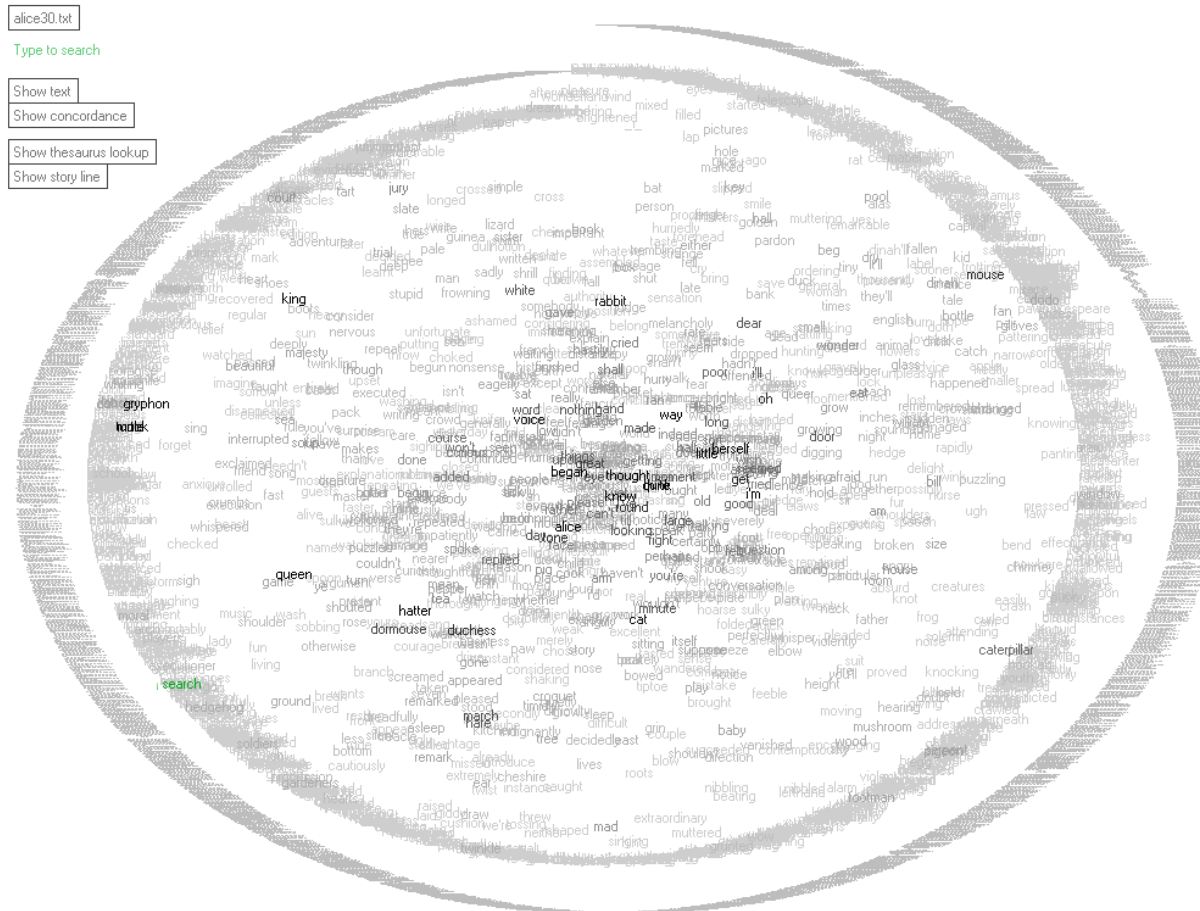


## TextArc: Revealing Word Associations, Distribution and Frequency

TextArc is a tool designed to help people discover patterns and concepts in any text by leveraging a powerful, underused resource: human visual processing. It compliments approaches such as Statistical Natural Language Processing and Computational Linguistics by providing an overview, letting intuition help extract meaning from an unread text. Here, an analysis of Lewis Carroll's *Alice in Wonderland* demonstrates TextArc's structure and some capabilities.

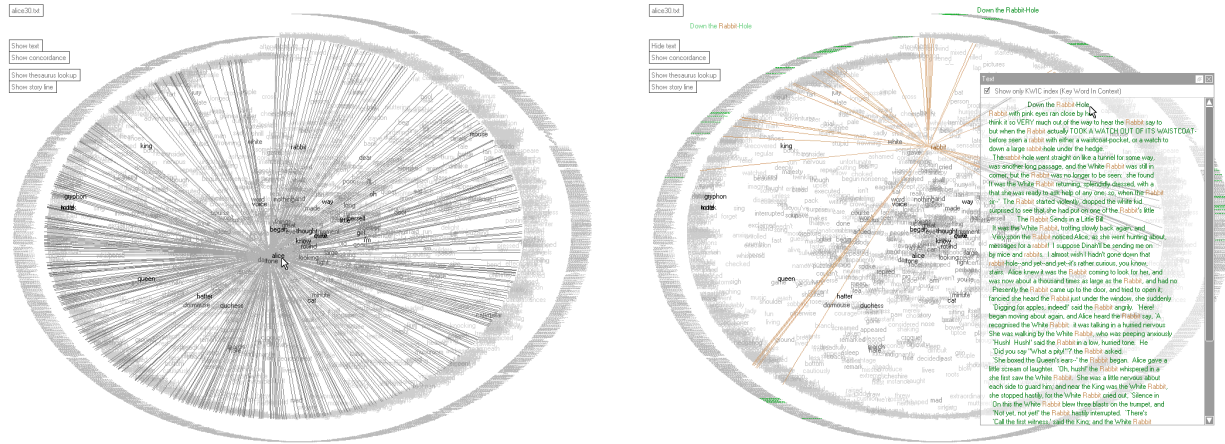


TextArc exposes the nature and style of a document's content, not by algorithmic winnowing but by arranging and showing every word. It taps into our pre-attentive ability to scan for brighter (here, more frequent) words, compare them, and let the eye read choose words in a balancing act between the . The eye and mind scan for ideas, then follow the ideas down to where and how they appear in the text.

TextArc represents the entire text as two concentric spirals on the screen: each line is drawn in a tiny (one pixel tall) font around the outside, starting at the top; then each word is drawn in a more readable size. Important typographic features, like the mouse-tail shape of a poem at about two o'clock, can be seen because the tiny lines retain their formatting. Frequently used words stand out from the background more intensely. Note that key characters like *Alice*, *hatter*, *king*, *queen*, and *gryphon* stand out, as do other words evocative of the story: e.g. *poor*, *dear*, *door*, and *little*.

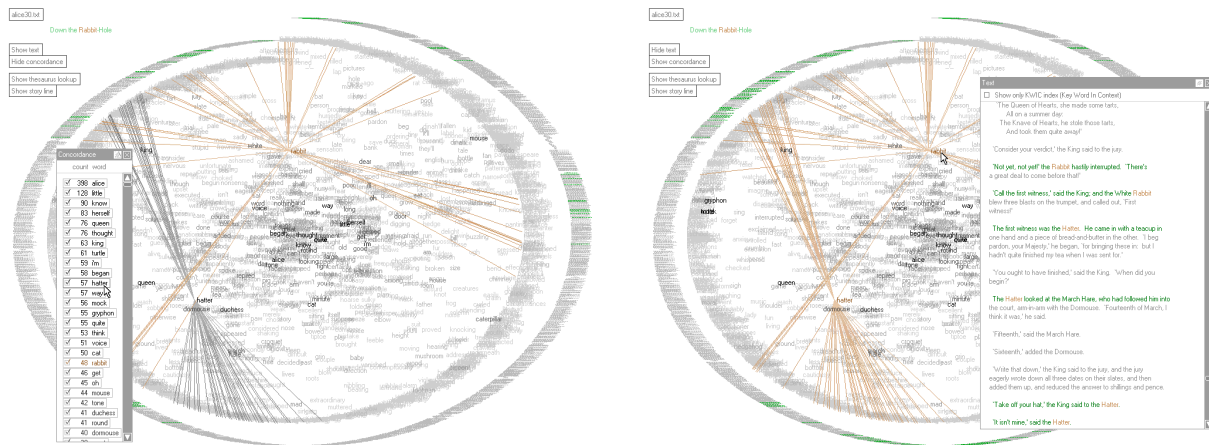
Some words appear inside the spirals. This is the key organizing structure of TextArc: words that appear more than once are drawn at their average position. Imagine each word attached to where it belongs around the spiral by a tiny rubber band; if the word appears in two places two rubber

bands are attached. The net result of this rubber band tug-of-war is that a word will appear closer to places where it is used more. Hence *gryphon* and *caterpillar* draw close to their chapters while *rabbit* stays more central. TextArc was designed with exactly this intent: words draw attention to where they appear in the document. Distribution information is revealed when the analyst points at a word: its “rubber band” rays become visible, linking it to every place it appears in the text. On the left below, we see that *Alice* appears evenly, throughout the story.



Clicking a word selects it, leaving its rays visible when the cursor moves away. A text view can show every line that uses that word: a KWIC (key word in context) index. Pointing at a line in the text view expands it in the main view, in context along the outer spiral, here just past the top.

Word associations only make sense in context. Here we select *rabbit* and point the cursor at *hatter* to see where they’re collocated and how they interleave. Note that we point at *hatter* in the concordance but its rays still appear in the main view. This tight integration between views directs one’s attention so people can fluidly follow their insights. For example, the text view on the right shows the full text, with the attended words and lines shown in exactly the same colors.



TextArc has dozens of features, and variations of it apply beyond single texts: to libraries of documents, news articles, and e-mail archives. Other enhancements can graphically code Statistical NLP results, and arrange extracted high-level concepts rather than words. Please contact W. Bradford Paley at Digital Image Design to see this and other “illustrative interfaces” in action—the true power of our tailored visualizations is the interactivity, clarity, and comfort they exhibit in use. Phone (212) 343-2442 ext. 229, or e-mail brad@didi.com.