1

INTERESTED COMPLICITIES

The Dialectic of Computer-Assisted Writing Assessment

Ken S. McAllister and Edward M. White

> She knew how difficult creating something new had proved. And she certainly had learned the hard way that there were no easy shortcuts to success. In particular, she remembered with embarrassment how she had tried to crash through the gates of success with a little piece on a young author struggling to succeed, and she still squirmed when she remembered how Evaluator, the Agency of Culture's gateway computer, had responded to her first Submission with an extreme boredom and superior knowledge born of long experience, "Ah, yes, Ms. Austen, a story on a young author, another one. Let's see, that's the eighth today—one from North America, one from Europe, two from Asia, and the rest from Africa, where that seems a popular discovery of this month. Your ending, like your concentration on classroom action and late night discussion among would-be authors, makes this a clear example of Kunstlerroman type 4A.31. Record this number and check the library, which at the last network census has 4,245 examples, three of which are canonical, 103 Serious Fiction, and the remainder ephemera. (Landow 1992, 193–194)

This excerpt from George Landow's tongue-in-cheek short story about "Apprentice Author Austen" and her attempts to publish a story on the international computer network, thereby ensuring her promotion to "Author," suggests a frightful future for writing and its assessment. The notion that a computer can deliver aesthetic judgments based on quantifiable linguistic determinants is abhorrent to many contemporary writing teachers, who usually treasure such CPU-halting literary features as ambiguity, punning, metaphor, and veiled reference. But Landow's "Evaluator" may only be a few generations ahead of extant technologies like the Educational Testing Service's e-rater, and recent developments in the fields of linguistic theory, natural language processing, psychometrics, and software design have already made computers indispensable in the analysis, if not the assessment, of the written word. In this chapter, we approach the history of computer-assisted writing

assessment[1] using a broad perspective that takes into account the roles of computational and linguistics research, the entrepreneurialism that turns such research into branded commodities, the adoption and rejection of these technologies among teachers and administrators, and the reception of computer-assisted writing assessment by the students whose work these technologies process.

Such a broad treatment cannot hope to be comprehensive, of course. Fortunately, the field of computer-assisted writing assessment is sufficiently well established that there exist numerous retrospectives devoted to each of the roles noted above—research, marketing, adoption, and use—many of which are listed in the bibliography at the end of this book. Our purpose here in this first chapter of an entire volume dedicated to computer-assisted writing assessment is to offer readers a broad perspective on how computer-assisted writing assessment has reached the point it occupies today, a point at which the balance of funding is slowly shifting from the research side to the commercial side, and where there is—despite the protestations of many teachers and writers—an increasing acceptance of the idea that computers can prove useful in assessing writing. This objective cannot be reached by examining the disembodied parts of computer-assisted writing assessment's historical composition; instead, such assessment must be treated as an extended site of inquiry in which all its components are seen as articulated elements of a historical process. This complex process has evolved in particular ways and taken particular forms in the past half century due to a variety of social and economic relations that have elevated and devalued different interests along the way.

In the following sections we trace this web of relations and suggest that theoretically informed practice in particular circumstances—what we will be calling "praxis"—rather than uncritical approbation or pessimistic denunciation ought to guide future deliberations on the place of computer-assisted writing assessment in educational institutions. Our hope is that by surveying for readers the technological, ideological, and institutional landscape that computer-assisted writing assessment has traversed over the years, we will help them—everyone from the greenest of writing program administrators to the most savvy of traditional assessment gurus—develop some historical and critical perspective on this technology's development, as well as on its adoption or rejection in particular contexts. Such perspectives, we believe, make the always difficult process of deciding how to allocate scarce resources—not to mention the equally dizzying process of simply distinguishing hype from

reality—considerably more straightforward than trying to do so without some knowledge of the field's history, technology, and "interested complicities."

INTERESTED COMPLICITIES

The process of designing computers to read human texts is usually called natural language processing, and when these techniques are applied to written texts and specifically connected to software that draws conclusions from natural language processing, it becomes a form of writing assessment. Raymond Kurzweil (1999), an artificial intelligence guru specializing in speech recognition technologies, has a grim view of natural language processing, asserting as recently as the end of the last century that "understanding human language in a relatively unrestricted domain remains too difficult for today's computers" (306). In other words, it is impossible—for now at least—for computers to discern the complex and manifold meanings of such things as brainstorming sessions in the boardroom, chitchat at a dinner party and, yes, student essays.

The disjunction between the desire for natural language processing and the current state of technology has created a territory for debate over computer-assisted writing assessment that is dynamic and occasionally volatile. It is possible, of course, to freeze this debate and claim that it is divided into this or that camp, but such an assertion would be difficult to maintain for long. To say, for instance, that there are those who are for and those who are against computer-assisted writing assessment might be true enough if one examines its history only from the perspective of its reception among certain articulate groups of writing teachers.[2] Such a perspective doesn't take into consideration, however, the fact that there are a fair number of teachers—and perhaps even some readers of this book—who are undecided about computer-assisted writing assessment; such people, in fact, might well like there to be a technology that delivers what computer-assisted writing assessment companies say it can, but who are ultimately skeptical. Nor does it consider the fact that natural language processing researchers frequently occupy a position that may be termed "informed hopefulness." Such a position neither denies the current limitations and failings of computer-assisted writing assessment nor rejects the possibility that high-quality (i.e., humanlike) computer-assisted writing assessment is achievable.

Another way the debate could be misleadingly characterized is as a misunderstanding between researchers and end users. Almost without exception, the researchers developing systems that "read" texts acknowledge

that the computers don't really "understand" what they're seeing, but only recognize patterns and probabilities. Of course, the process of reading among humans—and virtually every other sign-reading creature—also depends on pattern recognition and probabilistic reasoning, but the human brain adds to this a wealth of other types of interpretive skills—sensory perception, associative thinking, and advanced contextual analysis, for example—that makes a vast difference between how computers and humans read. Nonetheless, end users see the fruits of natural language processing research, which is often very compelling from certain angles, and declare such computer-assisted writing assessment systems either a welcome pedagogical innovation or a homogenizing and potentially dangerous pedagogical crutch. This misunderstanding is often exacerbated by the people who commodify the work of researchers and turn it into products for end users. The marketing of computer-assisted writing assessment algorithms and the computer applications built around them is an exercise in subtlety (when done well) or in hucksterism (when done dishonestly). The challenge for marketers dealing with computer-assisted writing assessment is that they must find a way around the straightforward and largely uncontested fact that, as Kurzweil (1999) said, computers can't read and understand human language in unrestricted domains—precisely the type of writing found in school writing assignments.[3]

Rather than trying to tell the story of the history of computer-assisted writing assessment as a tale of good and evil—where good and evil could be played interchangeably by computers and humans—we prefer to tell the history more dialectically, that is, as a history of interested complicities. The evolution of computer-assisted writing assessment involves many perspectives, and each perspective has a particular stake in the technology's success or failure. Some people have pursued computer-assisted writing assessment for fame and profit, while others have done it for the sake of curiosity and the advancement of learning (which is itself often fueled by the pressure of the promotion and tenure process). Some have pursued computer-assisted writing assessment for the advantages that novelty brings to the classroom, while others have embraced it as a labor-saving innovation. And some people have rejected computer-assisted writing assessment for its paltry return on the investments that have been made in it, its disappointing performance in practical situations, and the message its adoption—even in its most disappointing form—seems to send to the world: computers can teach and respond to student writing as well as humans. In its way, each of these perspectives

is justifiable, and for this reason we believe it is important to step back and ask what kind of conditions would be necessary to sustain such a variety of views and to attempt to ascertain what the most responsible stance to take to such a tangle of interests might be in the first decade of the twenty-first century.

The development of computer-assisted writing assessment is a complex evolution driven by the dialectic among researchers, entrepreneurs, and teachers. The former two groups have long been working to extend the limits of machine cognition as well as exploit for profit the technologies that the researchers have developed. Teachers, too, have been driven to shape the development of computer-assisted writing assessment, mainly by their understandable desires to lighten their workloads, serve their students, and protect their jobs and sense of professional importance. All of these people have motives for their perspectives, and some have more power than others to press their interests forward. As a dynamic system—as a dialectic—each accommodation of one of those interests causes changes throughout the system, perhaps steeling the resolve of certain opponents while eliminating others and redirecting the course of research elsewhere. In general, all of the participants in this dialectic are aware of the interests at stake—their own and those of others—and have tended to accept certain broad disciplinary shifts (from computer-assisted writing assessment as research to computer-assisted writing assessment as commodity, for example) while fighting for particular community-based stakes that seem fairly easy to maintain (like having a human spot-check the computer's assessments). It is for this reason that we see computer-assisted writing assessment as being a dialectic characterized by interested complicities: each group—researchers, marketers, adopters, and users—has interests in the technology that have become complicit with, but are different from, those of all the others.

The remainder of this chapter briefly narrates this dialectic beginning in the English department. It is there that the analysis of texts has been a staple of scholarly activity since long before the advent of the computer and where, despite its reputation for textual conservatism, innovative academics have more consistently acted as the hub of activity for the inherently interdisciplinary work of computer-assisted writing analysis and assessment than anyplace else on campus. Additionally, many readers of this book will be members of English departments seeking to engage their colleagues in discussion about the meaning and implications of computer-assisted writing assessment. Such readers will be more able to talk with their colleagues, almost all of whom have a

background in literature, if they are aware of the literary theories—theories of reading, as others may call them—that underlay response to and assessment of all texts.

NOTES FROM THE ENGLISH DEPARTMENT

When Lionel Trilling criticized V. L. Parrington in his 1948 essay "Reality in America," he did so in language that to proponents of computer-assisted writing assessment must now seem simultaneously validating and dismissive. Trilling notes cuttingly that Parrington's work is "notable for its generosity and enthusiasm but certainly not for its accuracy or originality" (1950, 15). To illustrate this criticism, Trilling complains that Parrington uses the word *romantic* "more frequently than one can count, and seldom with the same meaning, seldom with the sense that the word . . . is still full of complicated but not wholly pointless ideas, that it involves many contrary but definable things" (17). In this barrage of barbs, Trilling implies that accuracy, accountability, and stability are crucial characteristics of all good writing.

Further, Trilling here, as elsewhere, articulates the formalism that had come to dominate American literary criticism in the late 1940s and 1950s. Though based on older models of European formalism, this innovation in literary analysis was optimistically termed by American critics "the new criticism" because it eschewed such impressionistic matters as morality, biography, and reader emotion for intense study of texts as objects containing meanings to be discerned through detailed examination and close reading. Such reading, with particular attention to metaphor, irony, ambiguity, and structure, would reveal the deep meanings within the text and allow the critic to announce those meanings with a certain scientific accuracy based wholly on the words in the work of literature.[4] The few opponents of this approach complained that this dispassionate analysis was altogether too aesthetic and removed from the real and passionate world of literature and life, and that it rendered students passive before the all-knowing teacher who would unfold the meaning of a poem or a play as if solving a complicated puzzle that only initiates could work through. The charge of mere aestheticism, made fervently by Marxist and other critics with social concerns about the effects of literature, rings with particular irony now, as we look back to the new criticism as providing a kind of theoretical ground for computer assessment, an "explication de texte" also based on the belief that meaning—or at least value—resides wholly in the words and structure of a piece of writing.

Within twenty years after the publication of Trilling's essay, numerous articles had been published in the United States and Europe that treated literary texts as immutable objects available for semiscientific study; the characteristics that were projected onto the texts by a new breed of high-tech scholars—so designated because they eschewed the stereotype of the English professor by taking up computer programming and statistical analysis—were remarkably similar to the characteristics that Trilling and others had propagated a generation or two earlier. Rosanne Potter (1991), in her retrospective of the statistical analysis of literature, observes that early computer-using textual critics "took pride in discovering answers based on countable features in texts rather than on impressions" (402)—a phenomenon still quite apparent in the advertising literature for today's student essay evaluation software.[5] This pride came to its fullest fruition in 1968 when the Catholic University of Louvain opened its Centre de Traitement Electronique des Documents (CTED), an academically funded and staffed institution that had as its basic functions "developing automation in the field of the study of documents" and to act "as a training centre for the application of computing science to the human sciences" (Tombeur 1971, 335). By 1976, fueled by the early and ongoing successes of Ellis Page and the Centre de Traitement Electronique des Documents, early and encouraging developments in the field of artificial intelligence were potent enough to move prominent computational linguists Gerald Salton and Anita Wong (1976) to call for "a full theory of language understanding . . . which would account for the complete stated and implied content of the texts" (69). In other words, Salton and Wong, who were by no means alone, wanted an accurate linguistic model that could be superimposed by a computer, via a series of algorithms, over any given text to generate viable interpretations.

The post-structural theories of reading that have largely replaced formalism have done so by rejecting the narrowness and simplifications that restricted its reading of literature. Deconstruction, for instance, though fully committed to close reading of texts, emphasized the contradictions in them, the places where different meanings existed simultaneously, and replaced serious scientific analysis (or expanded on it) with new versions of reading as play and contest, both as a kind of insouciance (a book on Hegel opens with a chapter on eagles, or *aigles*, pronounced the same as Hegel in French) and as performance, in the sense that a musician "plays" and hence brings alive a musical composition. In fact, the performance of the critic virtually replaces the performance of the

author: authors die so that readers may live, according to one summary of the approach. Reader-response criticism restores the reader's role in creating meaning from a text, rejecting the new criticism's axiom that criticism should discover the best reading in the text; in this approach, every reader is entitled to—indeed, must eventually come up with—his or own text, since every reader is different from all others. And new historical readers proposed that textual meanings were to be obtained by situating the text in its social and historical contexts. In short, the limits that the new criticism placed on the experience of reading a literary text—limits that allowed those critics to reveal hitherto hidden meanings and connections within the text—were seen by the 1970s as far too restrictive for the reading and study of literature. But it was those very limitations, restrictions, and mathematical facts that ultimately provided the definition of reading that the early proponents of computer "reading" could use as they began to experiment with machine encounters with texts.

Consider the opportunity for computational analysis unwittingly pointed out by Richard Chase, an intellectual descendant of Trilling and the other exponents of the new criticism. Chase (1957) fortified Trilling's largely successful attempts to cement together a definition of "Romance" in his book *The American Novel and Its Tradition.* Chase expeditiously decrees romance characters to be "probably rather two-dimensional types, [and they] will not be complexly related to each other or to society or to the past" (13). In radically simplifying the meaning of romance characters, Chase observed that they were isolated from any real relationship with culture, history, or humanity. To most lovers of literature, such a withering review would be a lighthouse, warning unsuspecting readers away from the treacherous rocks of bad writing. But for literary hackers searching for ways to decode a living language, the flat characterizations in second-rate romance novels could prove a perfect schoolroom for computers learning to read. And so too, it might be argued, could the often formulaic writings of students. At the same time, deconstructionists expanded the range of critical attention to such texts as advertising, popular culture, television sitcoms, and, inevitably, student writing. Such texts, of no interest whatever to the new critics, offered much simpler writing for analysis than the Keats odes or Shakespeare plays favored by the formalists, and so allowed relatively simplistic readings to appear as critical insights. Computer-assisted writing assessment experimenters could use these apparently more simple texts for machine analysis while maintaining the limited scientific approach to texts they

inherited from the new critics. In this way, student writings came to be seen as a sort of proving ground for new reading and assessment algorithms, which could not approach the sophistication of aesthetic literary analysis. Nonetheless, computer-assisted writing assessment is a sophisticated project from a technical, if not an aesthetic, perspective and it is to some of these technicalities—and the researchers who pioneered them—that we now turn. In so doing, we hope to provide readers with a rough sketch of the principles and procedures upon which computer-assisted writing assessment began and upon which it continues to build, as well as to briefly characterize the historical and material conditions that provided the loam for this emerging bond among mathematicians, computer scientists, linguists, and writing teachers.

THE RESEARCHERS

Computer-assisted writing assessment is a subdiscipline of natural language processing, which is itself a subdiscipline of the field of artificial intelligence. The history of artificial intelligence research is a long and tragicomic one that involves a host of colorful characters, bitter enmities, stunning successes, humiliating failures, and more than a few hoaxes and practical jokes. Barring a look back at precomputer automata—chess-playing machines and mechanical fortune tellers, for instance—the field of artificial intelligence emerged and grew wildly during the cold war, from the 1940s through the 1980s. Natural language processing followed this same trajectory, though because of its more modest claims—and price tag—along the way, it did not suffer to quite the same extent that artificial intelligence research did when what is now commonly referred to as "AI Winter"—the period when federal funding for artificial intelligence (AI) projects was cut to a sliver of its former glory—hit in the 1980s.

During the cold war, there was a high premium on developing any and all technologies that could promote one side over the other; high-level military strategizing and force deployment occupied one set of artificial intelligence priorities, and natural language processing—in the form of universal language translation—was another. While these technologies did advance in important ways—pattern-recognition systems, neural networks, and the rudimentary translators found on the Web today are all fruits of this research—they never quite gave the return on investment promised by researchers.

With the collapse of the Soviet Union and the fall of the Berlin Wall, governmental urgency and the relatively lush funding that accompanied

it for developing highly advanced technologies fell away, and researchers were left to find new sources of funding, a situation that forced many researchers to become entrepreneurs. Thus, researchers became complicit in the interests of business—namely profit generation—and turned their attention to practical problems such as data mining, the automated translation of business and technical documents, and evaluating student writing. This move from largely self-directed research within the context of the military-industrial complex to the business world also made researchers complicit with the interests of adopters, who were, after all, their sponsors' clients. The consequences of these complicities included a new research focus on analyzing the genre of the student essay (instead of literature and documents secured by intelligence agencies) and a focus on interface design (the front end).

Ellis Page's research, with its successful trajectory from the 1960s into the 1980s, then its virtual disappearance until the mid-1990s when his Project Essay Grade (PEG) reemerged as a commercial product, exemplifies the spectacular rise and fall of artificial intelligence and its subdisciplines. There were certainly many others—Terry Winnograd, Henry Slotnick, Patrick Finn—who exemplify this history and whose work had to be adapted or abandoned in the face of this sudden funding shift. But before we describe how the entrepreneurs leveraged this change to their advantage, we wish to offer here a brief sketch of how natural language processing works, that is, of the research that underlies today's commercial writing-assessment products. It is fitting to include this here because, as noted earlier, a detailed description of the processes of commercial computer-assisted writing assessment applications is impossible to provide, not only because they vary from one implementation to the next, but also because virtually all of the most popular systems are protected intellectual property. Shermis and Burstein (2003) acknowledge this fact in the introduction to *Automated Essay Scoring: A Cross-Disciplinary Perspective* and also observe that this fact causes problems when one sets out to describe the details of such systems: "[T]he explanations as to why [computer-assisted writing assessment] works well are only rudimentary, subject to 'trade secrets,' and may not correspond well to past research" (xiii). Shermis and Burstein's book demonstrates this problem unfortunately well; despite the book's status as "the first book to focus entirely on the subject" of computer-assisted writing assessment, only three of its thirteen essays contain detailed descriptions about how their computer assessment applications work (see Larkey and Croft 2003; Leacock and Chodorow 2003; Burstein and Marcu 2004).

In our view, the technical details of computer-assisted writing assessment are an important component of its history because the technical details are the primary site of struggle for all of the players in the computer-assisted writing assessment game. Without an understanding of how these systems work generally, the work of effectively assessing the systems and their advocates and critics requires one to forego any claim on situational knowledge and rely almost entirely on instinctual and anecdotal evidence. For this reason, we offer here a brief overview of how computer-assisted writing assessment systems work in general, drawing from a single but highly influential source: *Natural Language Information Processing: A Computer Grammar of English and Its Applications* by Naomi Sager (1981). Sager, currently a research professor at the Courant Institute of Mathematical Sciences of New York University, is recognized as one of the founders of the field of natural language processing. Although she published several important studies in the early 1960s, her 1981 book *Natural Language Information Processing* is now one of the field's canonical texts and is considered the first relatively complete accounting of English grammar in computer-readable form. While many advances have been made in the field since the 1980s, Sager's computer English grammar remains the keystone in numerous computational linguistics projects around the world.[6] Sager and her team of researchers developed the Linguistic String Program, an application designed to read and analyze scientific and technical articles. Several medical research institutions use the Linguistic String Program to track patterns in everything from articles in the *Journal of the American Medical Association* to physicians' daily reports.

One of the original aims of natural language processing projects such as Sager's was not to assess writing but rather to gather content information from it. Because computers are able to process enormous amounts of data very rapidly, natural language processing researchers hoped that by making an automated system that could "understand" language, they would simultaneously create a tool capable of retrieving any sort of information from any sort of text faster than even an expert in the field could. Projections for the future of natural language processing have long included systems able to read and evaluate vast quantities of literature in a particular field—say all the articles that have been published in the *Journal of Astrophysics*—and then establish connections between all the articles, perhaps even discovering what D. R. Swanson terms "undiscovered public knowledge" (7).[7] Another natural language processing project that has long been energetically researched

is mechanical translation, the ability of a computer system to translate the prose of one language into the prose of another. And while it was Ellis Page's early work that is traditionally acknowledged as the starting point of the subdiscipline of computer-assisted writing assessment for the purposes of evaluating students, it was Sager's research that has led to some of today's most sophisticated natural language processing systems. The ubiquity of Sager's research in subsequent natural language processing projects from the 1980s forward suggests that at least some of that research lies at the heart of current proprietary student-writing assessment systems.

Sager's computer grammar of English is similar to the structure of transformational-generative grammar developed by—among others—Noam Chomsky. Sager and her team parsed out and coded hundreds of T- and PS-rules into their computer, depending on a single basic assumption about how natural language works, namely, that language is linear on the surface and this linearity is determined by grammar. Thus, if researchers could construct all the rules that dictate how "elements in well-formed sentences" may be combined, then in principle those rules may be translated into the artificial language of computers, thus enabling computers to understand natural languages like English, Cantonese, or Malayalam (Sager 1981, 4).

The way the Linguistic String Program, Sager's computerized grammar system, works can be briefly described as follows: first, the system identifies the "center sentence," or what we might call the basic sentence, as well as adjunct and nominalization strings (modifiers of one sort or another). It does this by proceeding one word at a time through the sentence from left to right. The Linguistic String Program applies the restrictions that are appropriate for it, as dictated in the lexicon, then it diagrams all possible syntactic forms and functions that term could be acting as; this diagram is called a "parse tree," and it is not unusual for the computer to generate numerous trees for each word. When the computer has finished making all the possible parse trees for one word, it moves on to the next word. Here the computer first generates all possible parse trees then compares this set of trees to the trees of the previous word(s). At this stage, the computer applies other restrictions that try to manage "local ambiguity," that is, semantically nonsensical but grammatical readings. By applying all these restrictions, the number of parse trees for each word is reduced. This process repeats until the program reaches an end mark, such as a period or question mark. At this point, the number of parse trees is usually very few, perhaps one or

two for each word. A final set of restrictions is applied to the sentence, which determines its final "meaning." This process of generating hundreds, even thousands of parse trees per sentence is very computationally demanding, and in 1981 it was necessary to run the Linguistic String Program on a Control Data 6600 minicomputer—among the fastest machines available at that time—just to get the syntactic positions of each word sorted out in a reasonable amount of time.

Finally, the Linguistic String Program analyzes the whole set of parse trees for particular meanings that a human user has asked the program to look for. The computer does this analysis by using the semantic entries in a digital lexicon and by using more restrictions that help the computer determine context. For example, consider the word "pulse." During the previous stage, the lexicon would have told the program that "pulse" can be either a noun or a verb, and the program, upon analysis of the sentence "The nurse recorded the patient's pulse as 75/120," would have marked "pulse" as a noun. But the lexicon also indicates that the noun "pulse" can refer to something physiological or astronomical. Now the computer must use the advanced selection restrictions to look at all the words in the current or previous sentences for signs about which "pulse" is meant; in this example, terms like "patient" and "nurse" indicate that "pulse" is physiological, not astronomical. The Linguistic String Program was also programmed to recognize the use of back-reference terms like "this," "the foregoing," and "thus," which stand in place of ideas mentioned in previous sentences. The reference rules made the Linguistic String Program both more complicated and much more versatile and powerful than any previous language-analysis program, because back reference is an extremely common trope in formal and informal communication.

A few years after Sager's landmark work and a few years before Ellis Page returned to his development on Project Essay Grade (a return made significantly easier by the now-easy access to powerful desktop computers), Yaacov Choueka and Serge Lusignan (1985) set out to develop software that would automate "the process of disambiguation," that is, software that could determine context. When they had completed their program, they used it to analyze Lionel Groulx's *Journal de Jeunesse*. Groulx was an early-twentieth-century Quebecois historian and ardent nationalist. Rosanne Potter (1991) describes their process and results this way: "The text [Groulx's] consists of 215,000 types, 17,300 different forms; the simple step-by-step process started when 31 ambiguous words were chosen as a test set; of these 23 words (75%) have two

different lemmatizations; seven (22%) have three, and one (3%) has four. . . . In 9 out of 10 cases, a two context is sufficient for disambiguation; even a one-context is sufficient in almost 8 out of 10 cases" (412).

George Landow, whose short story excerpt opened this chapter, never mentions the year in which "Jane Austen's Submission" takes place, but the response given by his fictional Evaluator program seems not much different from the actual response provided by Choueka and Lusignan's analysis program. Indeed, these responses do not differ markedly from the statistical outputs of current commercial applications such as e-rater, IntelliMetric, WritePlacer *Plus*, or the Intelligent Essay Assessor.

While there are numerous problems with Sager's (and others') natural language processing research—for example, their reliance on the examination, deconstruction, and reconstruction of "the well-formed sentence," and their exclusion of "colloquial or purely literary" usages of words (Potter 1991, 414; Landauer, Laham, and Foltz 2003, 108)—we have included this brief technical description only to give readers a sense of the basis upon which the rhetoric of the computer-assisted writing assessment discussion is founded. When writing assessment is reduced to tasks such as identifying "the relative frequencies of function words (expressed in words per million), [then] articles, pronouns, auxiliaries, prepositions, conjunctions, *wh-* words, [and] adverbs" become very important (Potter 1991, 412); their sheer number and the linguistic functions they serve become important in ways that seem startling to a human reader, for whom they tend to be more or less transparent. This importance has recently become marked by the proprietary ways in which such statistics are generated and processed, information that is increasingly kept under lock and key so as to protect the future revenues these algorithms might generate. Toward the end of Potter's retrospective, she suggests: "Each new generation of computing machines leads to increases in knowledge of linguistic regularities" (428). Similarly, Ellis Page, Dieter Paulus, Jakob Nielsen, and others have shown that each new generation of computing machines also leads to increases in knowledge about linguistic *irregularities*, a crucial element of writing-assessment software, from the simplest grammar-checker to the most sophisticated digital parser. It is the ability of researchers to corral and manage the regularities and irregularities of language, coordinated with the increasing demands on teachers and students alike and the defunding of artificial intelligence and natural language processing projects in the post cold war era, that paved the way for entrepreneurs to enter the picture and begin to turn writing assessment into a capital venture.

THE ENTREPRENEURS

Driving computer-assisted writing assessment's shift from federally funded to corporately funded research were entrepreneurs like Ellis Page, Jill Burstein, and Thomas Landauer (among others), and companies such as Educational Testing Service, Vantage Learning, Knowledge Analysis Technologies, Pearson Education, and Text Analysis International (TextAI). The dialectical shift their work represents is easily seen in the marketing materials they present, both in print and on the Web. Consider, for example, this blurb taken from TextAI's online corporate history: "Text Analysis International, Inc. (TextAI) was founded in 1998 to bring to market a new and pragmatic approach for analyzing electronic text. TextAI is a privately held software development company poised to take advantage of the surging demand for effective text analysis solutions with its groundbreaking VisualText technology. VisualText is the culmination of years of research and development in the field of natural language processing. The Company's products are based on software tools for developing accurate, robust, and extensible text analyzers" (2005).

Seeing a business opportunity in the abandoned work of government researchers, a raft of natural language processing entrepreneurs began writing business plans and designing practical applications and friendly interfaces to their (or their partners') complex work. The National Language Software Registry (2000) lists no fewer than 171 computer applications to analyze written text, for example, and lists dozens more in areas such as "spoken language understanding," "corpus analyzers," and "multimedia information extractors."

Ellis Page, traditionally recognized as the progenitor of computerized writing assessment with his 1966 Project Essay Grade, was a former high school teacher who saw computers as an opportunity to help struggling instructors: "Teachers in the humanities are often overworked and underpaid, harassed by mounting piles of student themes, or twinged with guilt over not assigning enough for a solid basis of student practice and feedback" (1968, 211). Page's work eventually became sponsored by the Educational Testing Service and the National Assessment of Educational Progress and has been proved to be both a reliable and valid way to assess certain aspects of student writing. We cannot speak to the differences between the current version and the 1966 form of Page's program, because, as we noted earlier, the code is proprietary. But the advertising is certainly more sophisticated. In Page's recent summary of PEG's migration to the World Wide Web, he notes with understandable

satisfaction that, in addition to the system's high correlation to human judges, a separate study had assessed PEG itself as a "cost-effective means of grading essays of this type" (2003, 50).

Similarly, the developers of the Intelligent Essay Assessor (IEA) (Knowledge Analysis Technologies/Pearson Education) and e-rater (ETS Technologies, Inc.) have capitalized on the federal funding crash of natural language processing research and developed their own successful commercial ventures. IEAs users include several major textbook and test-creation companies (Holt, Rinehart, and Winston; Harcourt; Prentice Hall) as well as an increasing number of defense-related customers. Knowledge Analysis Technologies' former president, Thomas Landauer,[8] is also a professor at the University of Colorado Boulder, and he has been deeply involved in computational linguistics for more than thirty years. The recent purchase of Knowledge Analysis Technologies (KAT) by Pearson Education (billed as "the largest education company in the world") promises to fund KAT's particular stripe of computer-assisted writing assessment for years to come, and is, says Landauer, "a dream come true for KAT. The founder's vision was to bring the enormous educational potential of our unique text-understanding technologies to the service of educators and students worldwide. The technology is now mature. The market is now ready. With the vast and varied strengths of Pearson Education and the other Pearson companies joined in the effort we now feel certain of success" (2005). KAT has found a lucrative niche that allows its research in the field of computer-assisted writing assessment to continue, albeit in directions probably unimagined in Landauer's early days.

Jill Burstein, codirector of research for Educational Testing Service's subdivision ETS Technologies, is another former English composition teacher. Unlike Page and Landauer, however, Burstein comes from a new generation of scholar/entrepreneurs, one in which the corporate context of natural language processing research is assumed. ETS has, of course, a very long history in writing assessment, dating back to the 1940s and 1950s. Despite this long history, however, it was not until the late 1990s—just like its competitors—that ETS fully committed to computer-assisted writing assessment by adopting e-rater "for operational scoring of the GMAT Analytical Writing Assessment" (Burstein 2003, 113). Due to falling computer costs and the rising expense of doing large-scale writing assessment with human labor alone, ETS began to invest in researchers like Burstein to find a way to cut costs and maximize profit. In a 2001 GRE Professional Board Report, Burstein and several of

her colleagues acknowledge this fact: "One hope for reducing the cost of essay scoring comes from longstanding efforts to develop computer programs that can, by modeling trained essay readers, evaluate essays automatically" (Powers et al. 2001, 1). The dialectic of computer-assisted writing assessment evolves at ETS in the same ways it does at Vantage Learning and Knowledge Analysis Technologies: rather than being driven by cold war politics and ideologies, marketability, usability, and profitability become the watchwords guiding research as well as funding its advancements and deployments in the public and private sectors.

Entrepreneurs, whether they are also researchers or are only funding the researchers, see the market potential for essay-assessment software and fill the void left by the National Science Foundation and other federal funding agencies. This reinvigoration of natural language processing research gives many scholars both some new liberties and some new constraints: there is money to pursue the sometimes highly abstract work of computational linguistics, but the upshot of this work must always be a significant return on investment. One practical consequence of this has been that unlike in other more academy-exclusive types of research, researchers doing corporately funded computer-assisted writing assessment must attend to the feedback given them by the adopters of the technologies they've developed. And because it is these adopters whose money ultimately funds their research, entrepreneurs are required to develop front ends and sets of documentation that make their systems "friendly"—that is, easy to use, cheap relative to some other assessment mechanism (such as human labor), and accurate according to some explicable standard—to both the adopters (who administer the assessments) and the users (whose work is assessed).

The language of the entrepreneurs' promotional materials suggests these constraints quite baldly through their easy-to-understand claims about validity, reliability, affordability, and accessibility. The consequences of the complicities among researchers and entrepreneurs are that (1) computer-assisted writing assessment and natural language processing research is channeled toward commodifiable ends (which may not be optimal from a research perspective); (2) the product is sold as a proven, rather than an experimental, technology; and (3) the assessment results (i.e., the results of the computer algorithms) must mimic human graders and appease the expectations of users rather than aim toward real interpretive complexity. In the last two sections of this chapter, we raise some of the issues these complicities catalyze with the adopters of computer-assisted writing-assessment systems—that is,

university and other institutional administrators—as well as for the people whose work is assessed by applications like e-rater and the Intelligent Essay Assessor.

THE ADOPTERS

Some teachers and administrators turn reflexively to technological solutions when funding for human labor is in crisis, as has been the case for education at all levels in recent years. They gain support from those who turn to technological solutions for other reasons, such as a genuine interest in new methodologies as well as the novelty and "coolness" factors they bring. In the majority of cases, however, educators ready to adopt computer-assisted writing assessment see it in terms of cost-effectiveness, efficiency, and perhaps in certain cases "customer satisfaction" (though this seems like an imposition of entrepreneurial rhetoric). The adopters' interests are complicit with those of researchers in that adopters need reassurance that they're getting what they pay for, that is, valid and reliable results (hence all the white papers on entrepreneurial Web sites). The adopters' interests are complicit with those of entrepreneurs in that adopters need effective solutions to labor and funding shortages and probably, in some cases, need lower-cost alternatives to continued levels of funding (i.e., the "downsizing" model). And adopters are complicit with the interests of users in that as education itself is increasingly commodified, students (and their parents) want evidence that their money is being well spent (or at least is not being wasted). Since a considerable amount of school funding is now tied to standardized tests and to the pressures of the job market, adopters and users share an interest in meeting those expectations by the most efficient and economical means possible. The consequences of these complicities may include a forfeiture of institutional control over writing assessment, a heightened sense of responsibility to users, who are suddenly subject to assessments delivered by a somewhat suspect source, and a decreased labor pool (which may temporarily reduce institutional pressure and minimize, for instance, the possibility of labor organizing or other mass-protest actions).

As this book goes to press, we can see these complicities at work in the introduction of written portions to the two tests taken by almost all applicants to American four-year colleges and universities: the SAT and the ACT. Both of these tests included short essay portions for the first time in 2005, with scores intended for use in the postsecondary admissions process. A perhaps unintended side effect is that the scores produced by these tests (mandatory for the SAT; optional for the ACT) are

replacing local tests designed for placement of students into various levels of first-year writing courses. Since the "writing" scores (derived from a combination of multiple-choice items, contributing three-quarters of the score, and a brief impromptu essay) are paid for by the student and are claimed to be valid, the expense and trouble of local testing seems unnecessary. That a single test can serve all students in all colleges for placement into all writing programs seems improbable, even with human scorers reading the essays. And although neither Pearson Education nor the American College Testing Service have declared that they will use computer-assisted writing-assessment technologies, and while both are actively recruiting human readers, it seems obvious that computer-assisted writing assessment will be pressed into service sooner or later (probably sooner) for the two million or so essays that their companies will need to score.

The adopters will be colleges and universities eager to have information on student writing abilities for admission deliberations and willing to abandon their own placement procedures—designed for their own students and their own programs—for a one-size-fits-all test of (at least to some) dubious validity. The entrepreneurs of the two large testing firms will be promoting the convenience and cost savings of the new tests, while the writing program administrators and the faculty will be raising questions about the cost-effectiveness of scores that may not relate to a particular campus writing program or its particular student profile. The entrepreneurs will tout the savings to adopters, because the scores will be delivered to the campus at no cost to the institution (though the students will pay and the testing services will make large profits), and we suspect that the faculty and the users will not have much of a voice in the final assessments or admissions decisions.

THE USERS

Students are the users whose writing is assessed and whose lives are affected by the results of these assessments. Their main interest, presumably, is to become better writers or at least to perform sufficiently well on their tests and in their classes to achieve the level of success they desire. Their interests are complicit, then, with the researchers through their desire to have their writing evaluated in a manner consistent with the expectations of the test writers or assignment givers. Users' interests are complicit with the entrepreneurs in that they need the costs of education to remain reasonable and under certain circumstances might be willing to sacrifice a certain amount of validity for a decrease in educational costs—as long as it doesn't cut into the bottom line of successful

testing. Finally, users' interests are complicit with those of adopters in that users recognize their dependence on the success of adopters and also recognize the obligation that adopters owe to them; these recognitions mean that users must both cooperate with and correct adopters' decisions—this is an integral part of the feedback process. The consequences of these complicities may include low user resistance to ineffective computer-assisted writing assessment, an inability to effectively assess computer-assisted writing assessment itself (and thereby effectively participate in the feedback process), and a sense that writing—as with Landow's Jane Austen—is not the art of saying something well but rather of saying something new using a set of preexisting rules.

CONCLUSION

The history of computer-assisted writing assessment, viewed dialectically, shows how there are a variety of sometimes contesting but always complicit interests that have shaped the direction of the discipline. These interested complicities are still at work, and writing teachers need to adopt a model of praxis—a process of critical (including self-critical) reflection and informed practice toward just ends—as they pursue their interests concerning computer-assisted writing assessment. This means that all complicit parties, but most particularly the faculty (which ultimately owns the curriculum), need to be aware of the history and profundity of the issues behind computer-assisted writing assessment. Those in the humanities should become informed of the ways literary formalism has laid the theoretical ground for computer-assisted writing assessment and also begin to understand the sophistication and complexity of modern computer-assisted writing-assessment algorithms. The time has passed for easy dismissal of (and easy jokes about) computer-assisted writing assessment; the time has come for reasoned and critical examinations of it. For instance, the questions about the validity of the SAT and ACT writing tests will not go away if, or when, the student essays are scored by computer. It will be up to humanists to demand or institute studies on their own campuses to answer these questions. At the same time, some local writing assessments may be so unreliable that computer scoring may have a role to play in improving them. If humanists do not take this step of critique, painful as it may be for many, they will be sent out of the room when serious discussion gets under way between the entrepreneurs and the adopters. If we fail to imagine the application of computer-assisted writing assessment to radically improve education, we may simply forfeit computer-assisted writing assessment to those who prioritize lucre above literacy.