The Poetry of Frequency: Using Voyant in English Pedagogy, Revision, and Beyond

E. Elizabeth Watkins
IS 578: Introduction to Digital Humanities
Prof. Zoe LeBlanc
18 December 2023

Table of Contents

Introduction

We analyze text subjectively and instinctually as we read, but what about quantitatively? What about story-wide trends in language that are so subtle that the average reader rarely catches on to them? Looking at an author or poet's entire body of work and identifying trends in word frequency or syntax between poems, novels, or eras across the writer's lifespan can provide invaluable insight into their process, habits, and craft. With the human eye alone, it may prove extremely difficult to identify these trends, and there is no guarantee that one's findings are entirely accurate because of human error. Further, in what ways could this type of analysis be applied to one's own writing, and what insights could be gained from looking at quantitative data generated by computers from looking at one's own work?

One might argue that this may result in a cold, purely data driven conclusion about what words one is drawn to, or that technology is attempting to replace human analysis. Using text analysis tools are not the best or only method for looking at text, by any means. This workshop is in no way attempting to devalue human analysis of text or replace it, but to offer an a new perspective for looking at literature and revision that can be seamlessly blended into one's existing process.

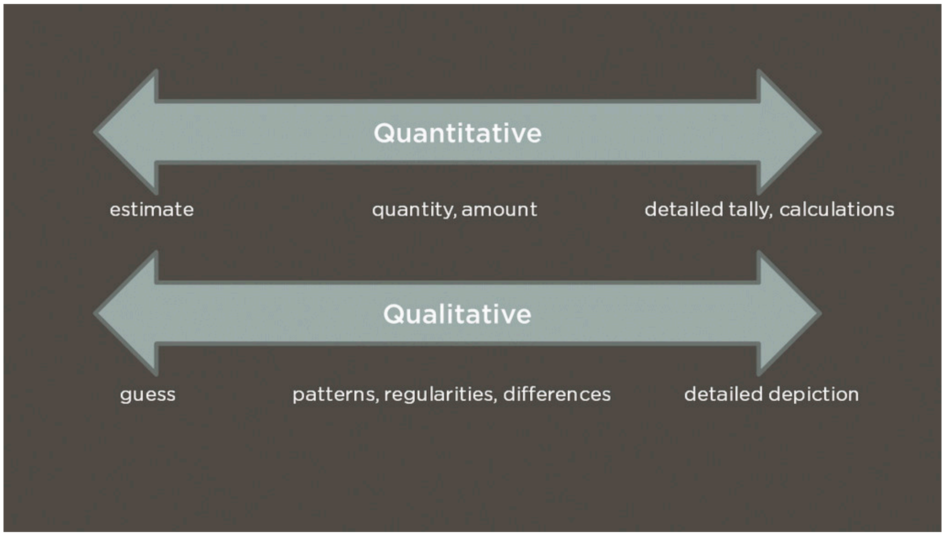
Workshop Goal

The ultimate goal of this workshop is not to make everyone experts in using Voyant; rather, it is to introduce the idea of using text analysis tools in English and creative writing classrooms, and to provide basic knowledge of how to use them.

It is perfectly acceptable to walk away from this workshop and decide that teaching students about the benefits of text analysis tools is not a good fit for your curriculum. However, it is my hope that at least introducing this concept to instructors in higher education may help them become more open minded to using digital humanities tools in general, as they are part of an emerging field that will undoubtedly pervade the academic careers of their students.

The Digital Humanities, Briefly Explained

There is a widely contested definition for what exactly the digital humanities are, and you're free to disagree, but it is the term we use for describing the integration of tools, software, and theoretical frameworks to process qualitative and quantitative data in humanities disciplines. For brevity, it is sometimes referred to as 'DH.' You may have noticed the use of the word 'tools' in that definition—essentially, DH tools are meant to present, interpret, generate, or somehow transform data to prepare it for analysis. This will make more sense in a bit when we get to working with Voyant, but it is important that you have this context for what DH actually is.



It may also be helpful to understand the concepts of quantitative and qualitative data, if you are not already familiar or have not taken a research methods class in a while. Quantitative data is numerical, and qualitative data is descriptive. For example, say that you have three red apples. The number of apples is a unit of quantitative data, versus the color of apple, which is considered qualitative data. Pretty simple, right? In this workshop we are going to be focused on a type of quantitative data called frequency, specifically within the context of how many times a word is used. Word frequency data can be useful for identifying recurring themes, emotions, or patterns within a text, but it requires context. Words can have multiple meanings, which can skew our analysis if we are not careful.

Introduction to Voyant

Voyant is text analysis platform that allows you to upload one or more texts and analyze them using a variety of tools. It's similar to a buffet-style meal where you can choose from a large variety of foods based on what you are hungry for. Instead of your craving for certain kinds of food, it's your personal research needs, and instead of the food, it's the tools in Voyant!

In Voyant, a single body of work is referred to as a document. As it is possible to analyze multiple documents at the same time, documents that are grouped together to be analyzed concurrently are called a corpus. If you plugged every single thing Jane Austen ever wrote into Voyant, for instance, *Pride & Prejudice* would be a considered a document part of the greater corpus of Austen's work.

The tools in Voyant generate graphs and tables to demonstrate the relationship between different words or phrases and how they're used in a particular document or across the corpus. Calculating word frequency is just one of the ways that Voyant can identify trends. To summarize, using Voyant is a way to identify trends in language and generate tables and visualizations based on those trends.

Voyant is just one of many text analysis tools; for a list including other tools, check out this LibGuide from the University of Illinois at Urbana-Champaign's library.

Things To Keep In Mind

Try to reflect upon your learning experience periodically throughout this tutorial. Think about how you feel about Voyant and digital humanities tools in general as you navigate this new frontier, and consider the following questions:

- ❖ What is your comfort level with Voyant right now? Do you think if you had more time to experiment with it that you could get used to it?
- ❖ What is the most interesting feature you've encountered? the most difficult or confusing thing?
- ❖ Can you identify an aspect of your own research where you might implement Voyant?
- ❖ Has your opinion of implementing digital humanities tools in your workflow changed?

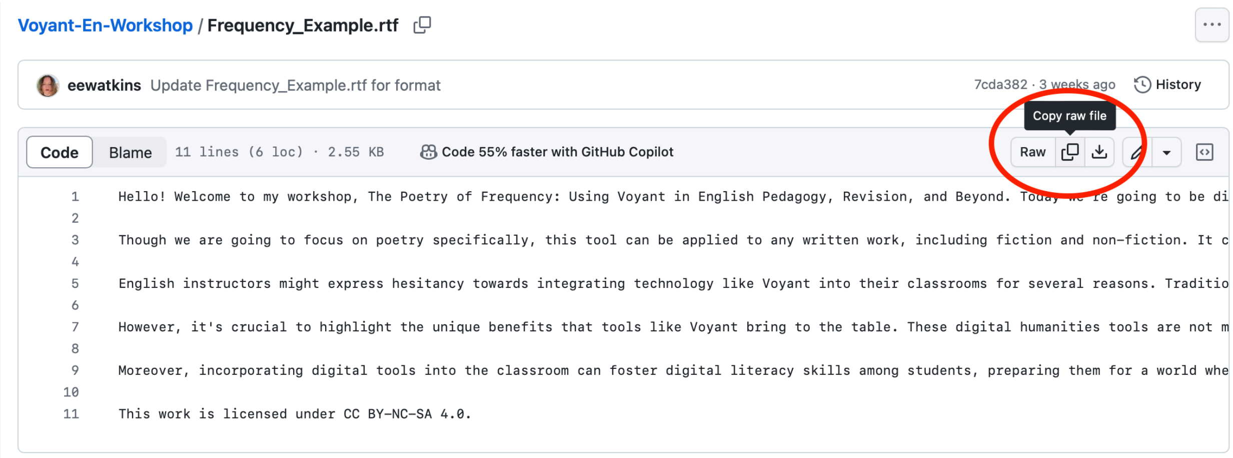
Accessing Voyant and Addressing Privacy Concerns

If you open this link, it will take you to this workshop's GitHub repository that contains all the files and other links we'll need for today. To get started, head to https://voyant-tools.org, Voyant's start page, or you can download a local version called VoyantServer using these instructions.

Are you unsure whether to use Voyant or VoyantServer? The difference between the two is that VoyantServer is a locally-hosted server—meaning all data is stored on your device—whereas Voyant is web-based, meaning that data is stored on Voyant's servers. If you are worried about the privacy of your work and uploading unpublished prose or poetry to a non-local/web-based server, that's a very valid concern! VoyantServer is the better option for you, as everything is stored on your personal computer and no data leaves it. For our workshop's purposes, since we are using work that is in the public domain, it may not be necessary to download VoyantServer, but you may do so if you choose. Keep in mind that this workshop was built using the online version of Voyant, so instructions may vary a bit.

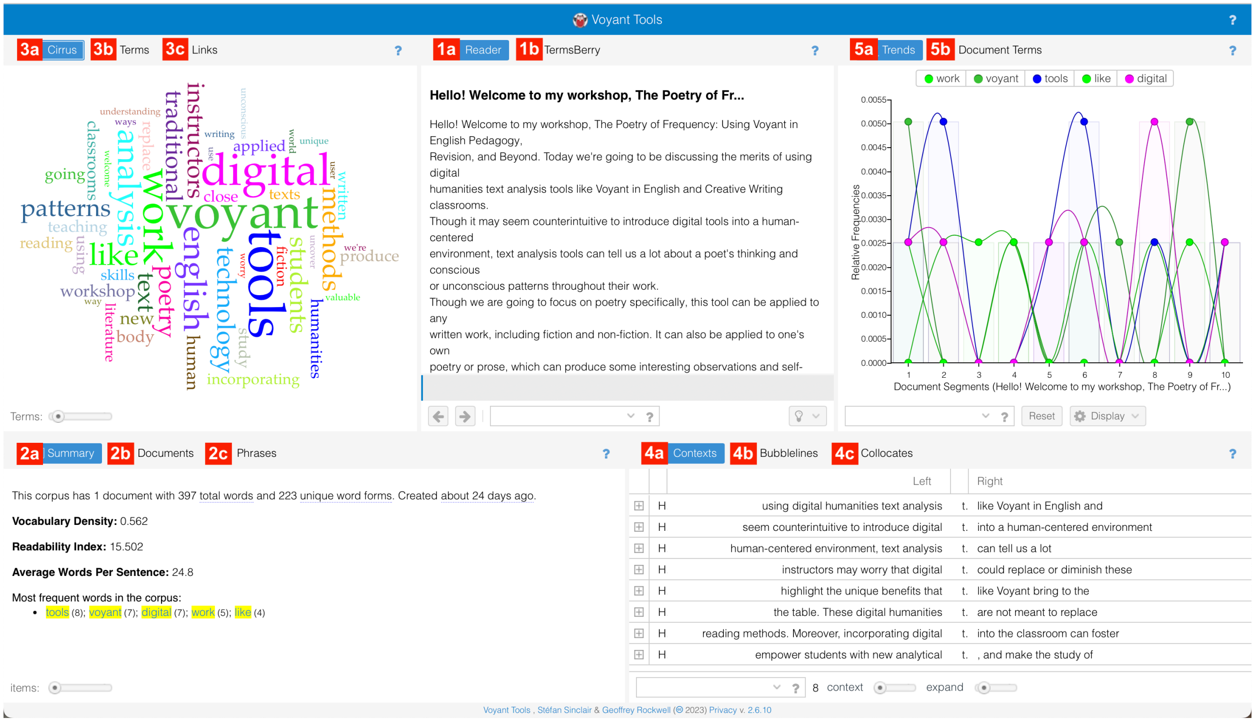
Uploading a Corpus and Getting Acquainted with Voyant

There are a few different ways to upload texts to Voyant; you can paste in text or a URL, or you can upload HTML, XML, PDF, RTF, or MS Word files. In this workshop, we will start out by copying the text of a short .RTF document to get a feel for the interface using a small, manageable corpus. An .RTF, for those unfamiliar, is a Rich Text File—all it contains is formatted text. Copy the raw file titled “Frequency_Example.rtf” by clicking into the file on the GitHub repository and clicking on the two overlapping squares on the righthand side of the page.


Voyant-En-Workshop / Frequency_Example.rtf
eewatkins Update Frequency_Example.rtf for format
7cda382 · 3 weeks ago
History
Copy raw file
Code
Blame
11 lines (6 loc) · 2.55 KB
Code 55% faster with GitHub Copilot
Raw

```
Hello! Welcome to my workshop, The Poetry of Frequency: Using Voyant in English Pedagogy, Revision, and Beyond. Today we're going to be di

Though we are going to focus on poetry specifically, this tool can be applied to any written work, including fiction and non-fiction. It c

English instructors might express hesitancy towards integrating technology like Voyant into their classrooms for several reasons. Traditio

However, it's crucial to highlight the unique benefits that tools like Voyant bring to the table. These digital humanities tools are not m

Moreover, incorporating digital tools into the classroom can foster digital literacy skills among students, preparing them for a world whe

This work is licensed under CC BY-NC-SA 4.0.
```

Paste this into the text box on https://voyant-tools.org and click on the blue button that says ‘Reveal’. The user interface, also referred to as the default skin, should look something like this:



Hello! Welcome to my workshop, The Poetry of Fr...

Hello! Welcome to my workshop, The Poetry of Frequency: Using Voyant in English Pedagogy,
Revision, and Beyond. Today we're going to be discussing the merits of using digital
humanities text analysis tools like Voyant in English and Creative Writing classrooms.
Though it may seem counterintuitive to introduce digital tools into a human-centered
environment, text analysis tools can tell us a lot about a poet's thinking and conscious
or unconscious patterns throughout their work.
Though we are going to focus on poetry specifically, this tool can be applied to any
written work, including fiction and non-fiction. It can also be applied to one's own
poetry or prose, which can produce some interesting observations and self-

This corpus has 1 document with 397 total words and 223 unique word forms. Created about 24 days ago.

Vocabulary Density: 0.562

Readability Index: 15.502

Average Words Per Sentence: 24.8

Most frequent words in the corpus:

- tools (8); voyant (7); digital (7); work (5); like (4)

		Left		Right
⊞	H	using digital humanities text analysis	t.	like Voyant in English and
⊞	H	seem counterintuitive to introduce digital	t.	into a human-centered environment
⊞	H	human-centered environment, text analysis	t.	can tell us a lot
⊞	H	instructors may worry that digital	t.	could replace or diminish these
⊞	H	highlight the unique benefits that	t.	like Voyant bring to the
⊞	H	the table. These digital humanities	t.	are not meant to replace
⊞	H	reading methods. Moreover, incorporating digital	t.	into the classroom can foster
⊞	H	empower students with new analytical	t.	, and make the study of

The various tools have been labeled in red to make them easier to identify. Here are a few quick explanations for each tool in the default skin for your reference:

1a. **Reader**: shows the selected document's full text. You can hover over a word for its frequency, and search for words throughout the document.

1b. **TermsBerry**: visualizes the relationships between words in the corpus and is useful for understanding semantic connections and the co-occurance of terms.

2a. **Summary**: gives general statistics about the selected document and corpus at large, like the readability index, total words, number of unique words, and more.

2b. **Documents**: lists all documents in the corpus with general statistics.

2c. **Phrases**: shows a table of common phrases and their frequencies.

3a. **Cirrus**: a visualization that shows how often words are used throughout the corpus with the size of the word indicating its frequency.

3b. **Terms**: shows all terms in the corpus and their frequencies. The trend column is useful for corpora with multiple documents to show trends over all documents.

3c. **Links**: shows proximal relationships between words; hover over them to highlight which words are used near each other.

4a. **Contexts**: shows every instance of the usage of a selected word and the words that surround it.

4b. **Bubblelines**: visualizes word usage over time; the size of the bubble indicates a word's frequency.

4c. **Collocates**: shows words that often appear in the vicinity of a selected word—similar to Contexts but a bit more 'diluted,' meaning the scope of close words is broader.

5a. **Trends**: generates a graph of the relative frequency of a few selected words over the entire document/corpus.

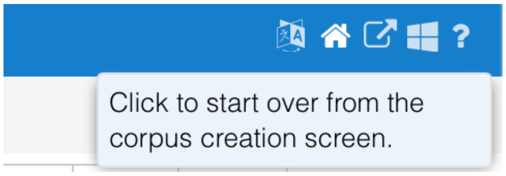
5b. **Document Terms**: very similar to Terms, but also shows trends in a single document, not the full corpus.

If you would like to find out more about these tools, Voyant provides handy instructional documentation for each one if you have trouble fully understanding what something does.

Take some time to mess around and experiment with the tools and see if you can think of any ways the tools or the statistics they provide might be applicable to larger documents/corpora. For example, think about why we might want to look at the context a particular word is found in, or why looking at a word's usage over the entire document is important or useful information. In the next step you will be taking a look at a large collection of poems written by a single author, so using this opportunity to familiarize yourself with the default skin while looking at a smaller and more manageable corpus will be beneficial.

Long-Form Example: Emily Dickinson’s Complete Works

Now that you’ve become at least acquainted with Voyant, let’s try using a more complex corpus! In Voyant, you can reset the interface by clicking on the ‘Home’ button on the top right of the blue navigation bar; if you hover over it, the little house icon will appear.



Click to start over from the corpus creation screen.

When prompted to make sure you want to start over, click ‘yes.’ This will take you back to Voyant’s homepage, and you will be all set to start over with a new corpus.

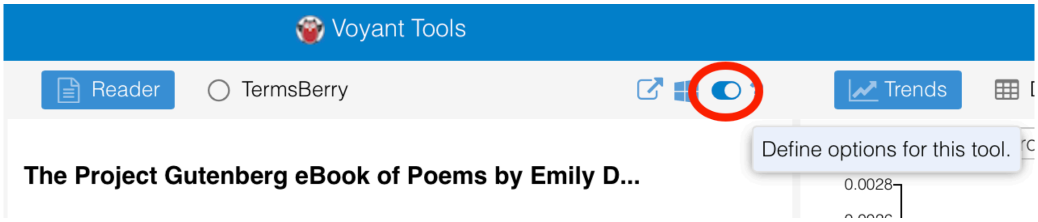
Voyant has a few corpora already uploaded that you can explore by default, such as Shakespeare’s plays, Austen’s novels, and Shelley’s *Frankenstein*. You’re welcome to peruse those on your own if you wish and can access them by clicking the ‘Open’ button on the left side of the homepage. For the next section of this workshop, we’re going to work with a text file that contains Emily Dickinson’s complete works, courtesy of Project Gutenberg.

For those unfamiliar with Project Gutenberg, it is a free, online library for out of copyright works that have entered the public domain and contains full texts from a host of different genres and time periods. It's a great resource for finding documents to plug into Voyant and other text analysis tools because of how easy it is to use, but it is important to be aware of the pitfalls of such a repository. The types of texts it contains skew heavily to works in English that were written pre-1923, due to copyright restrictions and general academic bias towards works written in English; if you work heavily in contemporary translation, for example, it might be a bit harder to find what you are looking for. Keep this in mind as you consider if text analysis tools are feasible for your research and pedagogical goals.

In the GitHub repository, use the same process as before for copying the raw text of the file named “EDCompleteProjectGutenberg.txt.” Paste that into the text box on Voyant’s homepage and click ‘Reveal’ again. This will take you to the default skin/user interface for this particular corpus.

You may notice that among the top words for this corpus are ‘project,’ ‘gutenberg,’ and ‘gutenbergâ’; these words are skewing our findings and are considered outliers. You can see the manifestation of this skew in the Trends graph—the word ‘project’ spikes towards the end of the document. Gutenberg’s terms of use unfortunately prohibit me from removing the Project Gutenberg front and end matter from the .TXT document I provided to you in this instructional setting. You are welcome to remove this excess on your own, but this is also a good opportunity to talk about stopwords.

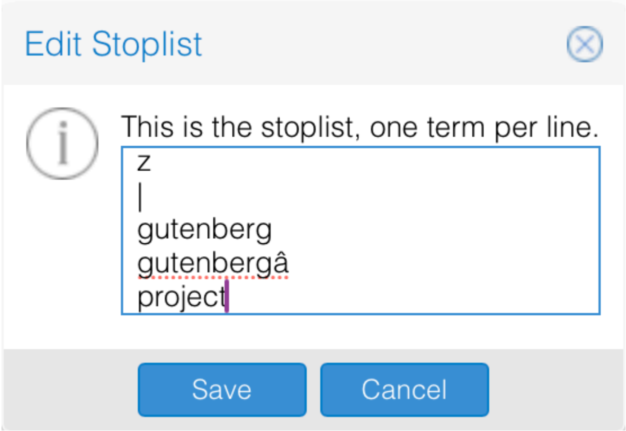
Stopwords are words that have been deemed to have little contextual value—like prepositions, for example—and are excluded from the results of our data. We can edit the list of default stopwords by hovering over and clicking on the switch button located at the top right of any tool in Voyant. It looks like this:



Define options for this tool.

The Project Gutenberg eBook of Poems by Emily D...

Once there, you'll want to click on 'Edit list,' and then type in the three outliers previously mentioned, each on their own line: 'project,' 'gutenberg,' and 'gutenbergâ'.



This is the stoplist, one term per line.

```
z
|
gutenberg
gutenbergâ
project
```

After clicking 'Save,' make sure that the 'apply globally' box is checked, as we want this change to apply to all of the tools we're working with. Once you've clicked 'Confirm,' your graphs and tables should shift and look a bit different, and hopefully more accurate.

Now that our data is a little more presentable, let's take a look at the Context tool. Say that we want to identify every instance of simile in Emily Dickinson's work; how might we do that? You'll notice that the most commonly used word, 'like,' is already selected in the Context tool. From here, we can see every time that Dickinson uses the word 'like' to produce a simile based on the context surrounding it.

'As,' the other word used to create simile, is a bit trickier to narrow down. The Phrases tool is a bit more useful for this, since 'as' is used in combination with different words like 'a' or 'the' to create a simile. Because it is on our list of stopwords, we'll first have to edit the list to include it in our data using the same steps we followed to add our outliers. Once you've deleted 'as' from the stopwords list, type it into the search bar on the bottom of the Phrases tool and select it from the dropdown. You can now adjust the 'Length' slider at the bottom to narrow down the number of words in the phrases that are included in the results. Using the Context and Phrases tools, we are now able to identify all instances of simile in Emily Dickinson's work! By locating every simile, we can do a thorough, systematic assessment of what context Dickinson uses it in and generate our own analyses of it. Finding this information manually would take ages. The use of text analysis tools in tandem with human analysis makes for a thorough, efficient, and intelligent review of any given writer's work.

Categories Feature

Throughout this workshop, you may have noticed that a select few terms on Voyant's interface have been highlighted in green and red. This is part of a new, currently-in-development feature called 'Categories' that aims to provide instantly recognizable connotational classifications to terms that Voyant's developers have identified as either 'positive' or 'negative' by labeling each respective category with a color. This can be useful for identifying tonal trends/moods and shifts throughout a document or corpus, and you can even create your own categories of words and color code them as desired.

However, as any librarian will tell you, the act of classification inherently imposes a power dynamic on the things being classified. Digital humanist Ted Underwood identifies this issue as "researchers [using] present-day patterns of association to define a wordlist that they then take as an index of the fortunes of some concept (morality, individualism, etc) *over historical time*" in his blog post 'How not to do things with words.'[1] When dealing with a text from an era or culture different from one's own, it is imperative to consider it in context and

[1] Ted Underwood, "How Not to Do Things with Words.," August 14, 2012, https://tedunderwood.com/2012/08/25/how-not-to-do-things-with-words/.

eliminate assumptions about a word's use whenever possible. Similarly to how word frequency needs to be considered with a grain of salt, so does the Categories feature. For a more in-depth explanation of how to use Categories, check out Voyant's instructional documentation.

Questions for Reflection and Implementing DH Tools in Instruction

You may wish to explore Dickinson's work a bit more on your own, and are absolutely encouraged to do so. While you explore, think about how you might implement it in your own classroom using these guided questions:

- How would you orient students to this new technology?
- What tools or features would you highlight as useful?
- What sort of assignment could you create to encourage students to become familiar with Voyant and text analysis tools in general?
- What has this workshop *not* covered that might be important to include in a tutorial?

Thank you for participating in *The Poetry of Frequency: Using Voyant in English Pedagogy, Revision, and Beyond*. It is my hope that it provides a basic tutorial for instructors new to digital humanities tools and digital text analysis, and a framework for how it may be implemented in the classroom. Questions about this workshop may be directed to emmaw5@illinois.edu.

Author's Note

Creating a digital humanities workshop for an audience that is sure to have a few Luddites is no walk in the park. I understand very well the urge to cower in fear when someone utters the words "digital humanities" or warns of the damage that generative AI can cause to the integrity of written work. It might be my creative upbringing at a small, rural, liberal arts college showing, but that fear has been plenty instilled in me. After the last five months of learning what the digital humanities are all about, I can say that there's not nearly as much to be afraid of as I thought. Text analysis tools are the next logical step for English and creative writing pedagogy in higher education.

Voyant can pick up on patterns that humans just miss. In *The Poetry of Frequency*, I attempt to show how computing can actually blend seamlessly with both analytical and creative endeavors by only replacing the excavation of data in certain types of contextual analysis. This only makes research more efficient with minimal effect to the text-reader relationship, and despite the fact that new DH users are often looking "to produce something new and dramatically different from what non-DH methods allow," I argue that it may be better for usage retention purposes to start by incorporating digital humanities tools into users' daily workflows.[2] As is often the case in librarianship, what the user wants is not necessarily what the user needs in reality; this discord is usually caused by a gap in knowledge of applicable resources, which is acknowledged not to disparage users, but rather to educate them on the full scope of services and resources available to them. This integrative approach could encourage users to get comfortable

[2] Paige Morgan, "The Consequences of Framing Digital Humanities Tools as Easy to Use," *College & Undergraduate Libraries* 25, no. 3 (August 7, 2018): 219, https://doi.org/10.1080/10691316.2018.1480440.

with the nature of digital humanities tools before attempting to add entirely new methods to their workflow, with the resulting failure potentially barring them from learning how to use DH tools entirely.

While my workshop is geared towards those with little to no DH experience, it was especially important to me during its composition and revision that I did not fall into the trap of explicitly promising that Voyant was extremely easy to use. I wanted to let the tool and the workshop experience speak for themselves in that regard, as the way tools are framed has significant impact on learning outcomes.[3] If a tool is framed as easy and someone finds it to be frustrating, they may abandon the workshop entirely out of frustration. To avoid this, I placed extra emphasis on making my instruction simple, clear, and sticking to the workshop's goal: educating instructors and providing a baseline of knowledge for text analysis tools.

Avoiding buttonology—the practice of simply giving a baseline tutorial of a tool's features without a critical or reflective angle—was a bit of a challenge while crafting this workshop, as any one-shot tutorial aimed at beginners is bound to fall into simple step-by-step instruction. By integrating metacognition into the curriculum by asking participants to reflect upon their experience in multiple instances during a workshop, instructors can guide participants to think about their own learning experience to "build self awareness" that leads to increased problem solving capability and a more complex understanding of the tool they are learning how to use.[4] This is also an important consideration when thinking about the tone of a workshop—

[3] Morgan, "The Consequences of Framing Digital Humanities Tools as Easy to Use": 214.

[4] John E. Russell and Merinda Kaye Hensley, "Beyond Buttonology: Digital Humanities, Digital Pedagogy, and the ACRL Framework," *College & Research Libraries News* 78, no. 11 (December 4, 2017): 590, https://doi.org/10.5860/crln.78.11.588.

being too patronizing can sour the experience or turn participants away from being open to learning with new technologies.

This workshop uses a semi-exploratory approach to achieve its learning objectives. While it does provide explicit instruction to get participants comfortable with Voyant's interface, I always feel it is important to also encourage spontaneous discovery of tools and features that may be useful to them. I had a mixture of Montessori-style and traditional educational methods throughout my own education, and while I do not explicitly favor one or the other, I feel that curiosity has been a driving force in how I choose to explore new concepts and it has been overall beneficial to how I learn. This approach places agency in the hands of the learner and forces them to think about how they are learning, which again ties in the concept of metacognition that leads to the development of critical thinking skills. This workshop does not simply throw participants into the deep end of the pool in the hope that they will eventually learn to tread water. Instead, it provides participants some guiding questions to facilitate the discovery process and gives structure to what might otherwise be an aimless and confusing journey.

The Poetry of Frequency does not pretend to be exceptionally revolutionary in its subject matter. It is a simple text analysis workshop geared toward instructors in higher education with little to no digital humanities experience. Where it is set apart is in how it approaches instructors; by meeting instructors where they are at without being overly patronizing, the workshop is able to provide an experience that is unintimidating, accepting of trial and failure, and comprehensive appropriate to their level of experience, hopefully leading to an increase in usage retention of digital humanities tools in higher education spaces.

Bibliography

Morgan, Paige. “The Consequences of Framing Digital Humanities Tools as Easy to Use.” *College & Undergraduate Libraries* 25, no. 3 (August 7, 2018): 211–31. https://doi.org/10.1080/10691316.2018.1480440.

Russell, John E., and Merinda Kaye Hensley. “Beyond Buttonology: Digital Humanities, Digital Pedagogy, and the ACRL Framework.” *College & Research Libraries News* 78, no. 11 (December 4, 2017): 588–600. https://doi.org/10.5860/crln.78.11.588.

Underwood, Ted. “How Not to Do Things with Words.,” August 14, 2012. https://tedunderwood.com/2012/08/25/how-not-to-do-things-with-words/.