Today, documents on nearly any topic on the web are accessible with a quick search. Type in a few keywords, and a staggeringly complex algorithm combs through data all over the globe in seconds. As far-reaching as these tools seem, they have a few major limitations. While they are great at finding content relevant to a given keyword or phrase, they can’t create an accurate summary of a piece’s content, explain its tone or intent, or find other subtle meanings that humans can extract at a glance.

“The general problem of search on the web is already solved,” states Ani Nenkova, associate professor in Computer and Information Science. “You can find articles on a given topic, but can you find only articles that are inspiring around a topic? Can you find articles that are well written and fun to read?” An ideal search method, she says, would pay attention not only to keywords, but also to deeper layers within a piece of writing. “You’d put in your usual query and then say, ‘I’m looking for general information on the topic,’ or, ‘I’m looking for detailed information on the topic,’ or even, ‘I’m looking for something that’s entertaining,’ and through this refine your result,” she adds.

From a computer science perspective, this turns out to be incredibly tricky. Narrowing a search by aesthetic preferences means creating software that can “read between the lines” to tease out what words really convey: emotional tone, narrative voice and the overall “quality” of a piece.

That last benchmark may be the most difficult to pin down. Aesthetics and quality are uniquely human ideas and remain so slippery to define that entire branches of philosophy exist to debate them. For a computer program to pick out “great” works, it would have to take into account the seemingly unknowable essence that makes Nabokov a master storyteller or Carl Sagan a legendary communicator of science—and put that into mathematical terms.

**Aesthetics and Quality Are Uniquely Human Ideas and Remain So Slippery to Define That Entire Branches of Philosophy Exist to Debate Them.**

Nenkova thinks she’s coming closer to solving the challenge. Language is, after all, a sort of code and in every code there are distinct and predictable patterns. By looking at enough written material of a specific genre, she says, statistical relationships begin to emerge in word choice and sentence structure, making it possible to sort works by subjective criteria.

**The Rules of Great Writing**

To test this idea, Nenkova and her former graduate student, Annie Louis, looked at works that are generally considered to be solid, clear and engaging, in the hopes of teasing out the informal “rules” good communicators follow when writing. “Annie and I started by looking at writing and journalism textbooks,” Nenkova recalls. “They all say things like, ‘make your writing visual,’ so we developed a rubric for scoring how visual a given word or topic was.”

With these textbooks as a guide, she identified a number of rules that are often followed, then used them to analyze more than a decade’s worth of science articles in *The New York Times*. 

Nenkova compared each of the pieces in the *Times* to a known source of quality work: an annually published anthology called *The Best American Science Writing*. Because pieces are selected for that publication by fellow science journalists, Nenkova reasoned that their judgment would serve as a proxy for “quality” writing in the field.

For the purposes of her study, if a piece appeared in the anthology, it was automatically deemed an example of “great” communication. Pieces by the same authors appearing in the *Times* were tagged as “very good,” and all other pieces in the newspaper were graded as “typical” writing. With articles lumped into these three quality “buckets,” Nenkova and Louis began looking for patterns that appeared in the works themselves: the frequency and complexity of words authors used, whether they used folksy colloquialisms, or if they relied on dense academic language, among other factors.

After analyzing hundreds of articles, the findings that emerged surprised her. Authors in each “quality bucket” used visual language—vivid, descriptive words—in predictable patterns. “Typical” writing used the most visual words overall, but “great” writing used those words more judiciously, sticking to specific categories or themes like “landscape” (trees, hills, mountains, lakes), or “shapes” (square, triangle, sphere, and so on). “In writing that stands out, there are fewer visual words but they exhibit a stronger pattern of organization,” Nenkova says. “They appear mostly at the beginning or end of the article rather than uniformly throughout, and stay within one coherent visual topic.” Unusual word combinations like “plasticky woman,” or “anti-spam operations” also appeared often in quality writing, she adds.

The biggest surprise, however, was that notable writing is lighter on detail than more pedestrian works, and is able to convey complex information without going into the weeds of a particular scientific finding. “It’s a lot easier to get excited about a topic when you’re not buried in the details,” Nenkova says.

Patterns like these are already helping Nenkova, who isn’t simply looking to grade journalistic works. She has her sights set on a bigger goal: uncovering the hidden linguistic “rules” that expert communicators use when writing. Knowing those rules, she notes, could have a far bigger impact than improving a search algorithm. It may actually improve people’s lives.

**A LOSS FOR WORDS**

For those with Autism Spectrum Disorder (ASD), communicating with others can be a daily struggle. They may have trouble telling appropriate jokes, understanding sarcasm, or choosing the right tone of voice to use with a boss or a close friend. Ultimately, these communication issues can lead to awkward and even disastrous social interactions, making it difficult to maintain a job or develop a romantic relationship.

Knowing when to use the right level of specificity, and whether to hold back on detail in some instances or add it in others, can be a factor in these exchanges, says Julia Parrish-Morris, research assistant professor of Psychology in Psychiatry at the Perelman School of Medicine and scientist at the Center for Autism Research at CHOP, who is currently collaborating with Nenkova.
“The idea of language specificity is what really intrigued me about Ani’s work,” she says. “If, for example, you’re talking to someone who knows a lot about the educational system, it makes sense to be very specific. But the decision about how specific to be calls on you to understand someone’s history and perspective, and to calibrate your level of interaction for that person.” That ability, she says, can be elusive for her ASD patients.

NENKOVA HAS HER SIGHTS SET ON A BIGGER GOAL: TO UNCOVER THE HIDDEN LINGUISTIC “RULES” THAT EXPERT COMMUNICATORS USE.

Teaching computers to identify clear communication might lead to software that helps people with ASD on a daily basis, such as a tool that offers subtle reminders of the right tone or specificity to use with a given audience, steering users through delicate social situations.

INFORMATION WITHOUT BORDERS

Difficulties in creating good communication aren’t limited to people with autism. Most of us at one point or another have struggled to be understood by our peers, even Nenkova herself. “I always liked reading books, but I was a terrible writer,” she admits. “I really wanted to know how great writers were able to move me.”

As a computer science professor, she’s starting to pick apart that question logically, and is helping students do the same. In the past, some of her undergraduates have adapted her tools to look at their own work, creating software that could gauge the clarity of their writing in real time. Others, like doctoral student Jessy Li, who will graduate in May, have applied Nenkova’s techniques to other languages, and are attempting to characterize good writing in Chinese as well as English.

Since every language and every culture has a different definition of how “quality” text should be organized, Nenkova and her students have their work cut out for them. “A Chinese person will have totally different interests when it comes to writing,” states Nenkova. “What does it mean to say what’s representative for that culture? We need more emphasis on understanding subpopulations to really tell.” The internet has no real borders, she notes, and a single search can turn up countless webpages in other languages, so being able to identify hidden traits that resonate with a given culture may help ease the spread of ideas worldwide. Although recognizing quality writing in other cultures is a tall order, she’s confident that the idea is gaining momentum.

“We’re making real progress with predicting text quality and style,” Nenkova says. “When I first started studying it several years ago, there was hardly anyone else working on the problem, and now it’s getting a lot more active. It’s amazing.”

By David Levin