TextArc: Showing Word Frequency and Distribution in Text

W. Bradford Paley
Digital Image Design Incorporated (didi.com)
brad@didi.com

Abstract

TextArc is an alternate view of a text, tailored to expose the frequency and distribution of the words of an entire text on a single page or screen. In texts having no markup or meta-information, one of the quickest ways of getting a feeling for the content of a text is to scan through the words that are used most frequently. Knowing the distribution of those words in the text can support another level of understanding, e.g. helping to reveal chapters in a text that concentrate on a specific subject. A structure and method of displaying an entire text on a single page or screen is presented. It reveals both frequency and distribution, and provides a well-understood and organized space that works as a background for other tools.

1. Introduction

TextArc was developed to help people deal with the ever-increasing influx of data they are forced to accept and integrate into their knowledge base. Much of that data comes in the form of raw text—e-mails, news stories, academic papers, and even a significant amount of data that could theoretically be categorized or indexed still comes to us as plain ASCII. TextArc was developed as a way to get an overview of a medium-sized body of raw text, e.g. the amount one might receive in a single day or week, and provide pointers into that text to let people more easily get to the things meaningful to their goals.

2. Existing Text Overview Methods

There are already many tools directed getting an overview of texts. Simple indices, concordances, lexicons, and other structured lists of words have been serving well for centuries. Computational linguistics techniques have recently added tools that generate automatic summaries, identify key ideas, and do semantic analysis. Several graphical techniques have also been developed to address this need, in the hopes of tapping into the vast visual processing capabilities of the human brain. Self-organizing maps have been deployed, as have multidimensional scaling techniques, to help users group similar concepts.

These approaches generally factor out one key dimension that has great meaning in a text: its original linear order. Since authors spend so much effort in crafting that order we tried to develop a technique that would respect and build on that order. This was done in the hopes that the new view to be complimentary to existing graphical and non-graphical text overview techniques, with the conviction that the expressive variety among views available to any knowledge worker is as important as the design of any one view.

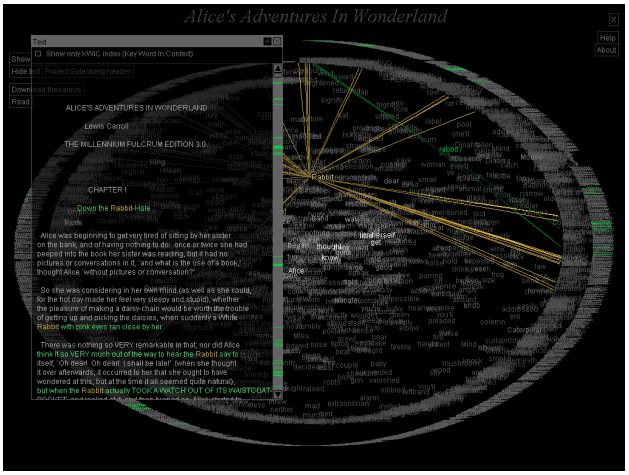

Figure 1: The interactive TextArc at textarc.org

3. TextArc Structure

A TextArc is a structure built entirely of the words in a text, generally placed in the same order that they appear in the text. Words that lose much of their meaning when taken out of context ("stop words" such as "and," "if," "the," e.g.) are not initially drawn, though they may be turned back on in a control panel in the interactive version. Words are also "stemmed;" grouped together based on their word stem (e.g. "jump," jumped," and "jumping" are represented by one word), though they can be ungrouped in the same control panel.

3.1. Text line placement

To create a TextArc first the entire text is drawn in an ellipse around the outside of the page or screen, line by line, in a tiny—potentially even unreadable—font. Lines are positioned at even increments around the ellipse: starting at the top center, keeping their baseline horizontal, and stepping each line's starting point clockwise around the ellipse. The steps around the ellipse are scaled to make the last line appear next to the first line: the angle of each step is roughly 360° divided by the number of lines in the text.

A radial baseline for lines would improve readability at the top and bottom, and is scheduled for future work.

Every line is drawn in order to retain the typographic structure of the document. Chapter breaks, headings, block quotes, poetry, references, and other typographically distinct features become visual landmarks to help users orient themselves at a global level.

3.2. Text word placement

The text is then repeated word by word along an inner ellipse. Words are as close as we can come to concepts in a raw text without making significant assumptions about language or using knowledge external to the corpus presented. TextArc was designed to work as well in Japanese as English, for example, or even in unknown languages or coding systems.

While lines strictly adhere to the rule positioning them around an ellipse, words do not. One additional rule causes the word scattering that is the key organizing principle of TextArc: if a word appears more than once in the text it is drawn only once, at the centroid of all of the points around the ellipse where it "should" appear. This fills the center of the ellipse with words that appear more than once. The averaging action of the centroid "pulls" words toward the center if they are distributed evenly throughout the text, or alternatively pulls them away from the center if they are not evenly distributed; placing them closer to chapters in which they appear more often.

3.3. Overall elliptical shape

The text is drawn in an ellipse to maximize the use of rectangular screen area, to minimize the amount of word overlap, and to ensure a relatively consistent "pull" on every word from each position it appears in the text.

The ellipse is "broken" at the top: we reduce the radius of the circle that is scaled to become the ellipse by just enough to cause an unmistakable visual gap at the top. This is done to make the beginning and ending points distinct, and to make it clear that the text is being written out following a clock metaphor. It is done at the expense of distorting the space that the word positions sample, but the tiny distortions introduced may be unnoticeable since user interpretation of word position is relatively coarse.

3.4. Word brightness and size

Words are drawn on a black background and get lighter as they are used more frequently. This is done on the assumption that a word used more frequently might be more important, so it should stand out from the background more distinctly. In the printed version, excerpted in Figure 2, type size also encodes frequency.

4. Interacting with the Prototype

TextArc has been implemented in the Java programming language, available for free public use at http://textarc.org. Interaction allows the relatively low resolution of current computer screens to express a good amount of the information TextArc extracts from a text. For example, when the user's cursor is over a word lines are drawn to the position each word was used around the arc of the text. In Figure 1, the word "rabbit" is highlighted, showing lines to every position it is used in Lewis Carroll's *Alice's Adventures in Wonderland*.

Brushing and linking also help make sense of the display. When an overlay concordance window is brought up, mousing over a word "arms" (brightens) it in that window and in the arc itself. Likewise, when an overlay window showing the original text is visible (as in Figure 1), words clicked and highlighted in the arc are also highlighted in the text.

In an interesting tradeoff of spatial resolution for interaction time, the star-like glyphs next to each word in the printed version simultaneously show every word's distribution for short texts. No mouse-scanning necessary.

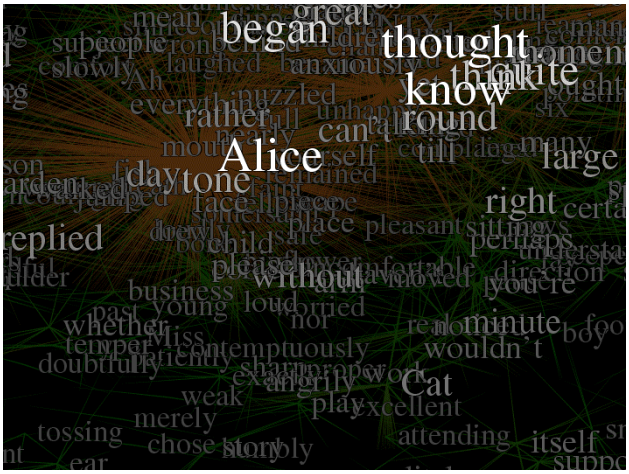

Figure 2: A printed TextArc (detail) with glyphs

5. Summary

TextArc's unusual structure is well defined and carries some information about a text to the viewer through the generally underused visual channel. It seems to show the distribution of a word in a document very effectively, filling a void left by previous text analysis techniques, and therefore may contribute a new view to help people understand text documents, using their perceptual as well as their linguistic abilities.