19
Text Analysis and Visualization: Making Meaning Count

Stéfan Sinclair and Geoffrey Rockwell

> Un des problèmes de la sémiotique serait … de définir la spécificité des différentes organisations textuelles en la situant dans *le texte général (la culture)* dont elle font partie et qui fait partie d'elles. (Julia Kristeva)[1]

Which Words are used to describe White and Black NFL Prospects?

In May of 2014 the sports website *Deadspin* carried an article about the words used by National Football League (NFL) scouts reporting on black and white prospects (Fischer-Baum *et al*., 2014). They found differences. White players were more likely to be called "intelligent" and blacks more likely to be called "natural." They had compiled a collection of texts – a corpus – and analyzed it with *Voyant Tools*.[2] Digital humanities methods and tools had come to sport journalism.

But *Deadspin* went a step further. Instead of discussing the difference in vocabulary they provided an "interactive" for readers to try comparisons (they use "interactive" as a noun, a ellipsis for something like an interactive widget). You type in a word to search for and the interactive returns a simple bar graph that you can drop into a comment (Figure 19.1), as hundreds of readers did. They used a simple interactive text visualization to make their point.

This chapter is about such text analysis and visualizations.[3] The analytical practices of the digital humanities are becoming ubiquitous as digital textuality continues to surround and overwhelm us. This is an introduction to thinking through the analysis and visualization of electronic texts. We start by asking again what an electronic text is in the context of analysis – a preliminary but crucial first step. Then we look at how

A New Companion to Digital Humanities, First Edition. Edited by Susan Schreibman, Ray Siemens, and John Unsworth.

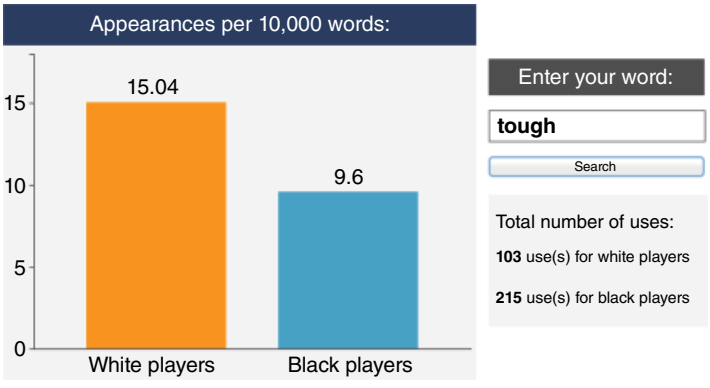

Figure 19.1 An interactive text analysis and visualization widget by *Deadspin*.

analysis takes apart the text to recompile it in ways that let you reread it for new insights. Finally we will return to how interactive visualizations bear meaning.

Ubiquitous Text

Text may be less flashy and less glamorous than other forms of communication such as sound, image, and video, but it remains the dominant way that humans communicate, discover, and process information. It is estimated that every day some 200 billion emails are sent and some 5 billion Google search queries are performed – and they are nearly all text-based.[4] The hundred hours of video uploaded to YouTube every minute would remain largely inaccessible were it not for text-based searches of the title, description, and other metadata. Even if we hesitate to join the poststructuralist theorists (like Kristeva, quoted above) in saying that *everything is text*, we can certainly agree that *text is everywhere*.

For humanities scholars and students working with texts as cultural artifacts, it is reassuring to recognize that people from every sector in our digital society are struggling with how to derive meaning from texts, from high-school students researching an essay topic to journalists combing through leaked security documents, or from companies measuring social media reaction to a product launch to historians studying diversity of immigration based on more than two centuries of trial proceedings.[5] The particular texts, methodologies, assumptions, and objectives vary widely between different applications, of course, but fundamentally we are all trying to gain insights from the vast amount of text that surrounds us.

We are unrelentingly bombarded by text in our lives and we have access to unfathomable quantities of other texts.[6] Yet for some, the problem is the opposite one: a dearth of readily accessible and reliable digital texts, whether because of legal reasons (like copyright or privacy), technical challenges (such as the difficulty of automatically recognizing characters in handwritten documents), or resource constraints that make it impractical to digitize everything (parish records scattered throughout the world, for instance). As a result, there is a significant inequality in the availability of digital texts, one that has a profound effect on the kinds of work that scholars are able to pursue.

When text *is* available there can be so much of it that we naturally seek ways of representing significant features of it more compactly and more efficiently, often through visualization. Visualizations are transformations of text that tend to *reduce* the amount of information presented, but in service of drawing attention to some significant aspect. For example, if you wanted to make an argument about the differences between the vocabulary used in mainstream commercials for toys targeted at girls compared with toys targeted at boys, you could simply compile examples from a sample set of about 60 advertisements and invite your reader to peruse the full texts. Or you could create *word cloud* visualizations for each gender, as Crystal Smith (2011) did (Figure 19.2).

Word clouds such as these have become commonplace in content such as advertising, posters, and presentations, which is to say that representations of data derived from analytic processes of digital texts have become normalized, they are not the preserve of an obscure branch of the humanities or computer science. Word clouds are especially conducive to wider audiences because they are relatively simple and intuitive – the bigger the word, the more frequently it occurs.[7] However, word clouds are usually static or very limited in their interactivity (animation for layout, hovering and clicking on terms). They provide a snapshot, but do not allow exploration and experimentation.

We have also witnessed in the past years an increase in the number of more complex text-oriented visualizations in mainstream media on the web. The *New York Times* in particular has produced several rich interactive visualizations of digital texts, including an interface for exploring American State of the Union addresses, shown in Figure 19.3.

It is worth drawing attention to several aspects of this interface:

1. The explanatory caption provides succinct context for the visualization and explicitly invites the reader to *analyze* the texts (a much more participatory activity than conventional newspaper reading).
2. The interface provides open-ended search capabilities.
3. It also provides suggested terms to explore.
4. There is a visual representation of the entire corpus – seven State of the Union addresses in what Ruecker *et al*. call a "rich prospect view" (2011) – with the distribution of term occurrences clearly shown.
5. For each occurrence of a term of interest, the surrounding text (context) can be displayed.
6. The frequency of terms can be compared, not only of the same term across multiple years, but also multiple terms.
7. There is a link to the entire 2007 State of the Union address.

With such rich and sophisticated analytic environments, do we even need to read texts anymore? Our reaction to this question reveals much about our purposes for interacting with texts. If we read text for pleasure – a compelling story, a nuanced description, a detailed account of an historical event, etc. – text analysis and visualization are unlikely to be satisfying in the same ways. If we are interested in examining linguistic or semantic features of text, analytic tools may be of help. In our (the authors') own practice as digital humanists, we have tended to combine these activities: we read

(a)

(b)

Figure 19.2 *Wordle* word cloud visualizations of vocabulary from commercials for (a) toys targeted at boys and (b) toys targeted at girls.

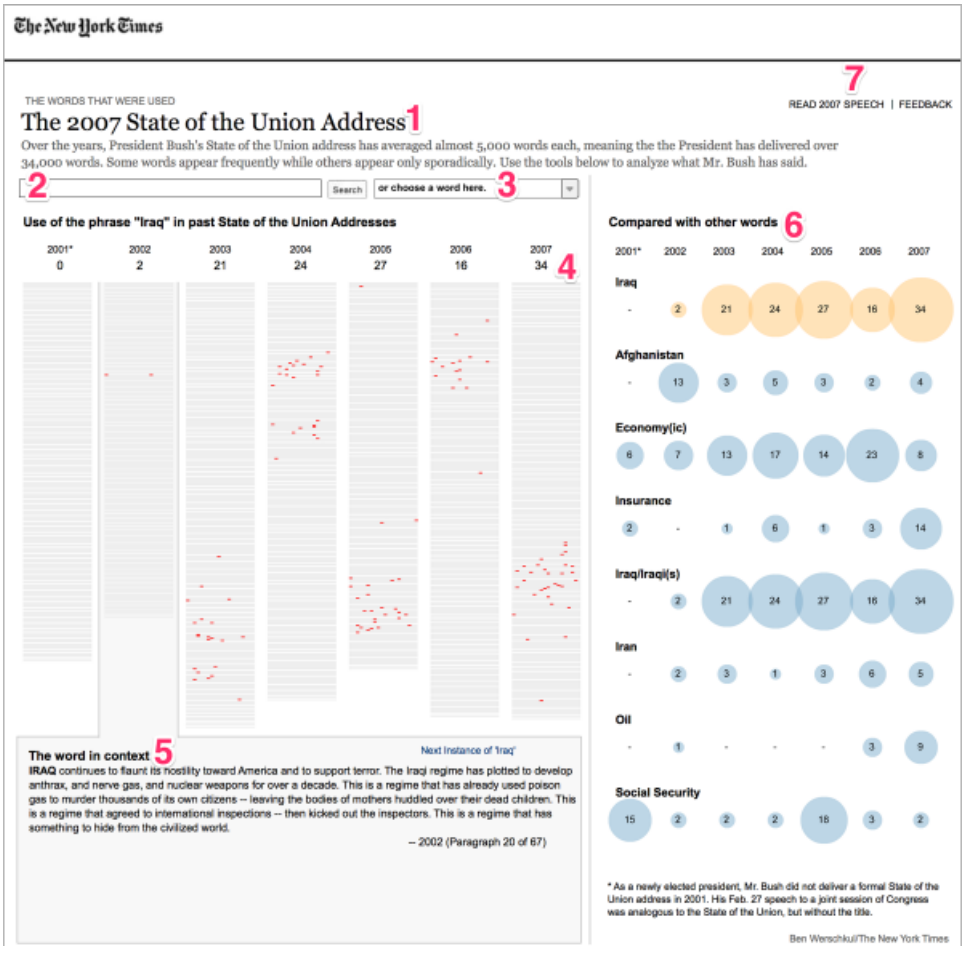

Figure 19.3 2007 State of the Union Address: an interactive text analysis and visualization interface from the *New York Times*.

texts we enjoy, we then explore and study them with analytic tools and visualization interfaces, which then brings us back to rereading the texts differently. This is what we call the *agile interpretive cycle*.

In the rest of this chapter we will explore this circling between reading, analysis, and visualization in more detail, but first we will have a closer look at what is a text.

What is a Text for Analysis?

The availability and prevalence of analytic tools and interactive visualizations can easily lead us to begin experimenting without a proper grasp of the nature and diversity of digital texts. For some purposes this naïveté is acceptable, but using tools effectively and creatively usually entails a full understanding of the materials used. Moreover, the history of digital humanities is as much about a rich tradition of reimagining text as

it is about algorithmic analysis – McGann's *Radiant Textuality* (2001) provides one of the most notable examples.[8]

Bits and Bytes

Digital text is fundamentally a sequence of characters in a string, which is to say it is composed of tiny bits of discrete information that are encoded with a chosen character set in a sequence. Typically we treat textual information at the character-level of granularity, whether it is a character in the Roman alphabet (upper- or lowercase *a* to *z*, an Arabic number (0 to 9), a Chinese ideogram (such as 三 or *sān*, meaning "three"), an Emoji character (like ☺), a control character (like a tab), or any other value from a predefined character set. There are many different character sets, so the crucial thing is consistency – if a text has been encoded with a particular character set, then any future processing of the text must use a compatible character set to avoid problems. This is especially the case for plain text formats where no formatting (and no character-set information) is stored with the text, which is only a sequence of codes from the set.

Unicode is a family of character sets that has helped resolve many issues related to incompatible character sets, but it is far from used universally (Mac OS X uses the incompatible MacRoman character set by default, for instance), and of course there are also huge stores of plain text files that predate Unicode. Character encoding is not an obscure technical issue in text analysis; it remains a common challenge for text analysis and visualization. Unfortunately, there is no reliable way to determine a plain text file's character encoding short of trying different character encoding settings in a text viewer (such as a browser) or plain text editor.[9]

Some character sets are limited to one byte per character, where a byte is composed of eight bits, and one bit is a binary value of 0 or 1. Other character sets (such as Unicode, and in particular UTF-8) can use from one to four bytes to represent a character. In other words, a single Unicode UTF-8 character may actually be represented by a cohesive sequence of up to 32 digits (0s and 1s). The character is typically the smallest unit of information with digital texts, but it is an atom composed of even smaller particles (and tools can misguidedly split an atom apart when character encoding mistakes are made).

Still, the magic of digital texts is that they are composed of discrete units of information – such as the character unit – that can be infinitely reorganized and rearranged on algorithmic whims. Extract the first 100 characters of a text? Sure. Reverse the order of characters in a text? OK. Isolate each occurrence of the character sequence "love"? Done. Digital text is conducive to manipulation – it invites us to experiment with its form in applied ways that print text cannot support. This is the essence of what Ramsay calls *algorithmic criticism*, made possible by the low-level character encoding of digital texts.

Format and Markup

Whereas plain text files only contain the characters of a text, other formats can also express information about character encoding, styling, and layout (on screen or in print), metadata (such as creator and title), and a variety of other attributes *about* the

text. Some file formats use a markup strategy to essentially annotate parts or the entirety of a text. Compare the different ways these markup languages indicate that the word "important" should be presented in bold:[10]

Rich Text Format (RTF)	This is {\b important}.
LaTeX	This is \textbf{important}.
HyperText Markup Language (HTML)	This is important.
Markdown	This is *important*.

It is worth noting that each of these formats can be readily edited with plain text editors, because the markup language itself uses a simple set of characters. Many other file formats are not editable in plain text editors, often because they are stored in a binary format (such as MS Word, OpenDocument, or PDF). Whether a file is editable in plain text or encoded in binary is independent of whether it is a proprietary (closed) format or an open standard. EPUB, for instance, is an open e-book standard that is distributed in binary form (as a compressed file) where much of the content is typically encoded in an HTML format. With concern for preservation and access, and deep roots in library culture, digital humanists have long favored human-readable (not binary) and open formats.

One of the crown jewels of the digital humanities community is the Text Encoding Initiative (TEI), a collective project founded in the 1980s to standardize markup for digital texts in a human-readable and open format.[11] Just as consistency and compatibility are crucial for character encoding, the same is true for other types of markup: how to encode a paragraph or a person mentioned in a text, for instance.

Although the TEI has traditionally been more focused on detailed encoding for preservation, there are definitely analytic benefits to the markup. Imagine we wanted to examine the term "lady" in Shakespeare's *Macbeth*. In a plain text file each character name is indicated before the speech, which means that a frequency count of the word "lady" might also misleadingly include "Lady Macbeth" the character name. With TEI, the character name is marked up with the <speaker> element, which makes it easier to reliably to filter out those occurrences. Conversely, we may want to only consider speeches by Lady Macbeth – again, a relatively trivial transformation of the text. Digital texts are infinitely reorganizable, and markup (such as TEI) serves to proliferate the number of logical moves that can be made, like extra grips on a climbing wall.

Despite all this, one of the first operations performed on a painstakingly marked-up text is often to strip out the markup. This is partly because many analytic operations do not benefit from the markup (indeed the markup can interfere with the proper functioning of the tool) and partly because there is still a dearth of tools that truly allow the markup to be exploited.[12]

Shapes and Sizes

Texts and text collections come in different formats, but also have different shapes and sizes, which also help determine what is possible and what is optimal.

A corpus is a *body* of texts (though a corpus can have only a single text). The kinds of text analysis operations that can or should be performed will of course be determined in part by the compatibility between what we call the *geometry* of the corpus and the design of the tools. One size does not fit all. A tool like *Poem Viewer* (Figure 19.4; ovii.oerc.ox.ac.uk/PoemVis) is intended primarily to assist in close reading of single poems, whereas the Google *Ngram Viewer* (Figure 19.5; books.google.com/ngrams) is intended to enable queries of millions of books (but no reading of text). These represent very different kinds of intellectual work, determined in part by the nature of the corpora.

Just as bits of a single digital text can be rearranged, texts within a digital corpus can be rearranged and sampled for a variety of purposes. Imagine a collection of articles from philosophy journals from the past 150 years[13] – this is a coherent corpus, but one that can spawn any number of other corpora based on a variety of logics for ordering, grouping, and filtering. For instance, we might want to have all documents ordered by year of publication and then author name, or by journal and then year and then author. Similarly, we might want to create new, aggregate texts that combine all articles by decade or by philosophical period. Or perhaps we just want to work with articles

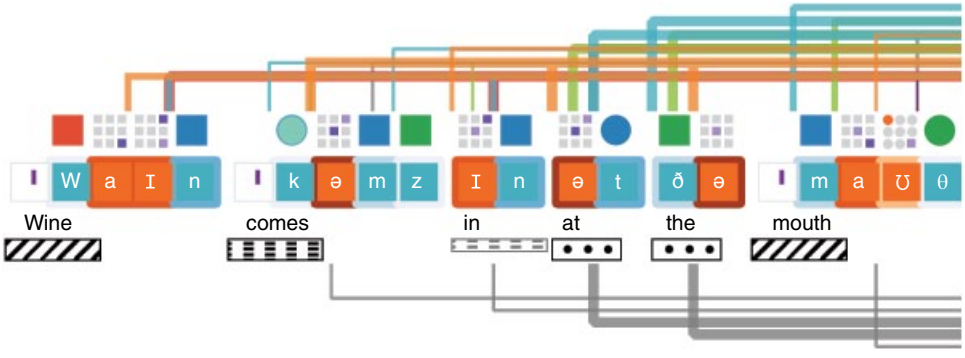

Figure 19.4 *Poem Viewer*, for close reading of linguistic features in poetry.

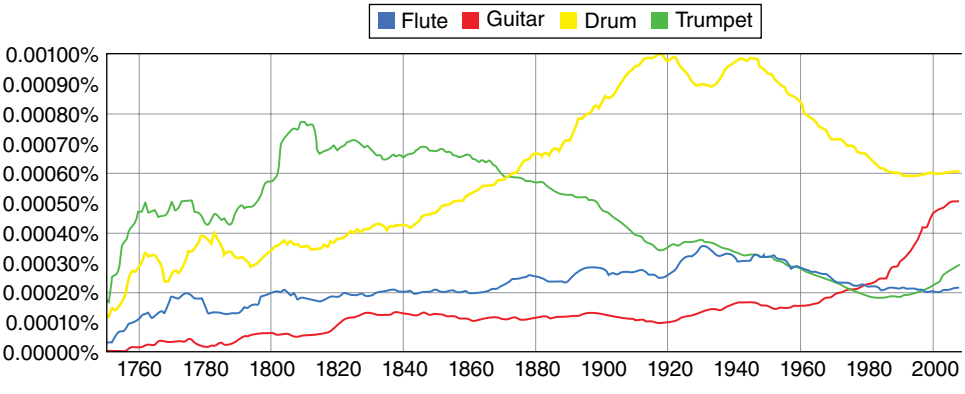

Figure 19.5 Google *Ngram Viewer*, which allows querying on millions of books.

published outside of Anglophone countries. In addition to corpus decomposition and reorganization, there are cases where a single text can generate a new corpus with many texts: all speeches from each speaker in a play in separate documents, for instance, or each item in an RSS feed becomes its own document.[14] A digital corpus is a bit like a bag of Lego where pieces can be built up in various configurations, but it is even better than that, since digital texts are trivial to clone and documents can exist in several structures a once (an infinite bag of Lego).

The presence of markup and of metadata is crucial for this kind of flexible and dynamic creation of corpora. Since the structuring and reorganization steps are often specific to the local research context (the available corpus and its format, the tools at-hand, the types of questions to ask, etc.), we have found that a bit of programming competency for parsing and processing document sets is valuable.

Analysis and Reading

In all these applications, the appeal to computers as an aid to processing texts can be largely summarized by two types of questions:

1. For texts with which I am already familiar, how can computers help me identify and study interesting things I had not noticed before, or things I had noticed but did not have reasonable means to pursue? Digital texts enable a proliferation of representations to explore linguistic and semantic characteristics and produce new representations and new associations, all of which can help to solidify intuitions we may already have had or generate entirely new perspectives.
2. How can computers help me identify and understand texts with which I am not familiar or which I cannot reasonably read? Human reading is time-consuming and selective, and retention of content is idiosyncratic. Computers can help extend human reading and understanding, especially for large collections of texts that you couldn't read in a lifetime. Computers can help identify what you might want to read.[15]

Of course, you have been doing text analysis all along. Readers on the web have become accustomed to embedded interactive analytics, like the *Deadspin* example we started with. We routinely use Find tools to search documents or web sites. It is common to see interactive word clouds in a blog that show you the high frequency words used in that blog at a glance. *Wordle* word clouds, like those shown in Figure 19.2, have become a common design feature for posters about digital humanities events. Newspapers like *The Guardian* have special data journalism units that specialize in gathering datasets and creating interactive widgets for readers to explore.[16] The question is, How we can use similar methods to study and represent historical documents, philosophy texts, or literatures?[17] To understand what we can do we need to return to strings.

The computer has a fundamentally different understanding, if we can call it that, of a text than we do. The computer "reads" (processes) a text as a meaningless string of characters. What it can do is operate on this string of characters, and it can reliably do very repetitive operations. For example, a computer can compare a short string like a

word to every position in a much longer string, like a novel. That is how searching works. The computer checks every word against what you want to find. It does this menial work quickly and reliably.

The computer can do more than just find words. The computer can find more complex patterns. Let's say you want to find either "woman" or "women" – the computer can be given a pattern in the form of a regular expression, "wom[ae]n."[18] Or you can do a truncation search that searches for any words that begin with "under" – "underwater," "understand," and so on. The regular expression for this, depending on the system, might look like "under.*" – where the "." means any character and the "*" means any number (of any character). Library database systems will typically assume that you want variants of your word, especially the plural with "s" on the end. One can, in fact, do a lot of text analysis just with regular expressions that describe the patterns you want to find and return the passages that match.[19]

But what is a word? We tend to think of a word as a unit of meaning that often has an analog in the real world. The word "cat" in "the cat is on the mat over there" refers to that furry thing I'm pointing at. A computer doesn't know what a word is and certainly has no sense of what words might refer to.[20] For a computer to handle words you need to define what the orthographic (written) word is in a string, and we typically do that by identifying the characters that demarcate a word. Words are usually bounded by spaces and punctuation, and a computer can be told to split a long string (text) into shorter strings (words) by looking for the demarcation characters – though this splitting up into words, a process called tokenization, is highly challenging in some languages that do not have characters to indicate word boundaries, such as Japanese and Thai. The rules for splitting a text into word tokens can get complex, and these rules vary from language to language, but this splitting or tokenization is a basic first step to text analysis since words are important to us, particularly since so many tools operate on the lexical (word) level, rather than other units such as phrases. Tokenization, it should be noted, is not a quantitative operation – it is a phase of text analysis that has to do essentially with symbolic processing and recognition of patterns, with some similarities to how humans read.

This brings us back to analysis, which etymologically means a breaking apart into smaller units. Text analysis, like any form of analysis, is a process of decomposition, and as such is a standard way of understanding something. When we try to understand any complex phenomenon, one way to start is to break it into smaller parts – ideally into atomic parts. Bodies can be understood in terms of organs and then cells. Histories can be understood in terms of epochs and events. Texts can be understood in terms of chapters, paragraphs, sentences, and finally words (even if meaning spans across these units). Where we can formally define these parts, the computer can help us decompose the text.

What then do we do with a text in tiny little parts? Well, we can build indexes for the end of the book or concordances that show each word in a line of context. Concordancing was in fact one of the original uses for computers in the humanities, as it is what Father Busa wanted IBM support for in the late 1940s (Hockey, 2004). Concordances, especially of the Bible, are tools with a history that goes back to the thirteenth century. They allow you to quickly scan all the instances of a word such as "love" in an important text. They are better than an index, which just tells you on

```
moon (29)
I.1/577.1  four happy days bring in | Another moon: but, 0, methinks, how
I.1/577.1    0, methinks, how slow | This old moon wanes! she lingers my
I.1/577.1       away the time; | And then the moon, like to a silver bow |
I.1/577.2   faint hymns to the cold fruitless moon. | Thrice-blessed they
I.1/577.2      to pause; and, by the nest new moon-- | The sealing-day
```

Figure 19.6 Key Word In Context (KWIC) of "moon" in *A Midsummer's Night Dream* from TACTWeb.

what pages you can find the word, because the lines of text containing the word that represents the concept of interest are arranged to make it easier for one to see patterns in the appearance of the word.

Searching for words and presenting them on the screen has evolved from the print concordance into very large search engines like Google. Computers can arrange the passages of text with the word concorded in different ways, like the Key Word in Context (KWIC), where the key word (e.g., "moon") lines up so you can see what words come before and after (Figure 19.6). Until personal computers and then the Web came along and there were easy ways of publishing electronic texts directly for the computer screen, batch concording tools such as COCOA and OCP were used to create large print concordances. Text analysis, up until the first interactive tools like ARRAS, was more a matter of taking apart a text and then rearranging it so that you could print the rearrangement. It was the print concordance that was then used as a study tool.

Another use of text analysis was to identify patterns of word usage by particular authors, a field called stylistics. Not only can computers find patterns, but they can count patterns and compare counts. By counting function words, which do not convey a lot of semantic content, but which are important syntactically (and which occur in greater numbers, making them more statistically significant), one can get a sense of an author's writing style. Writing style, once formally described, can then be measured and compared (Kenny, 1982), and you can even use it as one more tool in trying to identify anonymous authors like the Unabomber (Foster, 2000).

Text analysis is not just analysis, it is also synthesis. Text analysis tools such as concordances not only break apart a text, but they put it back together in new ways. These new ways range from KWICs to visualizations that are increasingly abstract representations of the text. Text analysis synthesizes a new text, like stitching Frankenstein's monster out of parts, and it allows you to study the original in a new light. It is the textual equivalent of sampling and synthesizing new musical works, or making a collage out of images cut up from elsewhere. This synthesis can be done for artistic purposes or it can be done for interpretative purposes. The emphasis on creativity and experimentation align well with contemporary maker culture and its core tenet that *doing* (constructive creation) fosters learning and discovery. Thinking through choices of how and what to create, as well as observing and critiquing what is created, can provide generative moments of insight. Moreover, the mere ability (or affordance) to perform actions on texts can be empowering for readers and serves to further unseat the notion of rigid, canonical texts (if any such notions remain after the rise of electronic literature and hypertext).

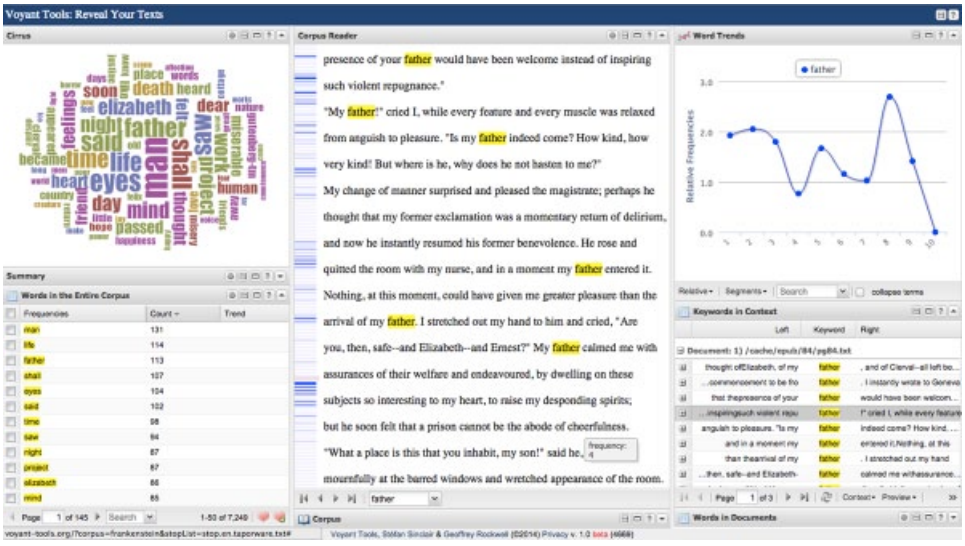

Figure 19.7 The *Voyant Tools* standard reading skin, showing Mary Shelley's *Frankenstein* for analysis.

We use the word *hermeneutica* for the interactive and interpretative analytical tools that facilitate the rearranging and manipulation of texts in order to better study and understand them. *Voyant Tools* (Figure 19.7), for example, lets you click on a word in the cloud (Cirrus) and then see the distribution of the word over the text (Word Trends). Clicking on the histogram shows the keyword in context, and clicking on an instance in the Keywords in Context panel jumps the full-text Corpus Reader to the right location. Each panel shows a different view on the text which can be used to control other views. Be careful, however, that you don't depend only on the stitch-ups. They are semi-automated rearrangements that should be questioned just like any other interpretation. Their very existence depends on a wide range of human choices, from the encoding of the digital text and the programming of the analytic tool to the parameters selected by the user and ways that results are read. Text analysis and visualization data are taken, not given, as Johanna Drucker reminds us, in her poetics of computer-mediated humanistic inquiry (2011).

Analysis and Visualization

Both print and digital text is represented visually for reading, and typography is about the graphical representation of characters in a particular medium.[21] In this simple sense, text is already a type of visualization, an instantiation of a more notional text that is not concerned with specificities like page numbers or scrolling position.[22] Emphasizing displayed text as visualization has the benefit of allowing us to take into account a full spectrum of text visualizations. Consider a text with only slight stylistic changes, such as having all adjectives display in green. Is this a text or a visualization? It is both.

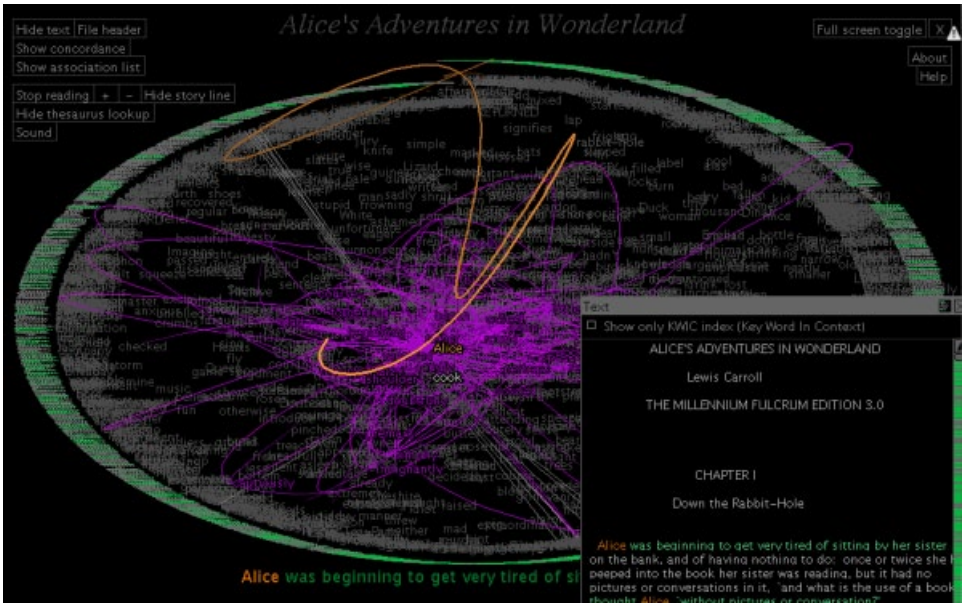

Figure 19.8 *TextArc* by Bradford Paley, showing *Alice in Wonderland* as text around the perimeter and as distributed terms within the perimeter.

We can iteratively add additional variations to the text rendering to change other stylistic attributes (italics, size, orientation, etc.) and even begin unhinging words or other lexical units from their original sequential position. A classic example of rich text visualization is Bradford Paley's *TextArc* (textarc.org), where words from a text are actually displayed twice, once in linear order arranged around the perimeter clockwise from the top (hovering over the tiny representation of a line causes a more legible version to appear), and then again by plotting each content word within the circle as if each occurrence in the perimeter pulled the terms toward it gravitationally (also called a centroid). As a result, the location of the word conveys information about its distribution in the document – "king" and "queen" occur more in the last third of *Alice in Wonderland*, for instance (Figure 19.8).

The spectrum of text visualizations thus includes a variety of stylistic and positional transformations, but also more abstract representations of textual attributes. One example of this is the *Knots* interface in *Voyant Tools*, which represents lexical repetition by introducing a kink in a line every time a selected term occurs. The more "knotted" a line, the greater the repetition (Figure 19.9).

Even though *Knots* is a more abstract and qualitative expression of repetition, it is only possible because of underlying data and algorithmic operations. The apparent dichotomy between the quantitative and qualitative can be misleading, particularly since text visualizations depend on a symbiosis between them.

Text visualizations can use a very wide spectrum of graphical features, from subtle typographical attributes in a sequential text to complex geometric forms produced from textually derived data. Reading practices are equally expansive: we read text to understand or experience something, and the same can be said about reading data visualizations.

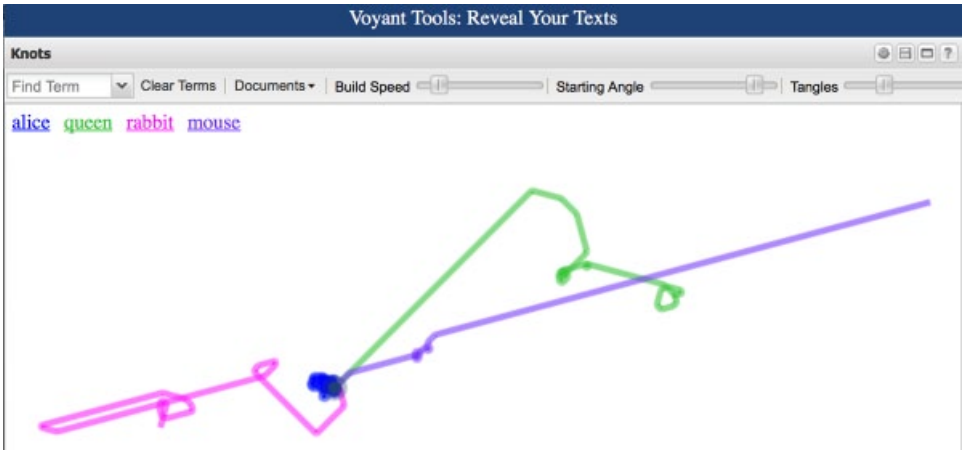

Figure 19.9 *Knots* visualization of *Alice in Wonderland* – some terms repeat often and regularly (e.g., "Alice," near the middle) while others occur very locally (e.g., "mouse," which shoots off to the right).

There is an important distinction between how to read a text visualization and how to interpret what is being visualized: understanding the mechanics of consumption compared to the understanding of what is being consumed. Once we have learned to read text in a language, we should be able to read most texts in that language, though the text may not always make sense to us. The same cannot be said for all text visualizations – we know when we are looking at text, but with some text visualizations we may be led to ask "what are we looking at?" Visualizations make use of a visual grammar, just as language requires a linguistic grammar, and we need to be able to parse what we see before attempting to analyze and understand it (see Tufte, 2001, for foundational work in studying visual information in graphs). We have developed common visual literacies for representations such as simple charts, maps, and timelines, but other representations (like *TextArc* and *Knots*) are likely to require explanation. The effectiveness of a visualization will depend in a first instance on the ability of the reader to decipher what is being seen, either because of familiarity with the visual paradigm or through a willingness to become familiar with it. One way we often make sense of the visual features is to play with the parameters or interactive controls, which is why interactive visualizations can be easier to understand. With interactives, the play becomes a way of understanding the rearranged text, but also the tool as text.

Making Meaning Count

It would be convenient if there were a reliable set of text visualizations that were guaranteed to produce new insights, but interpretation is never that formulaic (thankfully). Sometimes the relative simplicity and sparseness of a word cloud is useful to get an overview of a text, at other times a simulated 3D representation of term clusters in a scatter-plot graph showing correspondence analysis results is an effective way of studying a corpus.[23]

We have found two principles to be important when engaging with text anlaysis and visualization tools – they may seem obvious, but they are worth stating:

- *Don't expect much from the tools*. Most tools at our disposal have weak or nonexistent semantic capabilities; they count, compare, track, and represent words, but they do not produce meaning – we do. When you don't expect much from tools, it shifts the interpretative responsibility for making sense of the rich variety of ways that texts can be represented.
- *Try things out*. Taken individually each tool may not do much, but accumulating perspectives from many tools can be beneficial. One tool may help you notice something that is worth exploring in more detail with another tool. Within each tool there may be settings that are worth tinkering or playing with for different effects (Sinclair, 2003). We use tools not to get results but to generate questions, so the more things we try, the more questions we're likely to have. Ramsay (2014) calls this the screwmeneutical imperative.

These two principles are expressed in part in the *Voyant Tools* environment that we have developed: the individual tools are designed to be simple and modular in order to favor interaction *with* and *between* the tools. The tools are intended to facilitate the augmented hermeneutic cycle by enabling navigation between reading text, analysis, and visualization at various scales ("differential reading" that slides between close and distant reading practices – see Clement, 2013).

Voyant Tools has the benefit of being readily accessible on the web and relatively user-friendly, but there are many other tools and interfaces that are worth exploring. For text analysis and visualization from a digital humanities perspective we suggest exploring resources listed on the *Text Analysis Portal for Research* (tapor.ca) and the text-mining section of DiRT (bit.ly/1sRGAuI).

The idea that text analysis and visualization are interpretative practices may seem paradoxical at first glance, since the digital is founded on matching and counting, but no amount of counting can produce meaning. On the other hand, digital tools do facilitate experimentation with the representation of digital texts, and those representations can lead us, as readers, to observe noteworthy phenomena and connections, some of which, we may argue, are meaningful. Sometimes we also get interested in the interpretation of these tools of interpretation, but that is another type of text analysis.

Notes

1 This quotation is from Kristeva's "Texte clos" (1968). We have added emphasis to highlight Kristeva's poststructuralist move to conceptually equate text with culture (everything is text). Here is an English translation: "One of the problems for semiotics is … to define the specificity of different textual arrangements by situating them in the general text (culture) of which they are a part and which, in turn, is part of them" (Kristeva, 1980:83).

2 *Voyant Tools* is a suite of text analysis tools we developed for the web. You can try them at http://voyant-tools.org.

3 This chapter is based on *Hermeneutica*, a forthcoming book on text analysis. See http://hermeneuti.ca.

4 The scale of the numbers is more significant here than exact values, which are notoriously difficult to determine. The estimate for emails comes from a widely cited report from the Radicati Group (2014), and Google search numbers are estimated from Google's own documentation and comCore statistics (http://bit.ly/1s3deqZ).

5 These examples are intended to be generic and representative but are inspired by specific examples such as (1) a high-school student doing text analysis on *the Game of Thrones* (bit.ly/1m6H9if); (2) an independent analyst parsing Canadian security documents leaked by Edward Snowden (bit.ly/1iyAWpC); (3) a car company like Kia tracking the response to a new model of vehicle (buswk.co/1mIsf4i); (4) a historian studying immigration using 200,000 documents from the Proceedings of the Old Bailey in London (bit.ly/1satlmL).

6 It would take well over 100 years to read just the 45,000 e-books in Project Gutenberg (gutenberg.org), assuming one could sustain the unlikely pace of one e-book a day.

7 Other aspects of word clouds may appear intuitive but are not – typically the position of words has little meaning, for instance. Word clouds have detractors who justifiably argue that they are often misused (when other visualizations would be more appropriate), insufficiently contextualized and reductive, and informationally misleading (like the color of words in some instances); see for instance Harris (2011).

8 See also Ryan Cordell's excellent "On ignoring encoding" (2014), which attempts to recalibrate the disproportionate attention paid to text analysis compared to digital editing and encoding practices.

9 Some text editors (like JEdit) and analytic tools (like *Voyant*) have built-in heuristics to try to guess character encoding, but in most instances it remains a guess, and it is best to specify the character encoding if it is known.

10 LaTeX may be the least familiar format presented here, but it is widely used as a document preparation format for scientific publications.

11 See Renear (2004) and Hockey (2000) for more information on the TEI.

12 One notable exception is TMX (textometrie.ens-lyon.fr).

13 We are beginning work on a corpus of philosophical texts from the past 150 years, provided by JSTOR.

14 RSS is Really Simple Syndication, an XML-based format that allows for multiple items (like news articles or blog posts) to be included in a single document.

15 Franco Moretti (2005) downplays reading in his description of distant reading, but we don't buy it: Moretti is still very much in the business of reading and interpretation.

16 See http://www.theguardian.com/media/datablog/2012/mar/07/open-data-journalism for an entry point into their Datastore and Datablog.

17 For an exploration of text analysis for teaching, see Sinclair and Rockwell (2012).

18 We are focusing on simple searching here, but of course it is also possible to have computers perform morphological analysis to find word variants that belong to the same family.

19 For more on regular expressions, see Stephen Ramsay's classic "Using regular expressions" (http://solaris-8.tripod.com/regexp.pdf). Ramsay also treats of patterns in *Reading Machines* (2011.)

20 Some of the challenges of natural language processing from the last half-century can be summarized by the difference in semiotic models between humans and computers: for humans, language refers to concepts that are learned through experience; for computers, language is a formal representation of lower-level binary data.

21 Braille for the visually impaired is an exception, because characters are represented for tactile rather than visual sensing.

22 The claim that there is a notional text prior to any printed or displayed instantiation will seem contentious to some, but we are especially interested in emphasizing that any form a text takes is already laden with visual specificities (font face, size, and color, page layout, etc.) that are bound to influence the experience of reading text.

23 A correspondence analysis graph for the archives of the Humanist Discussion Group listserv is a useful way to study shifts in concerns over time of the digital humanities community: see bit.ly/1ljh2BT, as well as Wang & Inaba (2009).

References and Further Reading

Clement, T. 2013. Text analysis, data mining, and visualizations in literary scholarship. In *Literary Studies in the Digital Age: A Methodological Primer*, ed. K. Price and R Siemens. New York: MLA Commons.

Cordell, R. 2014. On ignoring encoding. http://ryancordell.org/research/dh/on-ignoring-encoding (accessed June 20, 2015).

Drucker, J. 2011. Humanities approaches to graphical display. *DHQ: Digital Humanities Quarterly* 5 (1).

Fischer-Baum, R., Gordon, A., and Halsley, B. 2014. Which words are used to describe white and black NFL prospects? *Deadspin*. http://deadsp.in/1iNz1NY (accessed June 20, 2015).

Foster, D. 2000. *Author Unknown: On the Trail of Anonymous*. New York: Henry Holt and Company.

Harris, J. 2011. Word clouds considered harmful. *Nieman Journalism Lab*. http://bit.ly/QKNMdD (accessed June 20, 2015).

Hockey, S. 2000. *Electronic Texts in the Humanities*. Oxford: Oxford University Press.

Hockey, S. 2004. The history of humanities computing. In *A Companion to Digital Humanities*, ed. S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion (accessed June 20, 2015).

Jockers, M. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Kenny, A. 1982. *The Computation of Style*. Oxford: Pergamon Press.

Kristeva, J. 1968. Le texte clos. *Langages* 3 (12).

Kristeva, J. 1980. *Desire in Language: A Semiotic Approach to Literature and Art*. Trans. T. Gora, A. Jardine, and L.S. Roudiez. New York: Columbia University Press.

McGann, J. 2001. *Radiant Textuality: Literature After the World Wide Web*. New York: Palgrave.

Michel, J.-B., Shen, Y.K., Aiden, A.P., *et al*. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014), 176–82.

Moretti, F. 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. London: Verso.

Radicati Group. 2014. Email statistics report, 2014–2018. http://bit.ly/1o6GmQA (accessed June 20, 2015).

Ramsay, S. 2011. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

Ramsay, S. 2014. The hermeneutics of screwing around; or what you do with a million books. In *Pastplay: Teaching and Learning History with Technology*, ed. K. Kee. Ann Arbor: University of Michigan Press.

Renear, A. 2004. Text encoding. In *A Companion to Digital Humanities*, ed. S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Blackwell. http://www.digitalhumanities.org/companion (accessed June 20, 2015).

Ruecker, S., Radzikowska, M., and Sinclair, S. 2011. Visual Interface Design for Cultural Heritage: a Guide to Rich-Prospect Browsing. Burlington, VT: Ashgate.

Sinclair, S. 2003. Computer-assisted reading: reconceiving text analysis. *Literary and Linguistic Computing* 18 (2).

Sinclair, S., and Rockwell, G. 2012. Teaching computer-assisted text analysis: approaches to learning new methodologies. In *Digital Humanities Pedagogy: Practices, Principles and Politics*, ed. B.D. Hirsch. Cambridge: OpenBook. http://www.openbookpublishers.com/htmlreader/DHP/chap10.html#ch10 (accessed June 20, 2015).

Sinclair, S., and Rockwell, G. 2014. *Voyant Tools*. http://voyant-tools.org (accessed June 20, 2015).

Sinclair, S., Ruecker, S., and Radzikowska, M. 2013. Information visualization for humanities scholars. In *Literary Studies in the Digital Age: A Methodological Primer*, ed. K. Price and R Siemens. New York: MLA Commons.

Smith, C. 2011. Word cloud: how toy ad vocabulary reinforces gender stereotypes. *The Achilles Effect*. http://bit.ly/1osjjji (accessed June 20, 2015).

Tufte, E. 2001. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Wang, X., and Inaba, M. 2009. Analyzing structures and evolution of digital humanities based on correspondence analysis and co-word analysis. *Art Research* 9. http://bit.ly/1jLWnbX (accessed June 20, 2015).