3

CAN'T TOUCH THIS

Reflections on the Servitude of Computers as Readers

Chris M. Anson

Yo! I told you
U can't touch this
Why you standing there, man?
U can't touch this
Yo, sound the bells, school is in, sucker
U can't touch this

—M. C. Hammer, "Can't Touch This" (1990)

Consider, for a moment, what's going on. First, you're in a multidimensional context where you and I, and this text, share a presence, a purpose, and knowledge that delimit the interpretive possibilities and let you begin fitting into boxes what little you've seen so far, or maybe shaping a box *around* it: academic genre, essay in a book, trusted editors, a focus on machines as readers, the common use of an opening quotation (lyrics, or a poem, or a proverb, or a line of text from a famous work). This one's in a vernacular. Does its style provide the meaning you'll eventually construct as you read, or is there something important about the direct-address question? Or school bells? Or is it about M. C. Hammer—a rapper launching a most un-rap-like text?

Curious, you move on, absorbing each new bit information, activating memory and prior experience to make something more of this than random words or the mutterings of the mad. After all, the text is validated by its context; it's been what Pratt (1977) calls "preselected." And just then—that reference, with its scholarly-looking date, adds a soupçon of authority. *Soupçon.* A bit of high-minded lexis. Will there be a thesis? Possibly. Is it emerging here, toward the end of the second paragraph? Doubtful; it wasn't at the end of the first. But there is a *cumulative* sense of direction and purpose—the text is adding up to something, and you move on to test various hypotheses as you automatically forgive the intentional sentence fragments. Meanwhile, that old reflective turn, *metacognition,* has been disturbed and awoken from its usual reading

slumber: the text is calling attention to your relationship with it and making you think, *what's going on here?*

In reading and interpreting the lyrics and introduction to this point, you have employed a dazzling array of conscious and tacit, cognitive and social, discursive and structural, temporal and historical, linguistic and intertextual knowledge, tangled and interdependent. If you're a machine, our condolences: you'll need far more than a latent semantic analysis program to say anything of *significance* about the text—to interact with it, to converse with yourself through it. Alas, more information alone won't help: even if your data bank boasts a domain that includes M. C. Hammer, what in cyberspace will you do with the reference except spit out a response that there is no relationship between the chunk of his song—no doubt already flagged as "poor" on the grammar and sentence pattern scale—and the rest of the text? And what about the possibility of irony, of self-consciousness?

This essay argues that the processes humans use to read, interpret, and evaluate text can't be replicated by a computer—not now, and not until long after the written ideas of the current generation of learners and teachers are bits of archaic-sounding print losing their magnetic resonance on the disks and drives of antiquity. Machines are incapable of reading natural discourse with anything like the complexity that humans read it. This assertion—though obvious to all but the most impassioned believers that Hal is just on the horizon—suggests several important consequences for the push to create machine-scoring systems for writing, among them the relegation of meaning, audience, and rhetorical purpose to the trash icon of human literacy. In an unexpected turn of direction befuddling the coherence parameters of anything but a human reader, I'll then argue *for* the continued exploration of digital technologies both to analyze human prose and possibly to provide formative information that might be useful to developing writers. Unlike the use of computer technology to make judgments on writing for purposes of ranking, sorting, or placing students, such applications are neither premature nor of questionable value for the future of composition in general and reading, responding to, and evaluating student writing in particular.

AI: WHAT WE LEARN FROM ITS BRILLIANT FAILURES

A number of goals have been proposed for the development of machine-scoring systems that can "read" essays produced by humans and analyze, rate, or evaluate the essays. The results could be used, for

example, to provide information about applicants for positions requiring some degree of writing ability, to place students into writing or other courses appropriate to their skill level, or to yield one of a number of indices that can decide whether a student should be accepted to a particular college or university or should pass from one level to the next in elementary and secondary education. The allure of such programs is obvious: computers could scan and evaluate an unlimited number of brief texts written by novice writers and provide the same results as human readers but with greater consistency and much greater efficiency. So optimistic are the advocates of some machine-scoring programs that they even flirt with anthropomorphic descriptions of these programs' capacities. Streeter et al. (2002), for example, assert that the Intelligent Essay Assessor "*understands* the meaning of written essays" (1; emphasis added). If such a claim were true, there would be no further need for teachers to read and respond to student writing—ever. In place of human readers, machines could understand texts in dozens of different settings. The government could rate entire school systems on the basis of machine-scored performances and allocate funding accordingly. Computers might even be able to read the transcripts of court hearings and reach verdicts that would determine the fate of human defendants, and do so without all the usual interpretation, discussion, and negotiation—all those messy subjectivities to which humans are prone.

But before we can speculate about such applications (or horrors), we need to explore what is meant by the capacity to "understand." What is involved in understanding written text? What are some of the processes humans use to do so, and does it seem likely that computers could be programmed to replicate those processes? Some answers to these questions can be found in the pioneering work of artificial intelligence (AI) and natural language, whose cycles of failures and successes did much to illuminate human language and reveal some of its astonishing complexity.

The development of AI in natural language has focused on different but related goals: to simulate the human *production* of language and to simulate the human *comprehension*—or, more commonly termed, "processing"—of text (see Wagman 1998). Throughout the 1970s, as burgeoning technologies inspired new speculation and experiments, AI experts investigated whether computers could do anything meaningful with "natural language," the text produced and interpreted by human beings in the course of daily life. In a series of fascinating explorations at the Yale Artificial Intelligence Laboratories, Roger Schank and his

colleagues and students set out to simulate the processes of both human language production and interpretation. As Schank (1984) articulated it, the goal of getting computers to begin "understanding" language was as much about coming to a fuller description of what humans do when we make and use language: "If we can program [a computer] to understand English and to respond to sentences and stories with the kind of logical conclusions and inferences an average human would make, this would be quite an achievement. But before we even tackle such a problem we will have to learn how humans understand such sentences and form their responses to them. What is language and how do humans use it? What does it mean to understand a sentence? How do humans interpret each other's messages?" (14).

To answer these questions, Schank and colleagues began creating sophisticated programs designed to use and manipulate natural language. Some programs worked with simple, prototypical characters and plots in order to create brief but coherent narratives. *Tale Spin*, for example, created fable-like stories using a stock set of animal characters, props, scenes, and motives. As Schank and Abelson (1977) describe it, this program, created by Jim Meehan, "makes up stories by simulating a world, assigning goals to some characters, and saying what happens when these goals interact with events in the simulated world" (210; see Meehan 1976). Each time the program created a story, however, its mistakes revealed certain kinds of knowledge fundamental to human language processes that computers lacked and had to be given. Although the chronicle of these failures is too extensive to be summarized here, a few examples are instructive.

In an early iteration, *Tale Spin* produced the following story from its many programmed roles, actions, and scripts:

> One day Joe Bear was hungry. He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe threatened to hit Irving if he didn't tell him where some honey was. (Schank and Abelson 1977, 83)

In this output, it became clear that the system needed to know what it had just said. The computer "was capable of answering the question *Where is honey found?* But it could not look back at *beehive* and see that is where honey can be found" (83). When we use language, our texts are not linear; the assertions and ideas represented in each bit of text cumulatively (and exponentially) complicate and inform both further text and prior text. In reading (or listening), we look to previous text to

interpret incoming text and make predictions about text yet to come. We move backward and forward at the same time.

A later attempt yielded the following tale:

> Once upon a time there was a dishonest fox and a vain crow. One day the crow was sitting in his tree, holding a piece of cheese in his mouth. He noticed that he was holding the piece of cheese. He became hungry, and swallowed the cheese. The fox walked over to the crow. The end.

This time the program, which had been set up to create an Aesop-like plot, had been given the knowledge that the animal characters could be aware that they were hungry, and that would activate a goal to satisfy their hunger. But this awareness is not automatic, or else every character would try to satisfy its hunger immediately whenever food is present. The program had to allow the crow to hold the cheese but not eat it.

Natural language processing experiments at Yale also produced a number of programs designed to read and "interpret" existing text for routine purposes, such as taking a full news story and condensing it into its essential ideas. For example, SAM (Script Applier Mechanism) was a prototype designed to answer simple questions about texts it had processed. In the many false starts in this and other programs, the researchers uncovered dozens of types of world knowledge applied by humans to natural discourse: actions, roles, causal chains, properties, possibilities, and plans. Schank and colleagues had stumbled on a central problem: the need to account for the linguistic and psychological ubiquity of *inferencing*. Inferencing occurs in natural language constantly, providing the connective tissue between assertions and yielding meaning and interpretation. It works all the way from simple word-level semantics to the level of entire discourses in context. For the former, Schank offers the following example of the need for computers to "know" almost limitless permutations of word meanings.

> John gave Mary a book.
> John gave Mary a hard time.
> John gave Mary a night on the town.
> John gave up.
> John gave no reason for his actions.
> John gave a party.
> John gave his life for freedom. (1984, 93)

Or, to use another example, consider what a computer programmed to understand "hand" (in, say, "Hand me a cookie") would do with "John

had a hand in the robbery" or "John asked Mary for her hand," or "John is an old hand" (94). Simply programming the computer with alternative meanings of a word wasn't sufficient to get the computer to know what meaning to apply in a given case—which requires knowledge of the word's surrounding sentential and discursive context.

While such word-level semantic puzzles were not a major programming obstacle, work in AI at the level of discourse began revealing that humans bring to text a vast, complex storehouse of prior knowledge and experience. To simplify and categorize some of this knowledge in order to continue programming computers to work with natural language, AI researchers proposed several domains of inferencing, such as scripts, plans, and goals. As Robert Abelson conceived them (Schank 1984, 143), scripts constitute stereotypical world events based on typical experiences within a known culture or context. A common example is the restaurant script activated in the following narrative:

> John went to the new fancy French restaurant. He had coq au vin, a glass of Beaujolais, and mousse for desert. He left a big tip.

Most readers who have dined at fancy restaurants will fill in many times more information than is presented in this brief text by accessing scriptual knowledge. A fancy restaurant script includes certain roles, props, and actions. A diner is seated, often by a host or maitre d'. A waiter brings a menu. (Sometimes multiple waiters play different roles; a sommelier might provide advice on the wines.) Ordering is done from the table, where the check is also paid and tips are left. Assumptions about John's experience—that he ate with a knife and fork, or that the wine was poured into a glass and not a plastic cup or a chalice—come not from the text on the page, but from knowledge the reader brings *to* the text. No one reading this text would infer a scenario in which John leaves only the tip, without paying the bill, or leaves the tip in the bathroom sink, or goes into the kitchen to order and pick up his food. "John heard a knife clatter to the floor" will activate certain further interpretations and responses (for example, it's a mild social gaffe to drop your silverware in a fancy restaurant).

But compare the script for a fast-food restaurant, where the roles, props, and actions are altogether different: ordering at a counter, paying before eating, taking your own food to a table, and so on. No one reading a text about John going to Burger King would infer a scenario in which John waits at his table for someone to show up with a menu. The line, "John heard a knife clatter to the floor" will, in the context of

a fast-food script, lead to a different set of interpretations (someone is up to no good, for example, since only plastic knives are found in fast-food restaurants). Unless something explicitly contradicts it, a script creates a mass of inferential knowledge literally not present in the assertions of the text, invisible to a parser and simple decoder. Without this knowledge, a computer can't begin to work with even simple narratives or accounts of events, much less sophisticated academic arguments or other essays expected of students in schools and colleges.

As Boden (1977) points out, interpretation "is largely a creative ability of filling in the gaps" (310). Yet although it seems to require more cognitive effort for humans to "read between the lines," inferencing also appears to be desirable in many kinds of texts. When humans are provided with all the information needed to fill in a script, the resulting text is unappealing. A text such as "John wanted to marry his friend's wife. He bought some arsenic" is rendered horrendously boring when the considerable inferencing is filled in, but this is precisely the level of detail that a program needs in order to make sense of the text. Even the following more explicit version leaves out much necessary information:

> John wanted to marry his friend's wife. To marry his friend's wife, John knew that he had to get rid of his friend. One way to get rid of his friend would be to kill his friend. One way to kill his friend would be to poison his friend. One way to poison his friend would be to give his friend arsenic without his knowing. John decided to get some arsenic. In order to get some arsenic, John needed to know where arsenic was sold. In order to find out where arsenic was sold, John had to consult the Internet. In order to consult the Internet, John had to go to his computer. In order to go to his computer. . . .

Although it is possible to program computers to work with simple scripts such as going to a restaurant or riding a bus, interpreting natural language also involves making inferences that don't rely on scriptual knowledge. Schank offers the following narrative as illustration:

> John knew that his wife's surgery would be very expensive. There was always Uncle Harry. He reached for the phone book. (1984, 125)

Schank points out that most people don't have a script for paying for expensive medical treatments. Yet such a situation is not unlike paying for college, making a down payment on a house, and so on—in a general sense, *raising a lot of money for an important family expense* (1984, 126).

But the problem is not solved by having *more* scripts: there will always be a new situation to which we can't apply an existing script. Rather than describing a stereotypical course of events such as riding a bus, this kind of discourse calls into play a set of *goals* and the *plans* required to achieve them. Some goals are far-reaching, requiring many sets of plans; others are quite simple—in Schank's example, "Fred couldn't get the jar lid off. He went down to the basement and got a pair of pliers." In applying world knowledge to texts, we bring to bear thousands of possible plans for achieving countless goals.

To program computers to work with natural language, AI researchers had to begin with simple goals achieved by simple plans. One goal, called CHANGE PROXIMITY, had several plans, such as USE PRIVATE VEHICLE, USE PUBLIC TRANSPORTATION, USE ANIMAL, USE SELF (Schank 1984, 127). Consider the example "Frank wanted to go to the Bahamas. He picked up a newspaper." On the surface, these two assertions are unrelated. It is only by inferring both a goal and a plan to realize it that the sentences can be related. In this case going to the Bahamas must involve changing proximity; changing proximity can be accomplished by using a private vehicle, using public transportation, and so on. In addition, a computer needs to know that not all the transportation plans available in its list will work to reach all goals of changing proximity. It needs to be able to fit action into a *model of the world*, to rule out, say, driving, kayaking, or riding a whale to the Bahamas, And then, using stored information, it needs to infer that there might be information in the newspaper about the remaining modes (boat, plane) that could create a plan to realize the goal (128).

This need for ongoing inferential processes becomes even more obvious when we add a third sentence to the text:

> Frank wanted to go to the Bahamas. He picked up a newspaper. He began reading the fashion section.

Any activated inferencing about transportation must be modified with the addition of the third sentence, since the goal cannot be accomplished by getting information from that part of the newspaper. Instead, perhaps Frank has the goal of obtaining light clothing for a warm climate (Schank 1984, 128), and an entirely new set of plans comes into play. The further textual information erases prior inferencing and replaces it with new inferencing—a new hypothetical goal and new hypothetical plans to reach it.

Other kinds of world knowledge essential for understanding text include "roles"—specific motives and actions assigned to people based

on their positions. We interpret the sentence "The police officer held up his hand and stopped the car" not to mean that the police officer, Superman-like, pressed his hand against the oncoming car and held it from advancing, but that the driver applied the brakes in deference to the police officer's authority. Inferences about the plans, goals, and actions of people also derive from their roles: bank teller, pharmacist, teacher, politician, nurse, habitual child molester working as a gardener at a private middle school. Given that multiple themes, plans, goals, and scripts are at work in even relatively simple discourse, it is not difficult to imagine the complexity of "making sense" of broader, less constrained texts. Each inference can produce further inferences, a process that led the early AI programs to create what Schank calls "a combinatorial explosion of inferences," some "valid and sensible," others "ridiculous and irrelevant" (1984, 141).

An even more complex kind of inferencing involves applying knowledge of *themes* to natural texts. A theme consists of the background information that we need to interpret that a person wants to achieve a certain goal. A *role theme*, for example, allows us to interpret what might motivate a particular person or character in a text to do something. If we read that a wild West sheriff is told that someone's cattle has been stolen, we might infer that he has a goal of recovering them and/or bringing the thieves to justice. As Black, Wilkes-Gibbs, and Gibbs (1982) explain, "role goals are triggered by the actions of other 'players' when these actions become known to the character of the role. Once such a goal is successfully triggered, the character's plans are much more predictable than if a non-role person had the same goal" (335). For example, if the sheriff has the goal of catching the thieves, it's likely that he'll saddle up his horse, or enlist the help of a posse. If as readers we encounter, "Jack told the sheriff, 'My cattle are gone!' The sheriff went to the saloon to find his pals Slim, Ernie, Baldy and Pete . . . ," we assume the sheriff is rounding up a posse. But if we read the following line instead: "Jake told the chambermaid, 'My cattle are gone!' The chambermaid went to the saloon to find her pals Slim, Ernie, Baldy and Pete . . . ," we might be confused because the role member is acting in an unpredictable way (Black, Wilkes-Gibbs, and Gibbs 1982, 335). This violation of the role theme could be explained later, but notice that we do not *need* the action to be explained if is appropriate to the role member. Without this kind of information, a computer is unable to know whether certain information is redundant, necessary, predictable (and deleteable), and so on. When we consider presenting such an interpreting machine with

a text far more complex than these sorts of simplistic, stereotypical examples—such as a the paper of a high school student who critiques the ideology of a talk-show commentator by analyzing the false assumptions and faulty logic of his claims—we can begin to see the problems associated with creating machines that can reach any sound conclusions about the nature and quality of writers' prose, including the inability to judge how much information they have included relative to the knowledge of their audience, what kinds of logical chains they create, how their lexical and stylistic choices relate to their persona or ethos, and how appropriate that relationship is to the text's genre and context, or what informational path they lead the reader down in exploring or supporting a point.

Other interpretation programs ran into further problems, but the difficulties were not—at least theoretically—insurmountable. Based on the limited success of these early trial-and-error experiments, it seemed possible to create a sort of "mini world" programmed to account for hundreds of causal chains, the application of various scripts and plans, and so on, as long as the domain was limited and the computer had been given sufficient knowledge. In many ways, this is how current AI-based programs now "read" texts and provide certain types of assessments and feedback. This kind of limited application has at least some pedagogical potential because it works within a fairly stable domain with pragmatic purposes—practicing the textual process of summarizing a longer text, for example, where the "scoring" program has enough information about the longer text and the permutations of summary that it can determine the effectiveness of the student's attempt. (See Brent and Townsend, chapter 13 in this volume, for this type of application.) But the problem of assuming that, if given all this ability, machines might be able to interpretively extract something similar to what humans can ignores an essential characteristic of texts: that they are subject to multiple interpretations. A person reading an informative passage about pit bulls, in the domain of "domesticated canines," might see the text through the lens of being mauled by a pit bull as a young child. The results would be experientially different if instead of having been attacked, the reader had helped the family to raise prize pit bulls. For some texts, such as driving directions, a single, desired interpretation may be useful; but for most of the sophisticated texts that we want students to read, interpret, and produce, there is no "right way" for them to be read—a point thoroughly explored in reader-response theories (see Rosenblatt 1978 for a good theoretical introduction and Beach 1993

for an overview). Although machine-scoring advocates might argue that reducing interpretive possibilities takes a step in the direction of more consistent, reliable assessments of writing, to do so is to strip writing of its relationship with readers—that is, to turn it from writing into mindless bits of linguistic code.

One final concern of AI experts is worth mentioning. Computer programs usually operate *on* text, but they must be programmed to *learn from* it as well. As Schank points out, the AI programs he and his colleagues created had one serious flaw: "They could each read the same story a hundred times and never get bored. They were not being *changed* by what they read. People are intolerant of such boredom because they hope to profit in some way from their reading efforts. . . . To [change], an understanding system must be capable of being reminded of something it has stored in its long-term memory. But memory mechanisms are not random . . . . if we see every experience we have as knowledge structures in its own right, then thousands of structures quickly become millions of structures" (1984, 168).

The aim of computer-based text understanding is to produce a single output or assessment of the text's content and features; the machine can't read and interpret the text in the productively various ways that we want students to read and interpret, drawing on and applying an almost limitless fund of information, experience, and memory. (See the introduction, McAllister and White's chapter 1, Ericsson's chapter 2, and Jones's chapter 6 in this volume for descriptions of what the most common essay-rating systems are capable of doing and the levels—mostly surface—at which they do them.) This problem of computerized language processing is described quite simply by Wagman (1998): "Language-processing systems are constituted of structures that manipulate representations of objects and events. The constituted structures *do not understand natural language*, and their manipulation of representations accord to them the proper appellation of information-processing automata" (2; emphasis added).

Machines, in other words, are only machines.

IN SERVITUDE TO KNOWLEDGE: THE PROMISE OF COMPUTERS FOR THE ANALYSIS OF WRITTEN TEXT

While I have argued that the capacities of computers is nowhere near that of humans for reading and understanding natural discourse, the early experiments in artificial intelligence that helped to support that argument also reveal the potential for computer technology to serve as

an aid to research on language and writing development. Computers are far better suited to support advances in our understanding of human language processes than they are to relieve us of the need for human interaction so essential to people's learning and to the development of something as complex and socially determined as higher literacy. Computer analyses of text have yielded useful data that have furthered our understanding of many linguistic, textual, and even neurological processes. Machine analyses of style, for example, are well known, ranging from early studies of features in the work of specific writers such as Martin Luther King (Foster-Smith 1980) to documents such as the Declaration of Independence (Whissell 2002) to entire genres, especially for the purpose of document recognition and retrieval (see Kaufer et al. in press). In such increasingly sophisticated research, computers look for specific patterns in large quantities of prose and can correlate these patterns with other variables. In studies of the development of writing abilities, there is clearly much fruitful work to be done analyzing the prose of novice writers, especially longitudinally.

Correlational analysis facilitated by computers can also help us to understand the relationship between written language processes and other dimensions of human development, culture, psychology, and neurology. For example, in studies of women entering a convent in their twenties, Snowdon et al. (1996) found startling relationships in which "idea density" in the nuns' early writing, measured in part as a function of syntax, correlated with the results of cognitive test scores and the presence of Alzheimer's disease in the nuns' later lives, virtually *predicting* the development of the disease dozens of years before its onset. Content analyses also revealed that the nuns whose writing expressed more positive terms ended up living longer. Such new discoveries can be further tested on large numbers of texts using parsing, recognition, and content-analysis programs to identify specific features and variables, informing both neurobiology and language studies. Similarly, Campbell and Pennebaker (2003) have used Latent Semantic Analysis—a method commonly employed in machine-scoring systems—to relate stylistic features in subjects' personal writing to their overall health. In particular, they found that "flexibility in the use of common words—particularly personal pronouns . . . was related to positive health outcomes" (60). Changes in writing style across the subjects' texts were strongly related to wellness; the less the subjects' writing styles changed during a specific period, the more likely they were to visit a physician. The authors speculate that pronouns can be seen as "markers of psychological and physical

health, and, indirectly, people's thinking about their social worlds over the course of their writing." Their study provides "compelling evidence that the 'junk' words that people use in writing and speech reveal a tremendous amount about how they are thinking" (64).

As these and countless other studies show, the main advantage of computer technology for the analysis of text is not that it can do things that human readers or coders can't do—after all, the programs must be created to look for things that humans already know how to look for (though, as we have seen, human readers far outpace computers in terms of the *sophistication* of our reading processes). It is that computers can work through a single text hundreds or thousands of times, creating feature matrices, or they can examine tens of thousands of pages of text at lightning speed and, when programmed well, identify features with 100 percent accuracy, without the chance of human error. Studies have also shown that human readers can effectively "look" for only a few features at a time when they read, meaning that they must make many passes through the same texts to identify multiple features when asked to do so. Not so with computers, which can complete many tasks simultaneously without slowing down significantly. In time, cost, and accuracy for some tasks, computers trump human readers, which is presumably why there is so much interest in programming them to rate student essays on the basis of quality.

When used with large corpora of student texts, computers might provide us with information about student writing that has important implications for teaching and learning. Computer analyses could also yield relationships between such features and other aspects of students' education, such as their learning styles or attitudes. For example, in a study of approximately a thousand first-year engineering students' learning styles, preferences, study habits, and performance, several colleagues and I used sophisticated text-mining software developed by the SAS Institute to look for specific features in the students' weekly journal writing (Anson et al. 2003). As part of a one-credit Introduction to Engineering module, the students were required to write brief reflective electronic journal entries at a Web site. The entries focused on their learning and study experiences during their first semester of university life. Programmed to search for hundreds of potential relationships within and across twenty-seven thousand journal entries at lightning speed, the data-mining software allowed us to look for simple features such as word length or the use of punctuation as well as more sophisticated relationships between the students' texts and their learning preferences as

measured by the Learning Type Measure, a Myers-Briggs-type indicator. (Other work, notably Maid 1996 and Carrell and Monroe 1993, has also found relationships between learning style and features of student writing). Although it is not my purpose to report on the journal study here, it is interesting to note that among the simple measures, there was a relatively strong relationship between the average number of syllables in students' words (a crude measure of their lexical knowledge) and their grade point averages at the end of their first year of college. Female students used a statistically higher percentage of inclusive pronouns (*we, us*) than men, suggesting a profitable area for the continued study of gender variables in writing and learning. Other findings, such as the extremely low incidence of punctuation other than commas and periods across the twenty-seven thousand texts, remind us of the importance and determining influences of context and purpose: brief low-stakes journal entries instead of formal academic essays. Among the more sophisticated correlations generated by the SAS software, we found that the single most powerful textual predictor of first-year performance was the presence or absence of a single word: *physics*. Students who wrote about physics in their journal entries were much more likely to be in the high-GPA group than those who did not. This odd result reminds us as well that the results of computer analyses mean nothing until or unless humans can make sense of them in relation to other variables and aspects of the context in which the data have been gathered.

As computer programs further develop with insights from AI, linguistic theory, and areas such as computational semantics, there will be many opportunities to learn about written text and aspects of human physical, emotional, and cognitive development. Early proposals in the field of composition studies to use the insights of text analysis (such as cohesion, coherence, lexis, propositional structure, given-new information, and the like—see Cooper 1983) bore little fruit mainly because of the difficulty for human readers or coders to do the painstaking work of mapping such features across even small corpuses of texts. Technology now provides us with increasingly helpful ways to conduct such analysis, reopening abandoned pathways to new discoveries about the human capacity to write.

IN SERVITUDE TO LEARNING: FORMATIVE RESPONSE

The promise of digital technology to analyze human prose is not limited to research. To the extent that it can provide information to writers about their prose, it has some instructional potential. In this regard, it is helpful

to borrow a distinction from the field of assessment between *formative* and *summative* evaluation. Formative evaluation refers to information used in the service of improving performance, without any possible negative consequences for the person being evaluated; it is meant to bring about positive changes (Centra 1993). In contrast, summative evaluation refers to the assessment of performance after a period of time in which new knowledge, structures, or activities have been put into place; it is used to "make judgments about . . . quality or worth compared to previously defined standards for performance" (Palomba and Banta 1999, 8).

Computer-assisted formative evaluation of writing has not gained widespread acceptance or use, partly because the information it provides can be unreliable and one-dimensional and partly because the most sophisticated programs are not available for general pedagogical use. Simple feedback programs such as the sentential analyses provided by popular word-processing programs operate on text uniformly, without regard for the discursive community in which it is located, the intentions of its author, or the conventions expected by its readers. Indices such as sentence length or the use of passive constructions may have some limited use educationally in calling students' attention to certain linguistic features, but they fail to describe or respond to the relationship between such features and their appropriateness in certain kinds of discourse or the norms and expectations of certain communities or activities. More sophisticated programs, however, may be useful pedagogically to help students recognize textual or stylistic patterns in their writing and develop metacognition and metalinguistic ability in the improvement of their writing.

Pearson Technology's program Summary Street, for example, is a tool that purports to help young writers to learn how to summarize text more clearly and effectively. Students read passages and then try to capture the basic concepts, or the "gist," of the passages in a written summary. The computer then reads the student's summary, assigns it a score, and provides some boilerplate responses as well as comments on specific problems, such as misspelled words. Further attempts—for example, added information, clarified sentences, and the like—can show improvements in the score.

Any learner's earnest attempts to use such a system cannot be critiqued as intellectually bankrupt or of no pedagogical use. Many of the passages are interesting and well written, and the attempt to learn from them and summarize their contents requires rigorous intellectual and literate work. And, given the often horrendous workloads under which

many teachers of writing labor, especially in the schools, machine-aided practice and feedback for selected writing activities might provide some welcome relief. However, it takes only a little experimenting to reveal the limitations of such programs when compared with human readers and responders. After reading a passage about the ancient Aztec civilization that focused on their sacrificial practices, I wrote a summary designed to deviate from the original in noticeable ways. I located the sacrificial altar not at the top of the pyramid but inside a cave. Priests, not captives, were sacrificed in my summary, the purpose not to please the gods but pay homage to the peasants:

> The Aztecs believed that the peasants needed to be appeased constantly. As a result, they often sacrificed high priests to the peasants. They would take a captive into a dark and cavernous area they had hollowed out of the earth. There, to the sounds of beating drums and dancing, they would spear the priest with flaming swords.

In spite of the major differences in content between my summary and the original text about the Aztecs, Summary Street was unable to provide me with useful feedback. It questioned the line "A female peasant was then summoned from above," presumably because no females were mentioned in the original text (a relatively easy parameter to include in an assessment program). It flagged two misspellings (lower-case *aztec* and *disembowled*). In its final assessment, it assigned a high score to the summary, praising me for including so much extra information. Its canned response ended with the encouraging, "Good work, guest student!" Although the result took perhaps a millisecond to generate, the spurious response was in no way justified by its speed. A human reader, in contrast, would take a few minutes to read the summary but would offer a far more accurate assessment together with, if necessary, suggestions of far greater pedagogical value. (See McGee's chapter 5 in this volume for a similar and more extensive experiment using the Intelligent Essay Assessor—built on the same software that Summary Street relies upon.)

Such limitations of machine evaluation and response, of course, can be seen in the context of ongoing development: in a few years, programs may be sophisticated enough to simulate a fuller range of responses and judgments in domain-limited contexts for formative purposes. Considering the instructional potential of machine analysis of student prose, then, why should we object to the use of a machine-scoring program to determine students' writing ability summatively, for purposes

of accepting them to college, placing them in courses, or certifying their ability as they pass from grade to grade in the schools? If such applications can provide formative data to students, why shouldn't they provide to teachers, testers, and administrators some useful data about the student as product—like quality-control mechanisms on assembly lines? If all we really want to do with machines is look for a few simple measures, especially those measures that correlate with ability, even on a basic level, why not use them and avoid our own drudgery and human labor?

Although there is compelling enough evidence that computers can't read, interpret, and provide helpful feedback on a range of student texts in open domains, the answers to these questions take us beyond the potential for computers to enact those processes as effectively as humans and into the ethics of having machines read and rate students' writing to begin with. In the field of composition studies, scholars and educators have advocated purposeful, contextually and personally relevant occasions for writing, criticizing mindless, vacuous assignments and activities in a genre Britton and colleagues called "dummy runs" (1975). The rupture that machine scoring creates in the human activities of teaching and learning begins with the denial of a sentient audience for the students' work. Like Herrington and Moran (2001), whose experiments writing for a computer evaluator chronicled their disquietude with the rhetorical implications of not actually being "read" by anyone, I believe that this simple fact about machines as automata dooms them to failure in any contexts as politically, educationally, and ethically complex as testing students for their writing ability and using the results to make decisions about acceptance to, placement in, or exemption from a particular curriculum.

The point of writing in a course is for students to explore and reflect on ideas through language, convey their own interpretations and informational discoveries to others, and in the process intersubjectively create purpose and meaning. When they are aware of the subjugation of these human motives to an unthinking, unfeeling, insentient, interpersonally unresponsive, and coldly objective "reader"—even in a high-stakes testing situation admittedly already void of much intrinsic purpose—human communication is relegated to silence. This claim lies at the very foundation of the field of composition studies, traceable to its earliest commentary and theoretical work and infusing its scholarship ever since. In its genesis, research in the field showed that denying students purposeful contexts for writing had deleterious effects on their learning and on

the texts they wrote, and that the construction of purpose and audience is related to ability. Cannon (1981), for example, found that only the most proficient writers in the group he studied had developed higher-level purposes and showed a sense of ownership of and engagement in their texts. Similar conclusions were reached by Anson (1984), Emig (1971), Gage (1978), Newkirk (1984), Pianko (1979), and Nystrand (1982), who pointed out that without a full social context, writing "is not really discourse; it is [a] bloodless, academic exercise" (5). The centrality of human purposes and readers to written communication and to the development of writing ability is perhaps nowhere more important than in high-stakes assessment, since these rhetorical and social dimensions of writing appear to be so closely linked to performance. Largely unexplored empirically but of much concern to educators are the effects that vacuous contexts have on the manifestation of ability—a concern that shifts our focus away from whether machines can score writing as well as humans and toward what happens to students when they know they are not writing for flesh-and-blood readers. Until we know more about the psychological and compositional effects on performance of writing to and for computer readers/graders, we must proceed cautiously with their use in something as important and presumably humanistic as deciding the worth and value of people's writing.

CODA

As a reader, you have reached the end of a contribution to collective speculation about the subject of machine scoring. You have judged the validity of various claims, connected assertions and examples to prior knowledge and experience, affirmed and doubted, alternated between reasoned thought and emotional response. I have claimed that the process you've undergone cannot now, and probably not in the next several generations, be replicated by a computer, and that even if such a thing were possible, there is little point in doing so except for limited formative uses by developing writers. This and other contributions to the present volume, and continued national and international forums, conferences and meetings, published research, listserv discussions, blogs, and countless other opportunities for human interaction, will continue to create knowledge concerning writing development and instruction. Those contributions will proceed entirely without the responses, reactions, or ratings of machines, which are useful only insofar as they help *us* to make sense of our world and the nature of learning within it. For now, computers work best in servitude to the rich and

varied human interactions that motivate and captivate us. To substitute them for human work as important as the testing and judgment of other humans' literate abilities—grounded as they are in social relations and human purposes—is to assume that at least some dimensions of literacy are not worthy of our time.

And so to the machine, we do not risk affronting its sensibilities by telling it that *it* has nothing to offer this discussion, and that its rating is irrelevant. Or, more baldly:

> School is in, sucker. And U can't touch this.