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.

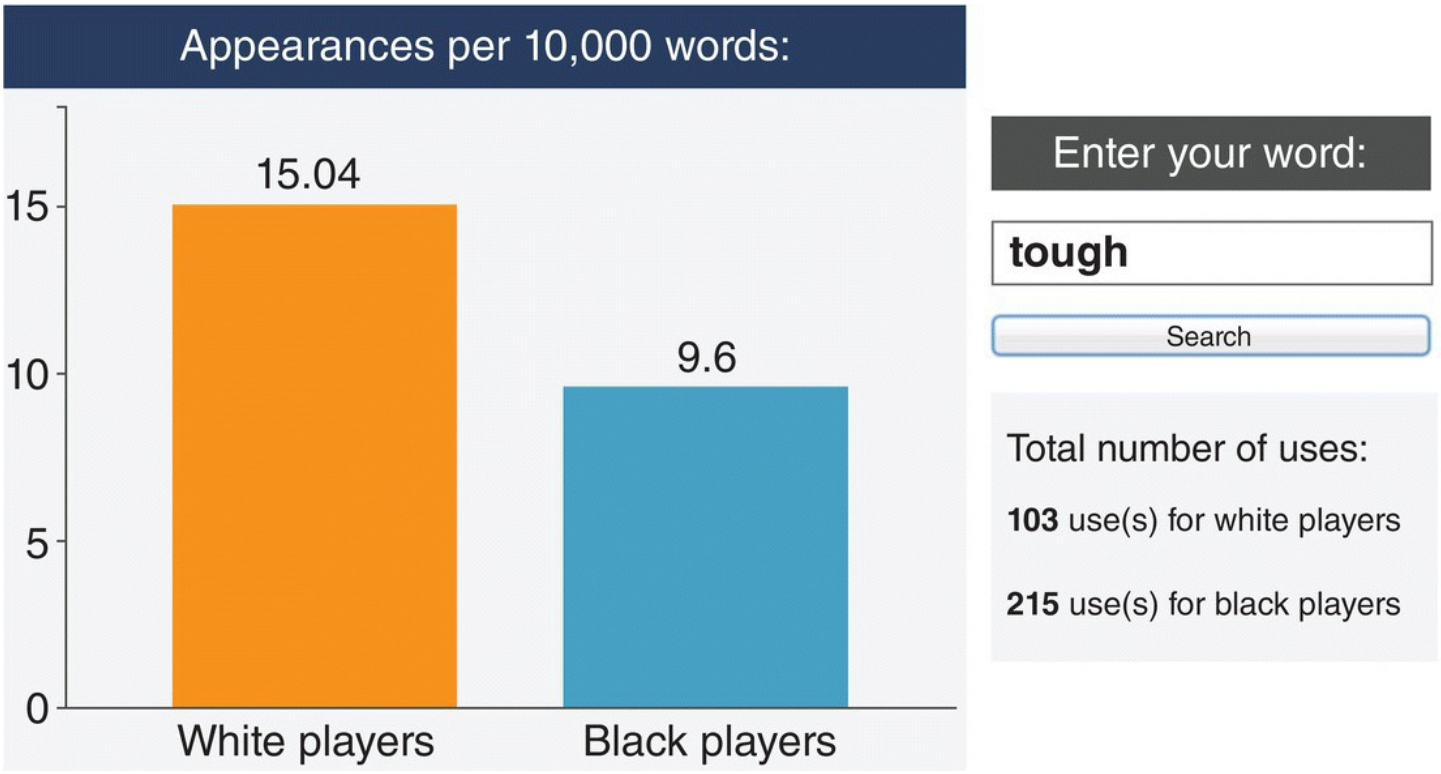

Figure 19.1 An interactive text analysis and visualization widget by *Deadspin*.

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 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).

(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.

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.

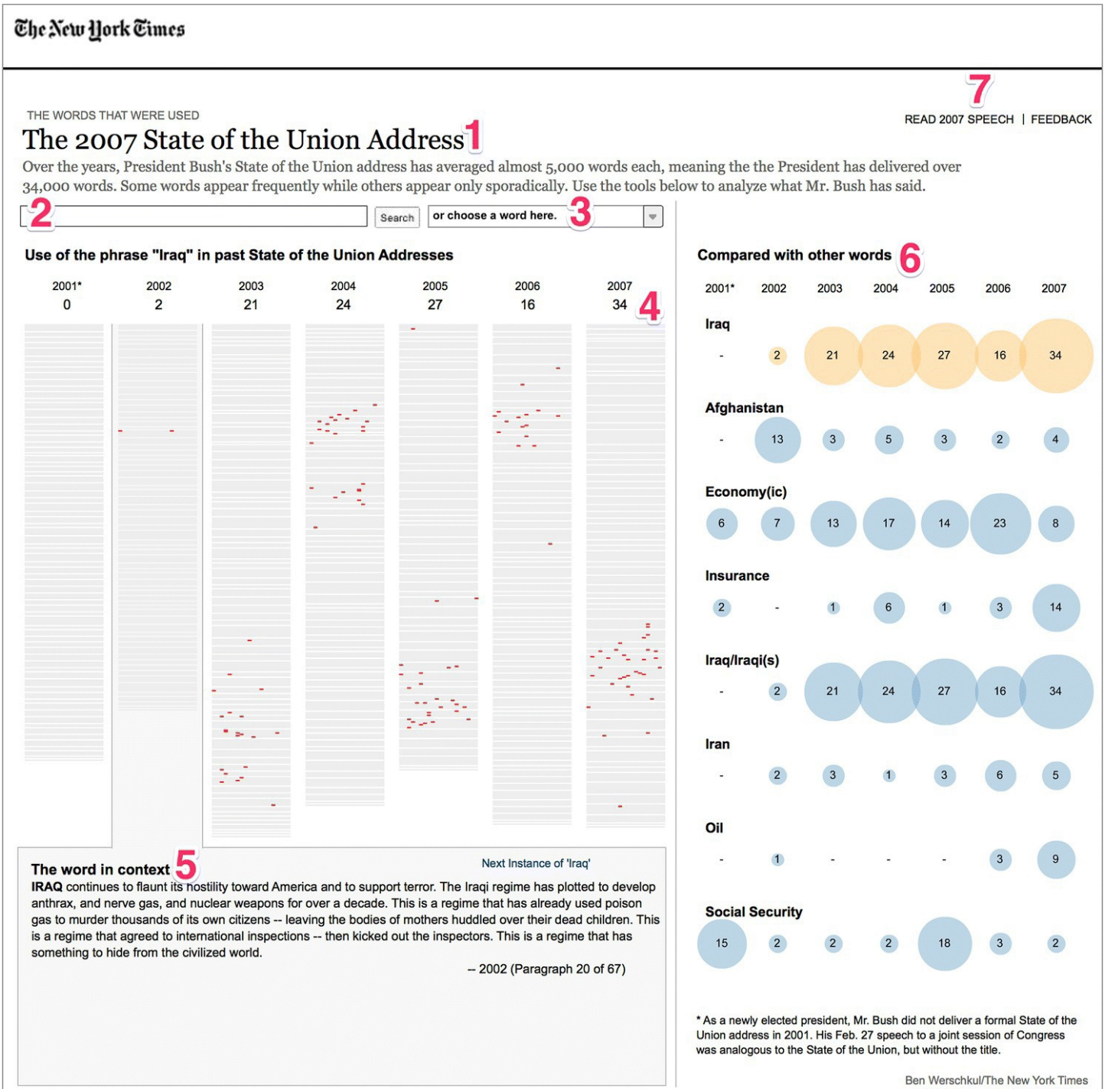

Figure 19.3 2007 State of the Union Address: an interactive text analysis and visualization interface from the *New York Times*.

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 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 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.

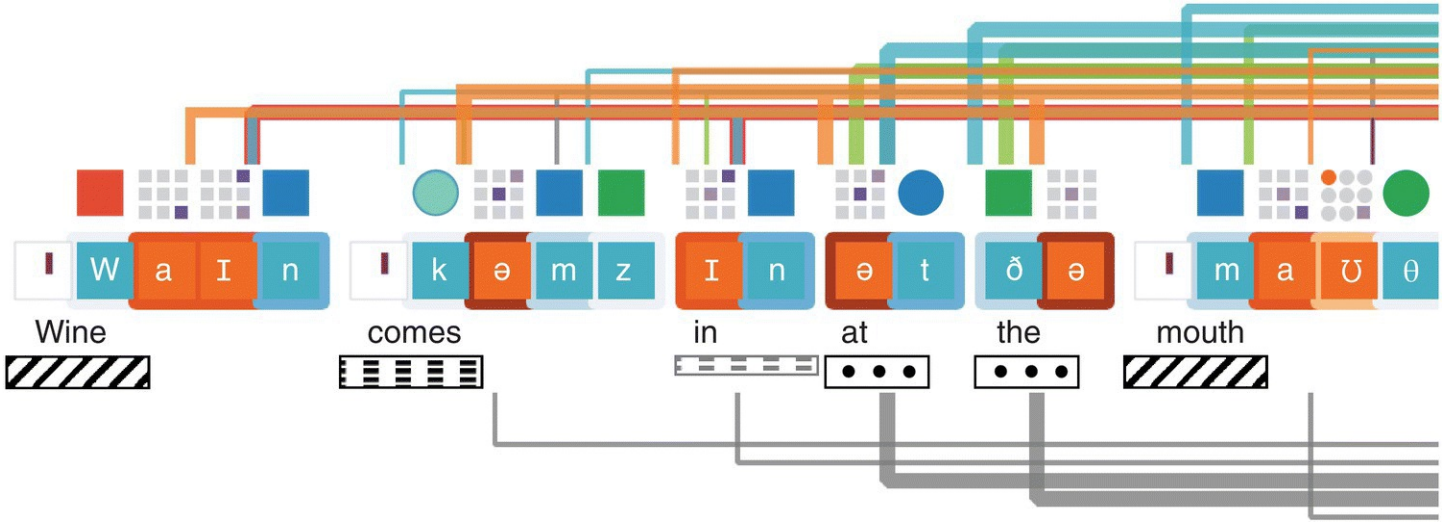

Figure 19.4 *Poem Viewer*, for close reading of linguistic features in poetry.

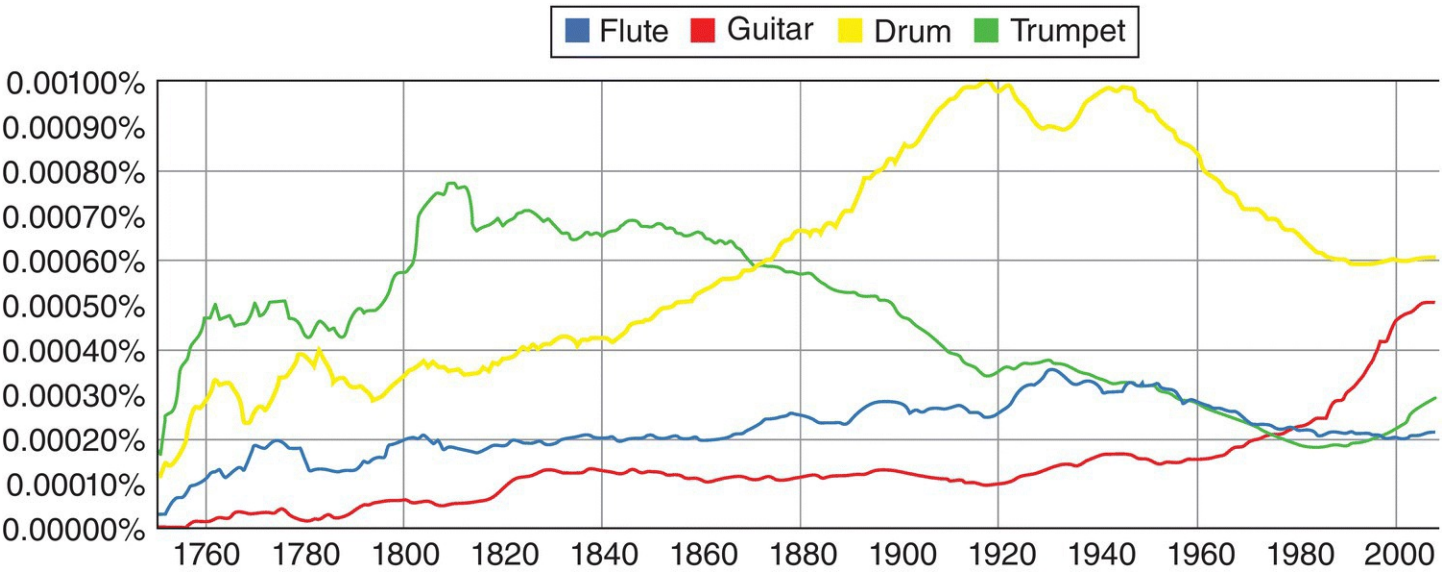

Figure 19.5 Google *Ngram Viewer*, which allows querying on millions of books.

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 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