Identifying Structures in Text Via Visualization

An intriguing question is whether this approach can be used to identify distinguishing features of texts composed within a literate culture and of texts composed within a non-literature culture. Are there rhetorical devices and structures indicative of oral composition, that have perhaps escaped notice previously because they were hidden in the sheer bulk of surrounding material?

(This paper is based in part on work by the authors, presented at the ASOR 2012 conference in Chicago, November 2012.)

By Gordon Rugg
Keele University, UK

And

David Musgrave
Amridge University, USA
May 2013

Declaration of interest: Gordon Rugg is co-inventor of the Search Visualizer software used to produce the illustrations in this article, and is a shareholder in the company that produced the software. All the illustrations in this article were produced using the free, online version of Search Visualizer.

Introduction

Sometimes, what's written between the lines of a document is more significant than what's in plain sight. This article is about ways of spotting those things that are hidden between the lines.

One place to start is with the choice of words. An example is the word "forty" in Homer's Iliad, about the Trojan War. It looks like an ordinary, unremarkable word, until you look at the places where it occurs within the Iliad. When you do that, you discover that it occurs frequently in one short section of the Iliad – but nowhere else.

There's something else that's odd about that section of the Iliad. It's known as the Catalogue of Ships; it lists the Greek ships that came to the war. If you've ever wondered about the description "The face that launched a thousand ships" then this is where it came from; the Catalogue lists a total of 1,186 ships.

The odd thing about this section is its description of where those ships came from. The Iliad as a whole was assembled in the Iron Age, long after the most likely date for the war it describes. However, the Catalogue of Ships refers to kingdoms and countries that had been gone for centuries by the time of Homer. How had those ancient names lasted through the ages?

Most scholars agree that the Catalogue comes from an older poem that had survived until Homer's time, and had then been incorporated into the Iliad, ancient names and all. The word "forty" most likely came as part of that package, perhaps referring to an archaic unit size for a naval contingent, perhaps using an ancient metaphorical sense of "forty" as meaning "an unspecified large number" in the same way that current English "a couple of dozen" is used as an approximate number, rather than exactly "24". The implications are far-reaching. This use of ancient names and numbering shows that information had survived from the Bronze Age down till Homer's time centuries later. This implies that there might be other parts of the Iliad that describe Bronze Age realities, giving historians and archaeologists invaluable insights into a long-gone world.

So, apparently minor choices of wording in a text can show apparent traces of an older past. The same is true of the structure of a text; how the themes in that text are chosen and ordered. These can tell the reader a lot about the world-view of the person who wrote a text, and also about older sources on which they drew, whether consciously or otherwise. In the Old Testament, for instance, there are themes such as the origin of sin, the concepts of clean and unclean, the juxtaposition of life and death, and many others. So, how are these structured?

Ancient authors used a wide range of rhetorical and literary devices that involve significant structuring of text. However, identifying these structures in texts of significant length is difficult, for two main reasons.

One reason is that the structures can be obscured by the sheer volume of surrounding text. Another is the language barrier, which is a particular issue for biblical studies, where several languages are involved.

If the text is short, then one way of identifying structures is to highlight the key words in color. In practice, though, this doesn't work well if the text is more than a page or two in length. You end up having to join the pages together to form a strip looking like a scroll, so that you can see the overall pattern, and if the strip is more than a few pages long, it's too long to hang on an office wall. Also, it's a laborious process, and it gets through a lot of hard copies if you want to try a range of keywords.

A more practical solution, speaking from experience of sticking together large numbers of pages with colored highlighter on them, is to use software. Using the highlight function in ordinary text processing software isn't very effective, because you're limited by the minimum size of the font. However, if you instead use a completely schematic representation, with each word in the text represented by a simple white or colored square, then you can shrink the square sizes down significantly, and thereby fit visualizations of long texts comfortably onto a computer screen. In addition, this approach bypasses the language barrier, since you are looking at patterns of word distributions, rather than trying to understand all the words in a text.

This approach is at the heart of the Search Visualizer software, first produced by Gordon Rugg and Dr Ed de Quincey (now at Greenwich University, UK), which was used for the visualizations in this article. This software is available online for use without charge, at www.searchvisualizer.com. There is a supporting blog, with

examples of using the software for different types of analysis, at searchvisualizer.wordpress.com.

Example: Textual structure in the Code of Hammurapi

A simple initial example of structure within a text is thematic structure, where different sections of a document deal with different themes in turn. As anyone who has ever graded student coursework will testify, not all written work follows this principle of location in a text being used for systematic division of that text by themes. So what happens if we look at very early written texts: do they show systematic thematic organization, or are themes mingled together haphazardly?

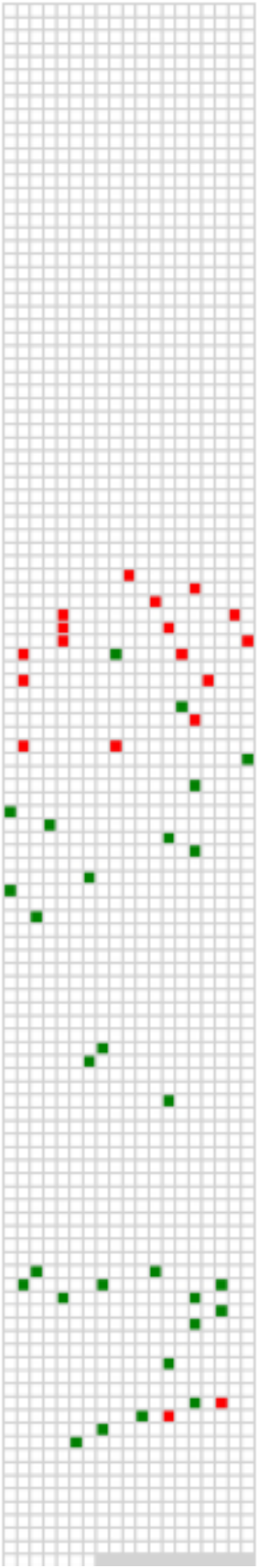
The illustration below shows the locations of specified key words within a short text, the *Code of Hammurapi*.

In Figure 1, each square in the figure represents a word in the text, reading from top left to bottom right in the same sequence that texts in English are read. This visualization omits line breaks, so the text is represented as if it were a single long paragraph. The text is the *Code of Hammurapi*, in the translation on the Sacred Texts site.

The figure below shows where the words “boat” and “hire” occur within the *Code*.

Figure 1: Mentions of “boat” (red squares) and “hire” (green squares) within the Code of Hammurapi

boat hire



This shows a systematic layout of themes in the text: first, the theme of boats, then the theme of boat hire, and then the theme of hire, with boat hire as the area of overlap. The two outlying mentions of "boat" toward the end of the text deal with the hire of boatmen, which is treated as part of the theme of hiring people.

The illustration above comes from an English translation of the *Code*. However, because this approach uses a purely diagrammatic representation, it can be applied to texts in any language: what you look at is the relative positions of the colored squares, not at the underlying text. This makes it possible to show the locations of specified words in the original language of a text. It also makes it possible to compare structures within documents across two or more languages – for instance, Greek and Hebrew texts in their original languages.

There are numerous potential pitfalls when working across languages, so it is advisable to do careful homework before using this approach on another language. In the case of our work on mentions of life and death in Genesis, for instance, the King James Version translation uses the word "life" as the translation for three different Hebrew words. We chose to treat them as sufficiently closely related in meaning to be all represented by the same color of square, but decisions like this require some knowledge of the language and concepts involved. Similarly, we focused on the abstract concepts of life and death, as opposed to the specific lives or deaths of particular people; again, decisions of this type require careful thought.

Example: Mirroring of themes across texts

The figure below demonstrates how structures can be compared across texts. It shows occurrences of the word "begat" in Genesis and in Matthew. It visualizes the Project Gutenberg copy of the King James Bible, on the grounds that this is readily accessible to readers who do not have access to the texts in their original language, or who are not fluent in those languages. Several books from the KJV are available for free on the Search Visualizer site, so readers can try their own searches there.

Figure 2 shows the occurrences of the word "begat" in *Genesis*. Although *Genesis* is a long text, most of the occurrences of this word are restricted to two main bands of mentions, plus a handful of other mentions, all within the first half of the text, and most very early in the text; there is strong thematic structuring within it. A very similar pattern appears in Matthew's Gospel.

Figure 2: Mentions of "begat" in Genesis

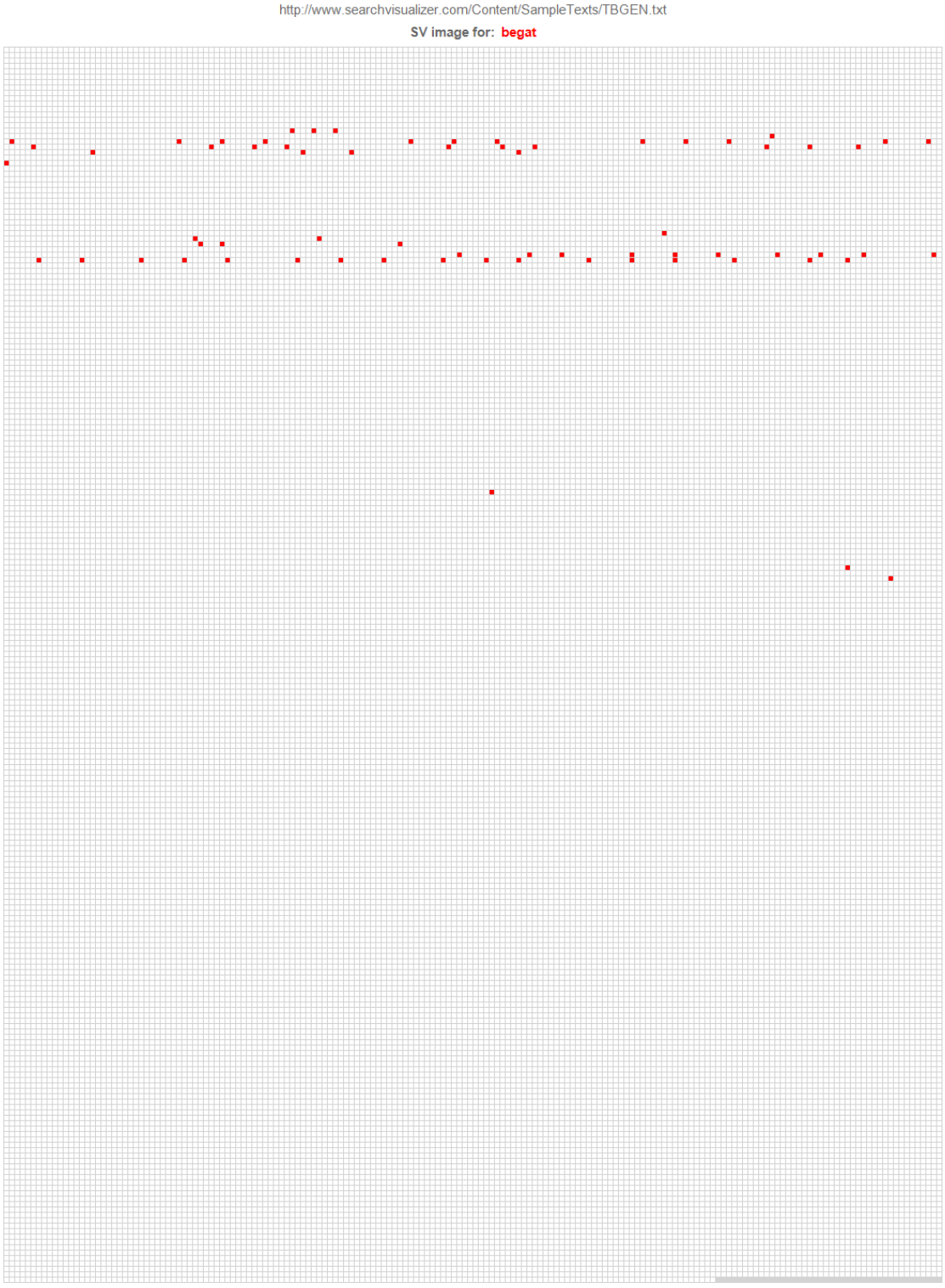
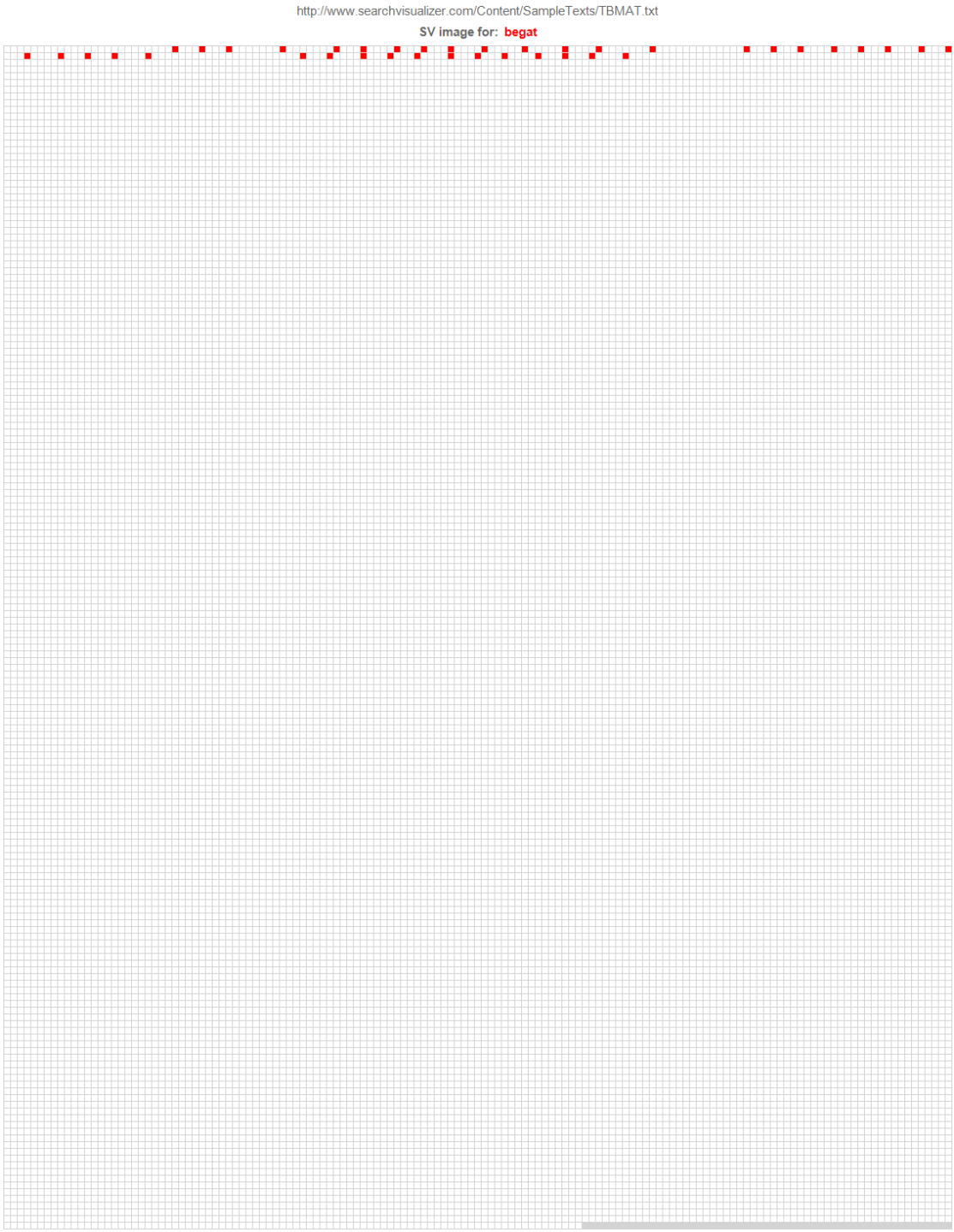


Figure 3: mentions of "begat" in Matthew's Gospel



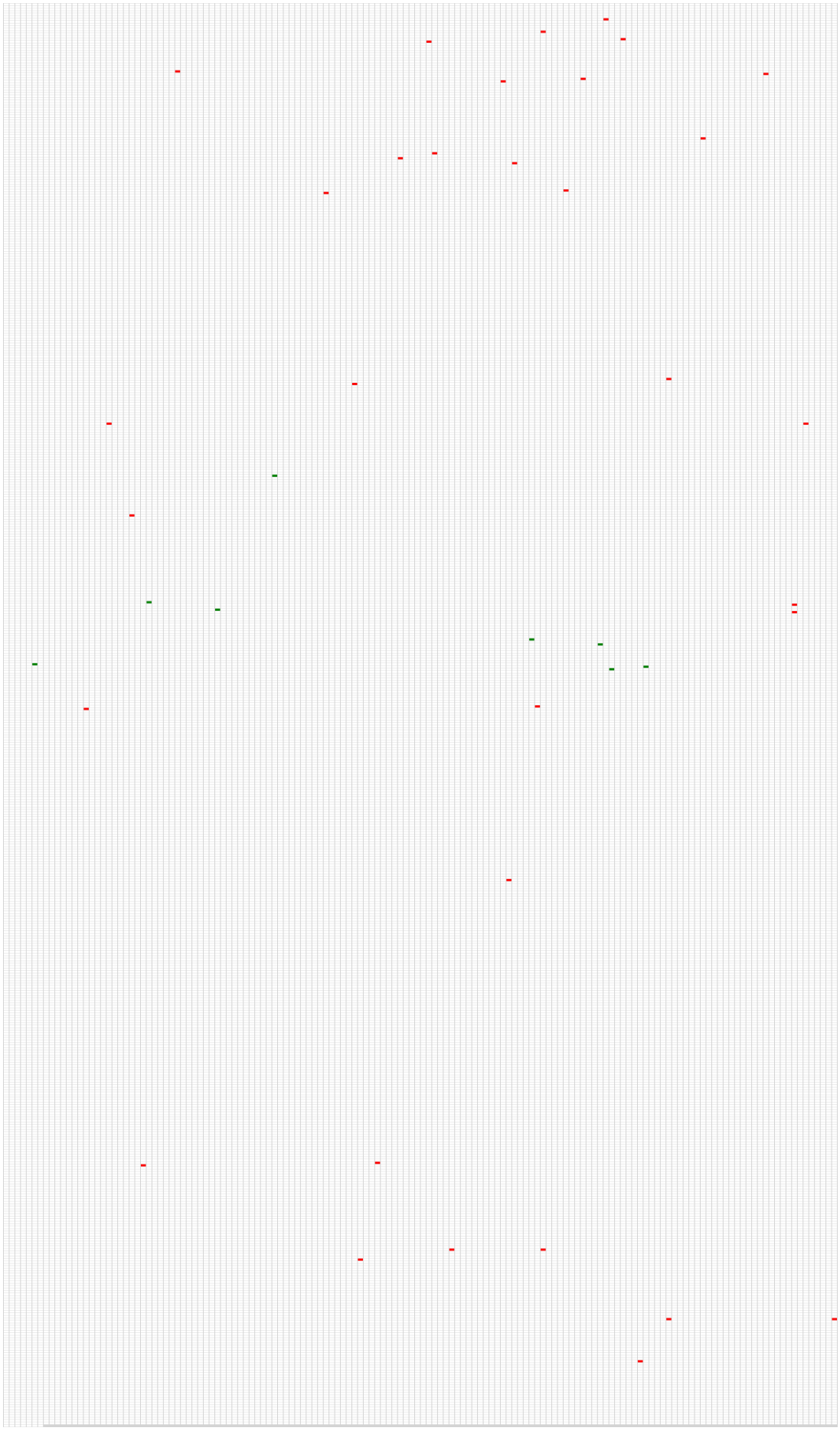
The similarities are striking: both books use the same, uncommon, word repeatedly in an intense band of mentions very early in the text. This is consistent with the way that the four canonical Gospels repeatedly allude to the Old Testament texts, both explicitly via quotations, and implicitly via repetition of Old Testament themes.

Example: life and death in Genesis

The examples so far all involve themes and structures that have long been recognized in the relevant literatures.

The next example shows a thematic structure that has received less attention, possibly because the structuring has been obscured by the sheer size of the text involved, and is only clearly visible in this type of visual representation.

Figure 4: Mentions of "life" and "death" in Genesis



It has long been agreed by biblical scholars that the twin themes of life and death are a central feature of Genesis, with the creation of life in the opening verses, and then the Fall and the arrival of death as key features of the creation story. However, when you look at these themes via a visualization, then you see something more: the themes are arranged in a symmetrical structure.

The structuring of these themes is suggestive of the rhetorical device of *inclusio*, or bracketing, where a particular theme is used to mark the beginning and the end of a section of text dealing with another theme. This is similar, but not identical, to other rhetorical devices such as *intercalation* (where two or more themes are interwoven) and *digressio* (where the flow of a narrative is deliberately interrupted by using a sub-narrative to build suspense).

In this case, the theme in the center of the inclusio is death, which is bracketed between mentions of life, like a book between two bookends, or the filling between two slices of bread in a sandwich (hence the media use of the term "The Genesis death sandwich").

Inclusio is widely used in the Old Testament, sometimes across substantial amounts of text – more than twenty chapters in the case of Jeremiah, where the bracketing theme at each end of the central portion is of almond rods and baskets of figs. However, this feature of Genesis had apparently not previously received widespread notice, if any.

This finding has implications for the debate about the authorship of Genesis, though it is far from closing the debate. This degree of structure is suggestive of a deliberate overall structuring of the text, rather than a random accretion of separate stories. However, it does not tell us whether such structuring was by an editor, an editorial team, or a sole author. For that question, it is necessary to use statistical analysis of the text – stylometrics – and traditional textual analysis.

Identifying structures in texts: Ways of asking new questions

This approach is a useful complement to stylometrics and to traditional textual analysis, since it can take words identified by those methods as potentially significant, and display their occurrence in a way that is readily comprehensible without the need for sophisticated statistical expertise.

A degree of caution does need to be applied. Sometimes, for instance, the pattern of occurrences of a word within a text reflects the chronological sequence of events rather than stylistic habits. A vivid example of this occurs the official war record for the battle of Gettysburg. That text opens with frequent mentions of the word "cavalry," reflecting the historical sequence of events. There is then a fairly abrupt transition point about a quarter of the way through the document, when cavalry are hardly mentioned. This continues until near the end of the document, at which point

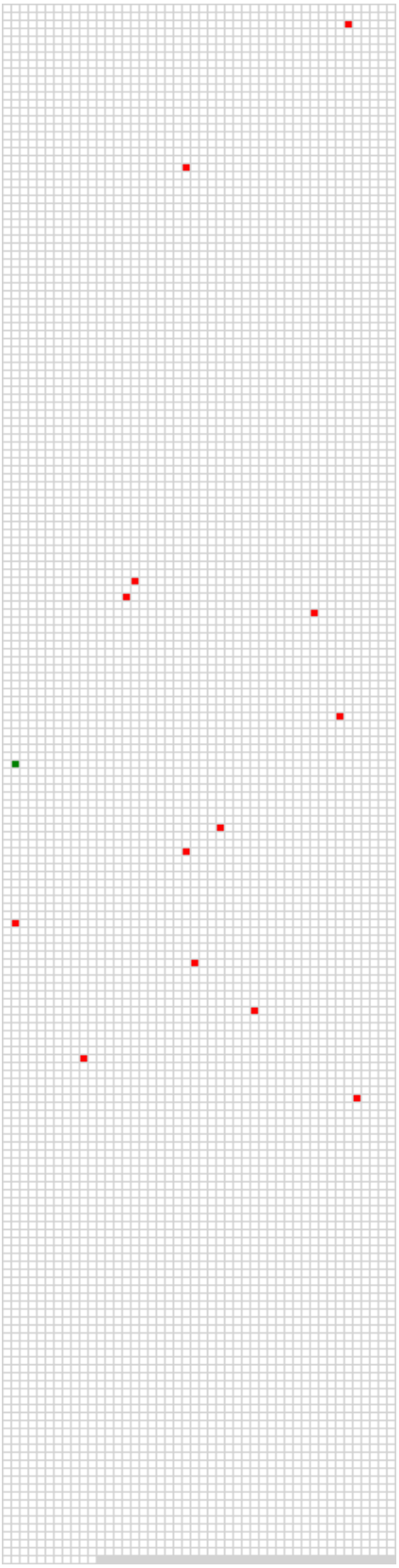
the word “cavalry” is again mentioned frequently, mirroring the usual sequence of events in a nineteenth-century land battle.

This is very different from the occurrences of the word “chariot” both in the Iliad and in the Irish epic *Táin Bó Cúailnge*. The word “chariot” occurs frequently throughout both these texts, from beginning to end. Whether this reflects military reality or artistic choice of theme is an interesting question.

An intriguing question is whether this approach can be used to identify distinguishing features of texts composed within a literate culture and of texts composed within a non-literature culture. Are there rhetorical devices and structures indicative of oral composition, that have perhaps escaped notice previously because they were hidden in the sheer bulk of surrounding material? An obvious place to start would be folk epics, such as Serbian oral poetry, building on the work of researchers such as Lord, Parry and Foley.

The closing illustration is from an early text, the *Epic of Gilgamesh*. The illustration shows how this approach can be used to compare texts across cultures and history. In this case, the chosen keywords are the same ones that were used earlier, on Genesis, namely “life” and “death”.

Figure 5: mentions of "life" (red) and "death" (green) in Gilgamesh



The text is dominated by mentions of life, but halfway through, sandwiched between mentions of life, is a single mention of death, mirroring the structure that we saw in Genesis.

In conclusion, this approach provides a useful new way of looking at texts, that complements traditional approaches. The examples above barely scratch the surface of what can be done. We hope that readers will find this approach interesting and useful for their own research.

Notes

The Search Visualizer software is available online, for use without cost:
http://www.searchvisualizer.com

The Search Visualizer site includes searchable copies of various texts, including the first five books of the Old Testament and the four New Testament gospels (all from the Project Gutenberg copy of the King James Version).

An earlier version of this paper is available on the Search Visualizer blog site:
http://searchvisualizer.wordpress.com/2012/11/16/visualising-structures-in-ancient-texts/

The translation of the Code of Hammurapi used for the visualization in this article is the one on the Sacred Texts site:
http://www.sacred-texts.com/ane

The version of Gilgamesh used in this article is from *The Electronic Text Corpus of Sumerian Literature*:
http://etcsl.orinst.ox.ac.uk/cgi-bin/etcsl.cgi?text=t.1.8.1.5
and is from the *Gilgamesh and Huwawa* (Version B)