This chapter reviews progress on a new corpus-based text analysis technology developed at Carnegie Mellon, called DocuScope. The technology includes a pattern matcher recognizing hundreds of millions of language strings indicating micro rhetorical acts. It also includes a visualization environment that allows researchers, teachers, and students to see and explore these patterns in textual models and in student drafts. While the tool has been used for writing education, it has not yet been optimized for educational environments and its most successful applications thus far have been supporting statistics-based research in the data mining of electronic archives (including samples of student texts and writing models) for rhetorical features. The first half lays out an overview of analytic choices we have made for conducting textual research. The second half explores how the knowledge derived from the DocuScope tool can benefit writing education and the evaluation of writing curricula.

1 Introduction

Textual resources are becoming increasingly digitized over the Internet, a situation creating greater demand among textual researchers for digitized tools that can take special advantage of digitized text. This chapter describes a new technology used to support research and education involving digitized text, especially corpus-based rhetorical analysis and on-line writing education. Significantly, the mere availability of digitized texts and tools cannot replace a substantive framework for conducting textual research or delivering writing instruction, be it paper-based or electronic. The first half lays out an overview of analytic choices we have made for conducting textual research. The second half explores how the knowledge derived from the DocuScope tool can benefit writing education and the evaluation of writing curricula.
2 Frameworks for Textual Research

We have long been interested in addressing questions that relate language choice and reader experience. Our research seeks to account for the wide variation in the experiences texts afford. How do the writer’s small and recurring choices at the point of utterance matter to the whole text experience of the reader? Answering this question is important to rhetorical analysts who wish to understand how the plasticity of language choice affects the plasticity of responses to rhetorical situations. Answers are also important to writing educators who wish to understand the wide palette to which students must be exposed in order to master writing across a range of genres and situations.

Our research is also addressed to the kernel of rhetorical theory by exploring micro–macro connections in language design. In classical rhetorical theory, the canons of invention and style are described as separate arts. We wish to understand the deep interconnections between these arts, the extent to which local decisions traditionally associated with “style” aggregate to inform global plans associated with “invention.” In what sense can micro-selections of text contribute compositionally to a text’s overall genre features? Our questions focus on the constellation of language choices that provide one or another reading experience to an audience or user. Underlying this variation are what we have called “design” choices (Kaufer & Butler, 1996, 2000; Kaufer, Ishizaki, Butler, & Collins, 2004). Viewed from a theory of writing as a design activity, these design choices include, but are by no means exhausted by, the following list of decisions:

- Story-telling perspective (e.g., first-person vs. third-person),
- narrating vs. informing vs. arguing,
- subjectively tinged observations vs. “objective” referential/descriptive reporting,
- temporal orientation (past vs. present vs. future),
- writing from personal authority vs. invoking a shared authority,
- acknowledging and guiding the reader vs. leaving the reader unaddressed, and
- keeping readers in a scene or cutting across events and scenes.

What we call types or genres of writing are deeply informed by the implicit decisions writers make along these and many other choice points. Our research has sought to uncover what these implicit choice points are and the various ranges of response they afford. These choice points and responses to them cannot be uncovered by studying single texts or even single types of text in isolation. They can only be uncovered by studying any text (or type of text) alongside other texts and types as they are distributed across the language.

A member of our research group, Collins (2003) has found, for example, that detective mysteries rely on more narrative-like elements than informational or argumentative elements. But this result could not have been found by studying mystery stories by themselves, which, as a matter of fact, do contain a great deal of language associated with informing or arguing. The result was determinable only through comparative analysis, by learning that mysteries exhibit a higher proportion of textual elements associated with narrative, relative to elements associated with information and argument, than do other genres, such as government reports, instructional manuals, and legal briefs.

The need for comparative analysis led us to develop a computer-based analysis that would allow us to code language patterns across large numbers and kinds of texts. Our