# Theory and Practise in Literary Textual Analysis Tools

***Ray Siemens*** *(siemens@uvic.ca)*
*University of Victoria*
***Geoffrey Rockwell*** *(georock@mcmaster.ca)*
*McMaster University*
***Susan Schreibman*** *(sschreib@umd.edu)*
*University of Maryland*
***Matthew Jockers*** *(mjockers@stanford.edu)*
*Stanford University*

## Panel Description

Through discussion of several exemplary literary textual analysis tools, participants on this panel explore elements of the literary studies community's reaction to textual analysis computer tool development -- and, particularly, how theorists perceive the development of tools as an activity that supports, tests, models, and expands upon their work. Panel contributors challenge the oft-perceived disparity between the 'lower' criticism (enumerative, bibliographic, re-presentative, &c.) in which most computing tools that we use have their origins and the 'higher' criticism often associated with thematically-oriented literary critical theory.

**Geoffrey Rockwell**, McMaster U (presenter)
**Matt Jockers**, Stanford U (presenter)
**Susan Schreibman**, U Maryland (presenter)
**Ray Siemens**, U Victoria (chair and respondent)

### Interrupting the Machine to Think About It

**Geoffrey Rockwell**

> "A machine may be defined as a *system of interruptions* or breaks (*coupures*). . . Every machine, in the first place, is related to a continual material flow (*hylè*) that it cuts into." (Deleuze and Guattari 36)

Text analysis tools (and for that matter any form of analysis) perform two types of operations. They interrupt the flow of continuous analog information in order to break it down into samples that can be quantified and then they synthesize new eruptions out of the samples. Even the representation of a text in digital form is a matter of machined sampling and quantitative representation whether you chose to represent a printed page as pixels or characters.

This interrupting and breaking down is a process that constrains what computer-based tools can do and that is the first point of this paper. The sampling and quantization also makes it possible to develop synthetic processes that create new hybrid artefacts like text visualizations or sonoric representations, the second point of this paper.

Finally, the breaking down (and not the transparent functioning) is the (error) message of the textual machine. We know the machine when it fails, when it is in error, and when it delivers monstrous results. To stand back and look at a machine, as opposed to looking through it, is to think through ambitious failure.

Such a thinking through a computer is pragmatic theorizing in a tradition of thinking while tinkering - a thinking often provoked by what is at hand. What is proposed is a theory of computer assisted text analysis that addresses the way such ruptures stress interpretation. Development happens in rupture, both the programming development that scripts computers and the performance of thinking (about machines and texts) called developing a theory.

In the meantime, *The Bug* that mocks us and interrupts our demonstrations is also what provokes reflection and adaptation. We wouldn't want it any other way, except at the moment of machined interruption, for which reason a demonstration of TAPoRware text analysis tools will interrupt this paper.

### Bibliography

Deleuze, Gilles, and Félix Guattari. *Anti-Oedipus: Capitalism and Schizophrenia.* Trans. Robert Hurley. Minneapolis: University of Minnesota Press, 1983.

Ullman, Ellen. *The Bug.* New York: Nan. A. Talese, 2003.

Yan, Lian, and Geoffrey Rockwell. *TAPoRware.* Accessed 2005-03-22. <http://taporware.mcmaster.ca/>

### Visualizing the Hypothetical, Encoding the Argument

**Susan Schreibman**

The *Versioning Machine* (*VM*) <http://www.mith2.umd.edu/products/ver-mach> was launched at ACH/ALLC 2002 as a tool to display multiple witnesses of deeply encoded text. It was designed as a presentation tool so that editors could engage with the challenging work of textual editing, rather than becoming experts in other technologies, such as XSLT, JavaScript and CSS, all components of the *Versioning Machine*. The application allows encoders who utilize the *Text Encoding*

*Initiative*'s Parallel Segmentation method of encoding to view their documents through a browser-based interface which parses the text into its constituent documents (at present the *VM* works best with *Internet Explorer* 6.0 and higher, but it also works with *Firefox* for PC and Mac). The *Versioning Machine* also provides several features for the end user to engage with texts, including highlighting a structural unit (paragraphs, lines, or divs) across the witness set, synchronized scrolling, and the ability to display a robust typology of notes.

The *TEI*'s Critical Apparatus tagset (as outlined in Chapter 19 of the *TEI*'s *Guidelines*) provides a method for capturing variants across a witness set. This highly structured encoding brings together in one document n number of witnesses which an editor considers the same work. The encoding enabled by parallel segmentation provides a typology for indicating what structural units of text, or parts of structural units, belong to each witness. In this way, content which appears in more than one version of the work is encoded once, with attribute values indicating which witness or witnesses it belongs to. It is an extremely efficient way of encoding in that the editor is saved the repetitious work of encoding the content which persists over multiple witnesses, as one would do if each witness were encoded as a separate document.

The apparatus element or `<app>` acts as a container element binding together the various readings, which are encoded within a reading `<rdg>` element. Attribute values indicate which witness or witnesses a particular structural unit (a paragraph or line, for example), or subunit, belongs to (See figure 1.).

```
<lg n="1">
<l n="1">
<app>
<rdg wit="a1 a2 a3 a4 pub">The sun burns
out,</rdg>
</app> </l>
<l n="2">
<app>
<rdg wit="a1">The world withers,</rdg>
<rdg wit="a3 a4">The world
withers,<milestone unit="stanza"/></rdg>
<rdg wit="a2 pub">The world
withers<milestone unit="stanza"/></rdg>
</app> </l>
```

*Figure 1. A fragment of parallel segmentation encoding*

When parsed in the *Versioning Machine*, the aforementioned fragment, the title of the text, along with the first few lines, is rendered as follows for the first three versions:

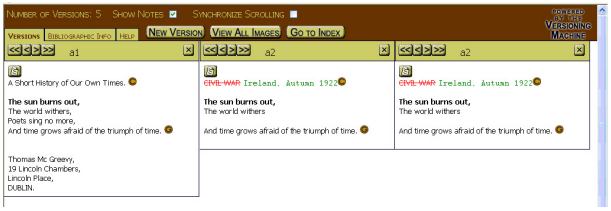


*Figure 2: The title of 'Autumn' rendered in the Versioning Machine*

In Lessard and Levison's 1998 article "Introduction: quo vadimus", they argue that computational humanities research has not achieved a level of acceptance because of the differences in "opposing intellectual paradigms, the scientific and the humanistic" . The scientific, they argue, is based on formulation of hypotheses, collection of data and controlled testing and replication. The humanistic paradigm, they argue is based on argument from example, "where the goal is to bring the interlocutor to agreement by coming to see the materials at hand in the same light" (263).

While the *Versioning Machine* was designed as a visualization tool, it is no less importantly an environment within which editors realize a theory of the text, bringing readers to an understanding of the work as embodied in its multiple witnesses. It can thus be seen within Lessard and Levison humanistic paradigm, as a tool for presenting a reading of the work through its editing and encoding, itself a primary theoretical event (McGann 75). Moreover, this primary event can illuminated and explicated though more traditional scholarly apparatus, such as annotation, adding an additional layer of textual analysis.

Thus the *Versioning Machine* provides a venue not only to realize contemporary editorial theory, but to challenge it. It meets the requirement that Stéfan Sinclair outlines in his 2003 article "Computer-Assisted Reading; Reconceiving Text Analysis" in that it is a tool which is relevant to literary critics' current approaches to textual criticism (178). The *Versioning Machine* is an active editing environment: it has been used by encoders editing texts as different as Renaissance plays and Dadaist poetry. The *Versioning Machine* is a tool which takes as its premise that the goal of much contemporary editing is not to create a definitive edition, but rather a "hypothesis" of the text (Kane-Donaldson as quoted in McGann 77), which can be read alongside an unedited edition of the text (that is, a reproduction of an image of the text in documentary form; McGann 77, Siemens). As such, it makes visible encoding as criticism, providing an environment to challenge our approaches to complex texts in terms of theories of encoding, as well as contemporary editorial theory.

## Bibliography

Lessard, G., and M. Levinson. "Introduction: quo vadimus?" *Computers and the Humanities* 31.4 (1998): 261-269.

McGann, Jerome. *Radiant Textuality: literature after the World Wide Web.* New York: Palgrave, 2001.

Schreibman, Susan, Amit Kumar, and Jarom McDonald. "The Versioning Machine." *Literary and Linguistic Computing* 18.1 (2003): 101-107.

Siemens, Ray. "'Unediting and Non-Editions' The Theory (and Politics) of Editing." *Anglia* 119.3 (2001): 423-455.

Sinclair, Stéfan. "Computer-Assisted Reading; Reconceiving Text Analysis." *Literary and Linguistic Computing* 18.2 (2003): 175-184.

Sperberg-McQueen, C.M., and L. Burnard, eds. *TEI P4: Guidelines for Electronic Text Encoding and Interchange.* Text Encoding Initiative Consortium, 2002. Accessed 2004-10-09. <http://www.tei-c.org/P4X/>

Vetter, Lara, and Jarom McDonald. "Witnessing Dickinson's Witnesses." *Literary and Linguistic Computing* 18.2 (2003): 151-165.

## Electronic Text Analysis and a New Methodology for Canonical Research

**Matt Jockers**

Using a combination of 'typical' text analysis tools (concordance and collocation) and other custom tools developed by the author, this paper demonstrates that conventional 'higher' criticism with its fashionable and thematically-oriented theoretical approaches fails as a means of assessing and generalizing about canons and genres of literature. Drawing on a case-study of the canon of Irish-American prose, the paper employs a quantitative and, indeed, scientific methodology to offer a radical reinterpretation of the canon.

In support of this research the author collected, coded, and categorized a database collection of prose literature including over 750 individual works written by some 280 different authors. The collection spans a period of 300 years and nears being comprehensive in terms of its scope and coverage of the prose canon and genre of Irish-American ethnic literature. In addition to the usual metadata associated with electronic archives, each work in the collection is tagged with metadata related to the nature of the work: metadata includes geographic setting (East or West of the Mississippi), regional setting (Northeast, Southwest, Mountain, Pacific, and etc), information about whether the work is set in an urban or rural environment as well as data specific to the author of each text. Using his own *Corpus Analysis Tools Suite* (*CATools*), a set of analytic tools developed using php and mysql for doing both semantic and quantitative text-analysis of materials specifically housed within a relational database structure, the author has mined the material in order to reveal latent chronological, semantic, and geographic trends within the overall canon since its beginning in the late 18th century to the present.

The results of this work not only challenge the best available scholarship on the subject of Irish-American literature but further challenge the efficacy of contemporary and fashionable theoretical approaches to literature that are based on the 'close-readings' of texts. In making the case for a re-evaluation of the Irish-American canon, the paper challenges the basic and fundamental methodology of traditional literary study, and demonstrates in clear and indisputable terms that a quantitative and, indeed, scientific analysis of the literary data is not only valuable to the study of a genre or a canon of literature but essential if we are to ever go beyond the mere 'readings' and interpretations of texts.

## Response

**Ray Siemens**