create a screenshot of your visualization and crop it through an image editor, or use **Windows Snippet tools** see *slide 20*
- You also need to support your results with visualizations exported from the tools

+

•

○

Textual Analysis with *Antconc*

How do we interpret and write up our results with AntConc?

- In addition, you need to write the processes you went through starting with:
- your initial hypotheses and research questions
- where you located your data
- how you prepared it & uploaded it on AntConc
- which tools you used in AntConc
- how you reached some analyses and answered your initial research questions
- and, if applicable, how you revisited and developed new research questions during the research process

Hands-on with Antconc: Loading a corpus for analysis



Download Antconc from: <https://www.laurenceanthony.net/software/antconc>

Go to the **File** tab and click on the corpus you saved for analysis

If you do not have a corpus, you can use the [Brown Corpus](#) untagged corpus version (1 million current American English words) for today. It's in this [folder](#).

Sub corpus: Brown_A.txt, Brown_B.txt, & Brown_C.txt, have Journalism English words

Make sure you uploaded them correctly by clicking on the files separately after you go to **File View** tool and checking that the words are normal looking with no strange characters

You usually start by making a **Word List** of the Corpus

So, click on the **Word List** tool, then click on **Start** in the lower left corner of the right blank area

You will get a list of words starting with the most frequent words at the top and their frequencies to their left

Hands-on with AntConc: Concordance

The Concordance Tool allows you to search for words or phrases you're interested in.

It will show you the **patterns** they appear in

Type the word you're interested in, in the search area at the bottom of the screen: **love**

Then click on **Start** or **return**

AntConc will now go through all the corpus and show you all the hits on the word **love** you just searched

It will also show you words to the right and to the left of the word that you just searched for

You can increase the number of words appearing to the **L** and **R** by clicking on **Search Window Size**

Choose **50** or **100** words. This way you will have more **context**

Hands-on with AntConc: Concordance Basic Features

If you want to see patterns, you will need to examine words to the **L** or to the **R** of your search word

Go to **KWIC sort** at the bottom. There are three levels of sort **1 to R**, **2 to R & 3 to R**

Click on that and you will see a new pattern appearing in the **Concordance**

Each word to the **R** has a color and is arranged in the alphabetical order of the first **R** word and then to the **2 R** and then to the **3 R**. You can **sort** to the **L**

If you want to see where your search word appears in the corpus, hover your cursor on the word and click on it. It will take you to the **file** and show you where it appears exactly and enable you to see more context

Type different forms of the word and see how they occur in the corpus: [love](#), [loving](#), [loved](#). You can also search for phrases: [loved by](#), [to love](#) and see how they are used in the corpus

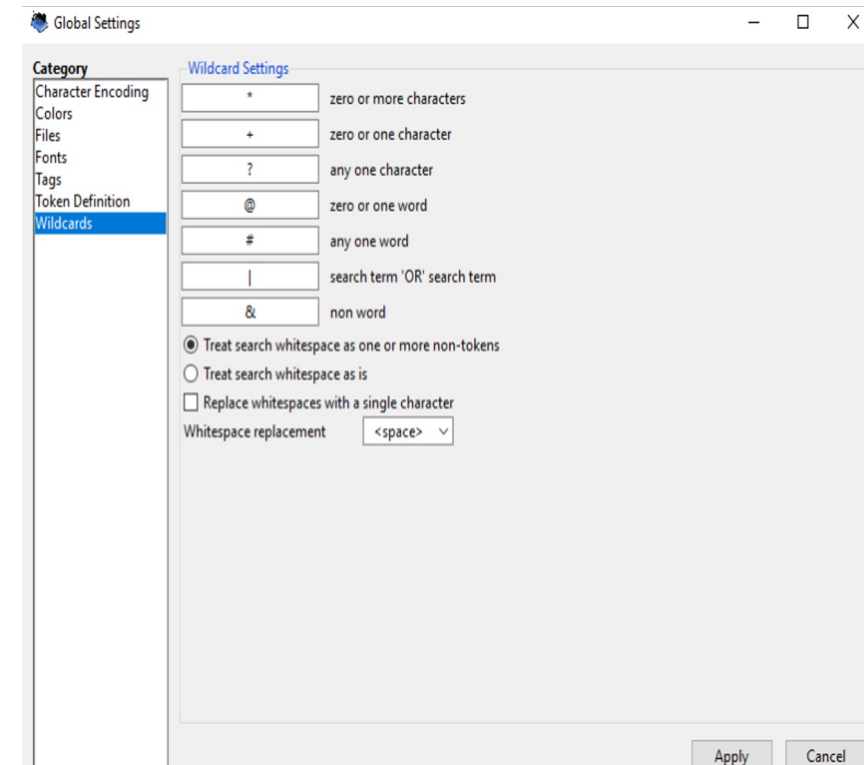
Hands-on with AntConc: Using Wildcards

A Wildcard such as ***** = **zero** or more characters following the previous letter

love* = love, loves, loved, lovers, loving, lovely

You can sort these results on **zero** word, center word, the **1 R** and **2 R**

Other **Wildcards** can be seen in the **Global Settings** tab at the top to the right of **File**



Hands-on AntConc: Using Concordance Basic Features

You can also search for strings and see them at the beginning, middle or end of words

lov or **ve** by unclicking **Word** to the right of search terms

When searching for a **word**, you will get lower and higher case options

If you want to search only for lower case **love** click on **case** option

To compare all the results that you got, use the **Clone Results** button and examine all your results in different windows to mark differences

Hands-on with AntConc: Concordance Advanced Features

As you to continue to search for words, you create a search history that can be accessed through the **up** and **down** arrows in the **search window**

You can export your results in a spreadsheet

By using **Shift+Click** you can select larger areas of text

Control+Click you can select non-sequential text

Alt+Control+A will allow you to select all the text in the results area

You can then paste it on a **Word notepad** or **excel sheet**

Control+C

❖ You can delete some of the lines

Hands-on with AntConc: Concordance Advanced Features

You can search for a number of words in the **Advanced Search** feature

Click on **Advanced**, then enter the list of words you want to search for

Click on **Use terms from the list**

Add a list of words manually or from a file

Then click on **apply**

You will realize that the word in the search window is greyed, because the application is searching for the list of words you just requested

You can even load a file of words and it will search for them all

You can also **Add** words to the context and search for **5 L** and **5 R** to the words: **said** and **say**

Click on **use context words** then **apply**, then **start**: they will appear in the vicinity of the search word **report**

Hands-on with AntConc: Concordance Advanced Features

Another feature in the **tool preference** option is the search by **character** instead of words. It will help with investigating spelling

You can **hide** the search term and all the search hits disappear

Put a **delimiter** around the hits in the **KWIC** display. The default is a tab so, adding a **delimiter**, will allow you to copy paste the search results in an excel sheet for further analysis

Another **tool preference** option is the **Line Break Replacement**: replacing it with a special character such as **%**. You can spot it in the concordance and it will help you search for words before or just after a line break

Wildcards can also help search for special words. You can change the **wildcards** symbols with other characters to prevent them interfering/overlapping with your results.

For example change the **?** wildcard symbol with **=**

Hands-on AntConc: Concordance Plot tool

Go to the **concordance plot tool**

In the **search term** window add **love**

The **concordance plot** tool will allow you to see how the word occurs in each file

Add **loved** and see the results

Now add **lover** and see the results in the second file

Clicking on the lines in the plot will take you to the file and where the word appears

Go back to the **Plot** and then go to **plot zoom** and see which area you're interested in

You can do all the searches with the **concordance** tool

You'll see how frequently words occur each of the files/texts

Hands-on AntConc: Saving Concordance Plot results

But this tool doesn't have a **save** or **clone** records option

So save by taking **screenshots** or use the **Windows snipping** tool to take snippings and save them so you can add them to your reports

Press **Window logo + Shift + S** keys

For Windows 10 Select the **Start** button, type **snipping tool** in the search box on the taskbar, and then select **Snipping Tool** from the list of results

Select the results that you want to include in your research report or presentation

Hands-on AntConc: File View

File view allows you to search for terms in the Word file uploaded on AntConc

You can search for a term or conduct advanced searches and move to see each occurrence across the whole **Word File**

If you get lost and can't find the term in a **Word file**, click on **Control + Click**

It will take you to the nearest search word in the **Word file**

If you click on any of the terms the cursor will jump to the **Concordance** tool (both are carefully linked together)

Hands-on AntConc: Clusters tool

Click on **Clusters/N-Grams** tool

This tool will allow you to find 2 or 3 word patterns in a corpus around your search term

The default setting of cluster size is **Min** and **Max 2**

Min frequency is also **1**. Change it to **2** and you'll get a different result

Min range of **1** means it occurs in at least **1 file** or more

Go back to default setting

Hands-on AntConc: Clusters tool

Most frequent cluster is **lover's** or **lover of**

Change search word to **loved** and you'll get different results

Click on any **cluster** and you'll see the patterns around them

You can **sort** the results by **range** in addition to **frequency**

Sorting by probability shows how likely the first word will occur with the second

You can use the **clone results** button to compare multiple searches

Hands-on AntConc: Clusters tool with N-Grams option

N-Grams allow you to find common expressions in a corpus without specifying a search term [so we can't access the search terms]

Select a **min of 4** & **max 4** (4 words) and then **min frequency to 10** (appearing ten times at least in the text) and **range of 3** meaning appearing at least 10 times in the 3 files of the corpus)

The power of the N-Gram tool is that it shows patterns. Click on one of these and it will take you to the **concordance** tool and then **sort** them if you want

You can **sort** in a number of ways. Explore the sorting options

If you go to **Tool Preferences** you will see the same options just referred to

Clone your results to compare them

Hands-on AntConc: Collocates tool

Collocates tool allows you to search for words that appear in close connection or association with a search term and determine the strength of that association.

Click on **Collocates** tab at the top

First type in a search word **love**

Stay with searching for single words

Set the span of words to the **L** and **R** of search words The common setting is **5L** and **5R**

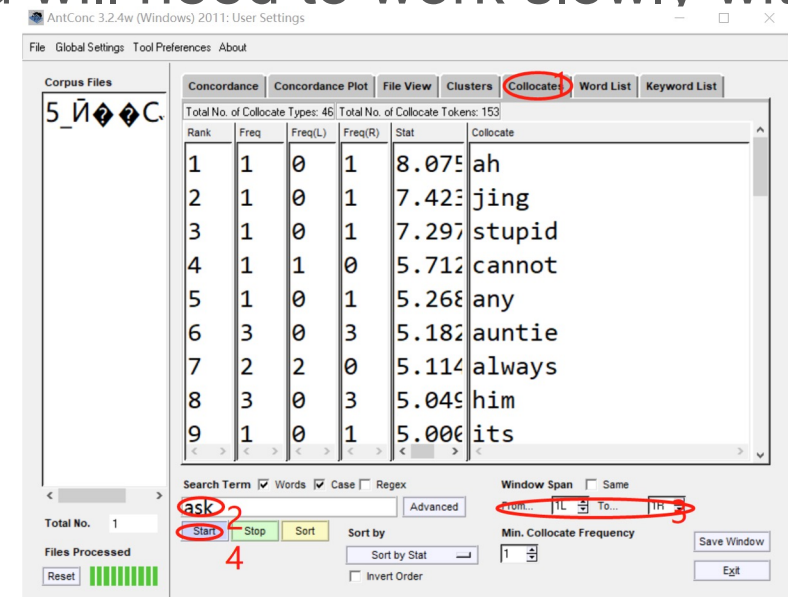
Min collocate frequency set to **1**.

Click on **start**, you'll get a warning to click on **word list**, Click **OK** and it will create the results

Hands-on AntConc: Collocates tool

The strongest association is a word with limited frequency which is a limitation here, but go to **sort results by frequency**

Click on any word and you'll go to the concordance for example: **denotes** for example and you'll see the interesting results. You will need to work slowly with this tool to get meaningful results



Hands-on with AntConc: Keyword lists

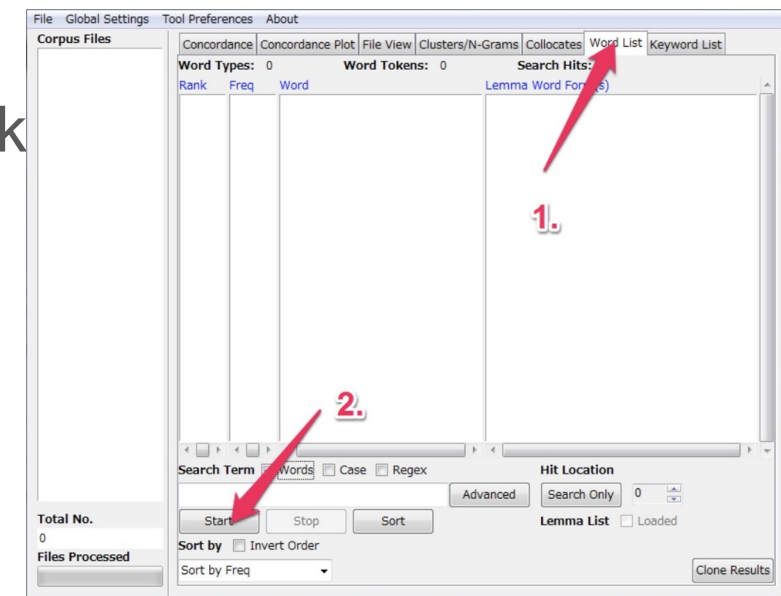
As mentioned earlier, the keyword list tool will show you the unusually frequent (or infrequent) words in a corpus in comparison with the words in a reference corpus.

This will help you to identify characteristic words/topics in a corpus, for example, create a corpus for your main author, and compare them with words/topics in another author's corpus.

Or to identify characteristic words/topics in a corpus of early works by your author, in comparison with later work.

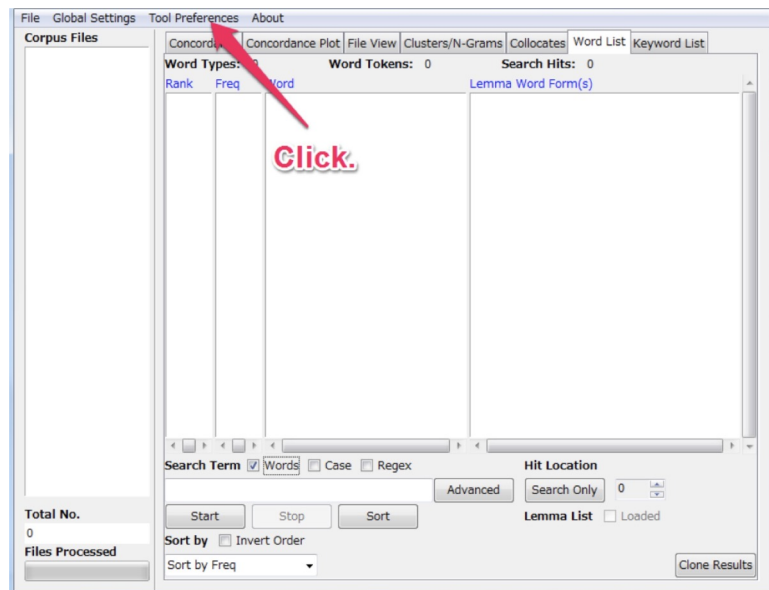
In AntConc load your corpus or corpora.

Go to **Word list** tab then click **start**.

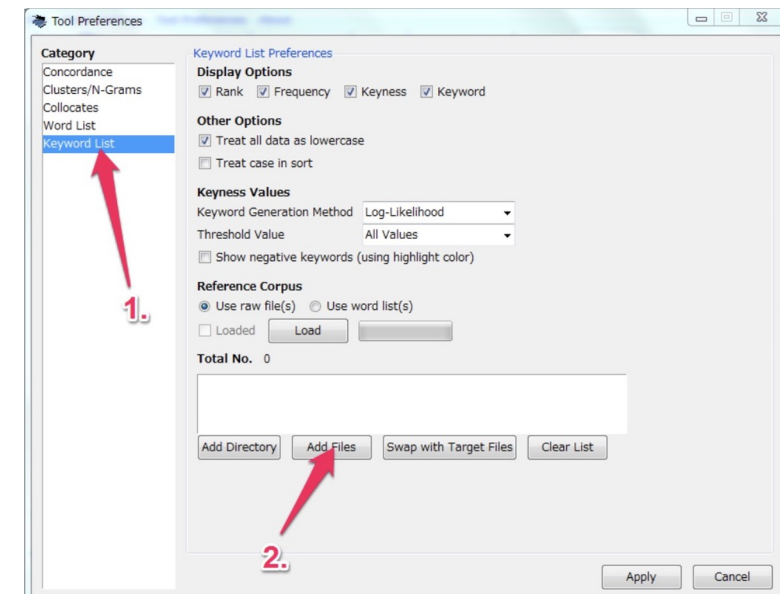


Hands-on with AntConc: Keyword lists

Select the Tool Preferences menu



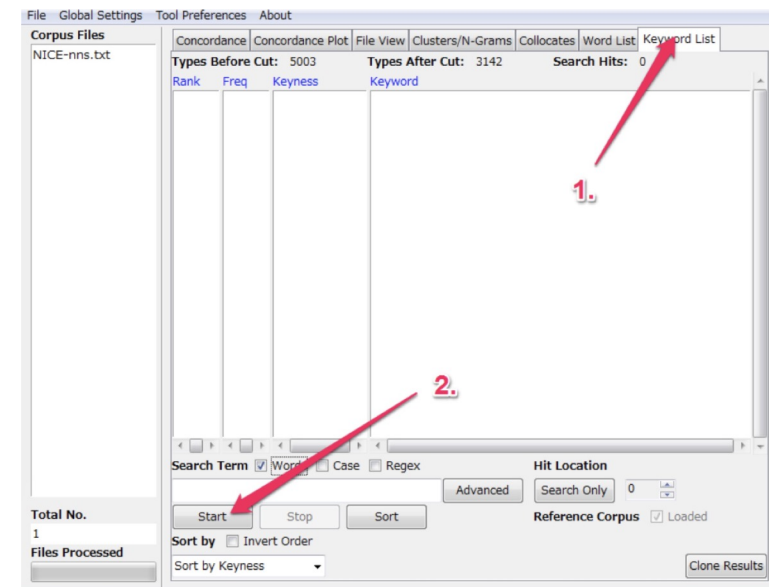
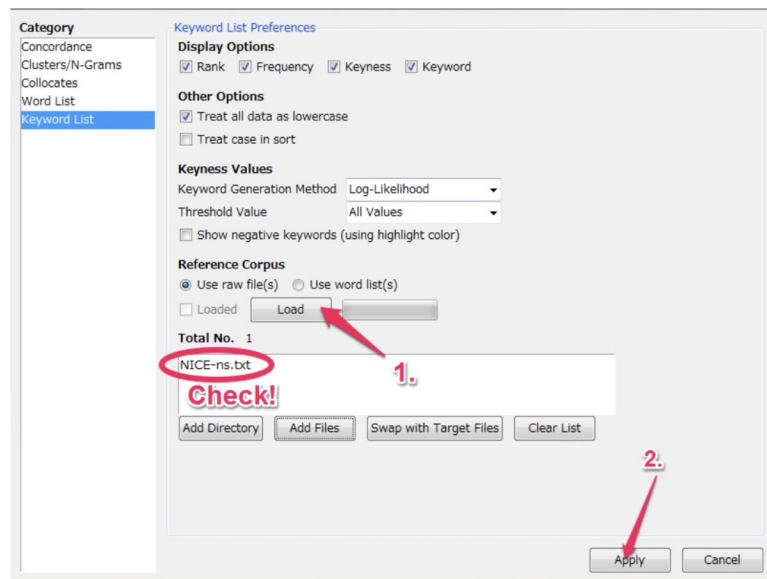
Click the 'Keywords List' option, 1.



Hands-on with AntConc: Keyword lists

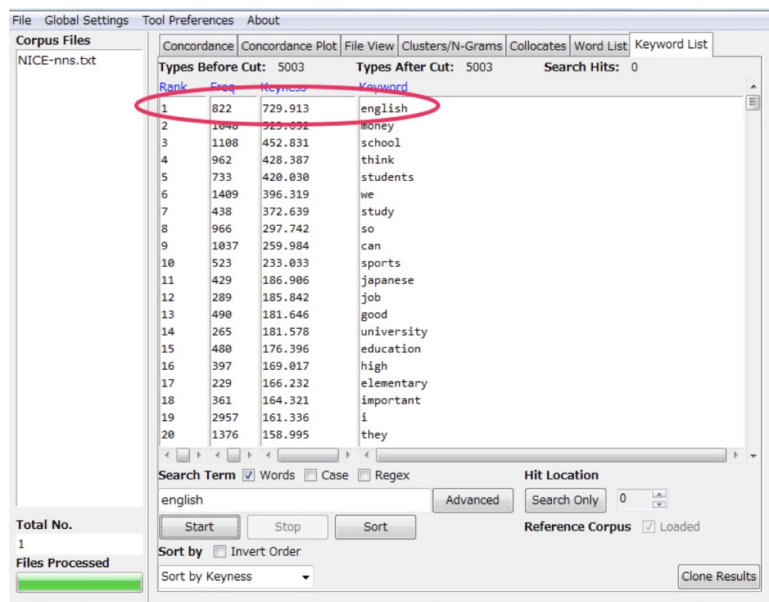
Check the desired file is there. Click 'Load' then click 'Apply'.

Go to 'Keyword List' tab then click 'Start'.



Hands-on with AntConc: Keyword lists

A list of types should appear like this.



Rank	Freq	Keyness	Keyword
1	875	833.8	english
2	1040	329.092	money
3	1108	452.831	school
4	962	428.387	think
5	733	428.030	students
6	1409	396.319	we
7	438	372.639	study
8	966	297.742	so
9	1037	259.984	can
10	523	233.033	sports
11	429	186.906	japanese
12	289	185.842	job
13	490	181.646	good
14	265	181.578	university
15	480	176.396	education
16	397	169.017	high
17	229	166.232	elementary
18	361	164.321	important
19	2957	161.336	i
20	1376	158.995	they

The keywords are ranked by default by the keyness. In this example the topranking type in “my” with a score keyness (in this example, chi-squared) of 833.8 (this is a combined score of both the target and reference type score). And it has a frequency of 875 in the target list.*

Hands-on with AntConc: Keyword lists

Using the **Treat all data as lowercase** option causes all words to be transformed to lower-case words. This is useful to get accurate counts of words in certain cases.

Using the **Treat case in sort** option causes capitalized words to appear before lower-case words.

Use the **Show Negative Keywords** option to view words that are unusually INFREQUENT in the target corpus compared with the reference corpus.

*Use the **Keyness Values** options to choose the statistical measure for measuring keyword strength, the statistical threshold used to decide a cutoff point, the effect size measure used to rank the keywords, and the effect size threshold to decide the cutoff of the effect size.*

Currently, two statistical measures can be used: Chi-Squared and Log-Likelihood (the default). The default option for the 'keyness' measure is recommended, however, you may want to experiment with different effect size measure depending on the corpus you are using. The selection of statistical measures and effect size measures are inspired by the work of Andrew Hardie of Lancaster University.

Hands-on with AntConc: Keyword lists

*Use the **Use raw file(s)** option to use raw reference corpus file(s) as the reference corpus.*

Use the **Use word list(s)** option to use word list(s) that correspond to a reference corpus. The word list(s) should be formatted as described in the tool explanation.

Click the **Add Directory** or **Add Files** buttons to select the reference corpus files.

Click the **Swap with Target Files** button to swap the main and reference corpora. Note that this will only make sense when raw corpus files are being used.

How do we interpret and write up our results with AntConc?

You need to test your findings before you interpret your results by clicking on the **Concordance** and closely reading/checking your texts/files

You also need to compare results, using the **Clone Results** button. This will allow you to examine all your results in different windows & mark differences

Some tools results can't be cloned (**e.g., Concordance Plot**), so you'll need to either create a screenshot of your visualization and crop it through an image editor, or use **Windows Snippet tools** see *slide 20*

You also need to support your results with visualizations exported from the tools

How do we interpret and write up our results with AntConc?

In addition, you need to write the processes you went through starting with:

- ❖ your initial hypotheses and research questions
- ❖ where you located your data
- ❖ how you prepared it & uploaded it on AntConc
- ❖ which tools you used in AntConc
- ❖ how you reached some analyses and answered your initial research questions
- ❖ and, if applicable, how you revisited and developed new research questions during the research process

+

•

○

Reflections on Antconc

- ❖ What can you do with Antconc?
- ❖ What can't you do?
- ❖ When and where will you be using it?

+

•

○

Computer-Assisted Textual Analysis tools Discussion

- How have you used, plan to use, are reluctant to use computer-assisted textual analysis methods and tools
- What do text analysis “platforms,” “web-based reading environments” provide you that you benefit from?
- What do they leave you wondering about?

+
•
○

Computer-Assisted Textual Analysis (optional) assignment for our next session

- Where in your syllabus would it make sense to introduce computer-assisted textual analysis methods of analysis?
- When in the syllabus do you want to integrate training for your students on the tool(s)?
- Write a draft of an assignment that you would want your students to do using one or all of the tools introduced in this session
- Draft a rubric for the above assignment

Resources and Tutorials used in the hands-on sections

1. <https://docs.google.com/presentation/d/1dH1xtthuyzETvRC671JZuHsg5aspF-Ju3EIElp3sdM0/edit?usp=sharing>
2. https://AntConc-manual.readthedocs.io/en/latest/keyword_list.html
3. <https://AntConc-manual.readthedocs.io/en/latest/clusters.html>
4. <https://AntConc-manual.readthedocs.io/en/latest/collocates.html>
5. https://AntConc-manual.readthedocs.io/en/latest/keyword_list.html
6. https://www.youtube.com/watch?v=2rvsBaM6W8Y&list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS_TZj&index=4
7. https://www.youtube.com/watch?v=gSDMNnRbm3M&list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS_TZj&index=6
8. https://www.youtube.com/watch?v=1FG9N3jqlHY&list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS_TZj&index=9
9. https://www.youtube.com/watch?v=7gMIEXfLYyE&list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS_TZj&index=8
10. https://www.youtube.com/watch?v=DV8YC0zKBOg&list=PLiRIDpYmiC0Ta0-Hdvc1D7hG6dmiS_TZj&index=7

QUESTIONS



njarkas@amicalnet.org