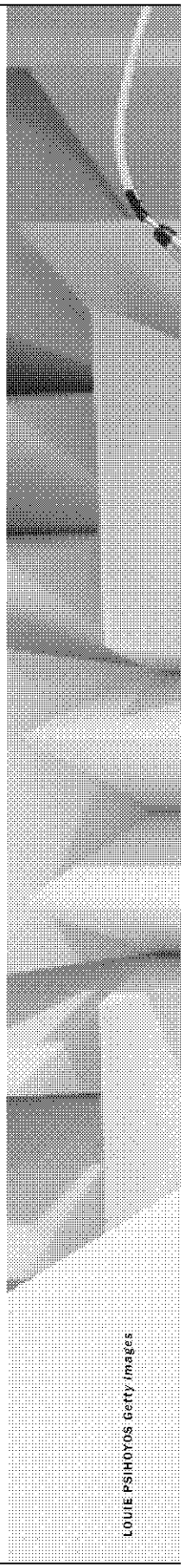
WHERE ARE THE TALKING ROBOTS?

Teaching a machine to speak has been a dream for decades. First we have to figure out how we know what we know about language

By Joshua K. Hartshorne



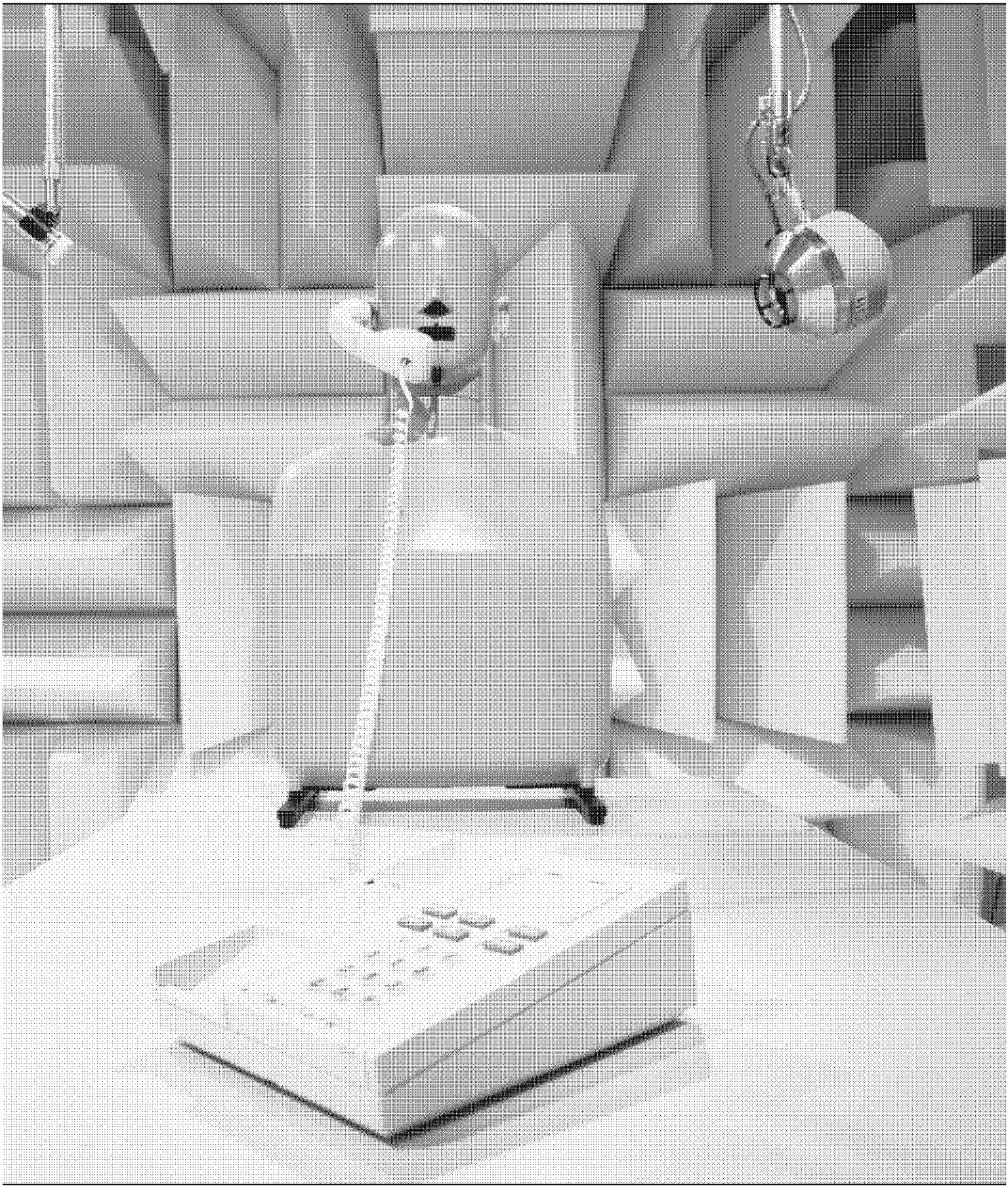
LOUIE PSIHOYOS Getty Images

Sulla, the world's first talking robot, was so adept at conversation—in four languages, no less—that a human visitor to the laboratory in which she was created refused to believe she was not a real person.

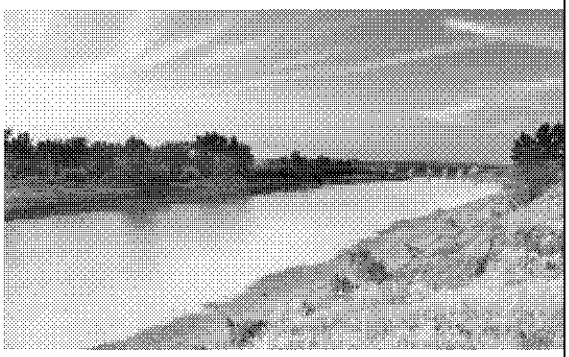
Alas, Sulla was not a real robot, either, but a character in Karel Čapek's 1921 play *R.U.R.*, which introduced the word "robot" to the lexicon. Ever since that debut, talking robots have seemed to be peeking around every corner, and not just in science fiction.

Almost as soon as modern computers were invented, researchers began to think about programming them to use language. In 1950 Alan M. Turing, one of the founders of computer science, predicted that by the turn of the century machines would be able to speak English so fluently that it would be difficult to tell a person from a machine—an achievement later dubbed the Turing test. Four years later a coalition of scientists at Georgetown University and IBM unveiled the 701 translation machine, which successfully translated 60 Russian sentences to English at the rate of two and a half lines per second, leading Leon Dostert, the researcher who dreamed up the technique used by the machine, to report confidently that fluent electronic translators were only "five, perhaps three years" off.

We are waiting still. After wave upon wave of optimistic prognoses followed by dismal failures, full-fledged talking robots seem no closer than other midcentury fantasies such as underwater cities and Martian colonies. If anything, the yearning for talking ro-



OUR ABILITY TO SOLVE TASKS **THAT WOULD SEEM TO BE** AS DISCERNING **THE MEANING OF SINGLE WORDS**—IS IN OF EVOLUTION. INDEED, WE ACCOMPLISH SUCH TASKS



Primed to speak: Research suggests that people quickly home in on the correct meaning of ambiguous words such as "bank" by taking cues from surrounding words: "swam" indicates the side of a river; "check," a financial institution.

bots is even more intense today because of our wish to replace the keyboard as our interface with digital services and ever smaller electronic devices.

Recent work in artificial speech has brought mixed results, giving us machines that can comprehend enough language to be useful (examples: Google Translate and the automated voice that answers your calls to customer service) while also confronting us with the limitations of the technology and its susceptibility to catastrophic failure (examples: Google Translate and the automated voice that answers your calls to customer service). Other projects are attempting to address these shortcomings by enlisting public participation via the Web so that we might learn more about how we choose our words.

But technology is not the only problem or even the biggest one: language has proved harder to understand than anyone had imagined. Our ability to perform such tasks as choosing the correct meaning of ambiguous words is in fact the fruit of millions of years of evolution. And we accomplish these feats without knowing how we do so, much less how to teach the skill to an artificial being. Indeed, as scientists try to codify grammar and tease out the subtle distinctions between similar terms, they are learning that meaning can be elusive and that the structure of language is a mystery even to we humans who have mastered it.

Old Rules, Broken

The earliest attempt to create talking robots was deceptively simple: to program them with the rules of grammar. This was IBM's strategy with its 701 machine, which was directed to translate Russian texts in its first public performance because of cold war interest in the Soviets. The 1954 press release introducing the project explains how the machine dealt with such language differences as word order. For instance, the English translation of the Russian *gyeneral mayor* is "major general." Whenever the machine encountered the Russian word *mayor*, its programming checked the previous word. If it was *gyeneral*, the 701 changed the order of the two words when it generated the English translation.

That such a straightforward system worked at all was partly because the 701 knew only 250 Russian words, so programming the machine to recognize every pair of adjectives and nouns in its database was not an onerous job. But many languages have hundreds of thousands of words, and English may have more than a million. If we make the reasonable assumption that half the words in English have multiple meanings, the programmer must consider 500 billion word pairs. At one word pair per second, writing the program would take nearly 16,000 years.

FAST FACTS

That Does Not Compute

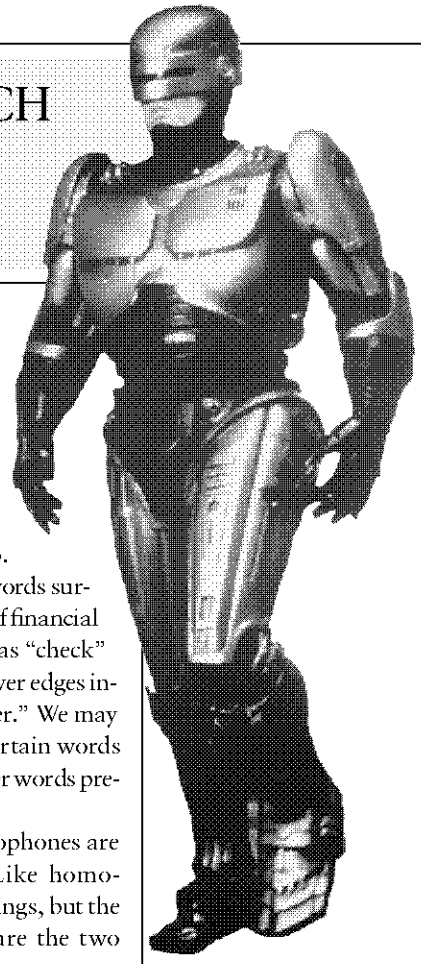
1. Programming a robot with the rules of English is difficult because we still do not know what all the rules are.

2. To help robots sort out ambiguity, scientists build language machines by feeding them billions of words tagged for meaning and parts of speech.

3. Researchers are using crowdsourcing on the Web to give robots a better sense of how human beings interpret and use language.

GILDO NICOLO SPADONI *Graphistock/Corbis (left)*; SAM CHRYSANTHOU *All Canada Photos/Corbis (right)*

COMPARATIVELY STRAIGHTFORWARD—SUCH FACT THE FRUIT OF MILLIONS OF YEARS **WITHOUT KNOWING HOW WE DO SO.**



As it happens, the phrase *gyeneral mayor* is actually an aberration—word order in Russian is generally similar to that in English, as opposed to, say, Spanish, where adjectives generally follow nouns. An apparent solution for a machine with a bigger vocabulary would be to program it with a rule such as "adjectives come before nouns in English and Russian but after nouns in Spanish" and append a list of rules for the exceptions. This strategy would not only vastly reduce the number of rules but also allow the system to handle new words. The problem is that the rules explaining the exceptions are likely to have exceptions, too. Although publishers of grammar books are loath to admit as much, scientists still have not found a set of abstract rules that fully explain English, Russian or any other language.

Yet the fragility of these systems lies not just in the imperfectibility of grammatical rules but also in the complexity of tasks as misleadingly straightforward as perceiving the meanings of single words.

Words of Many Meanings

One of the first problems encountered by a talking robot (and a talking robot's engineer) is that many of the words we use in everyday speech are homophones: they have multiple meanings. "Bank" can refer to either a financial institution ("John cashed a check at the bank") or the side of a river ("John swam to the nearest bank").

People quickly home in on the correct meaning when faced with such sentences. Psycholinguists Cyma van Petten and Marta Kutas of the University of California, San Diego, demonstrated this aptitude in a well-known 1987 paper about lexical priming—encountering a word primes people to process other words with related meanings. They found that just more than half a second after people come upon a homophone like "bank," only words related to the contextually appropriate meaning were still primed ("money" in sentence one above and "river" in sentence two).

This signature of normal processing breaks down in certain populations. In 2002 a team of neuroscientists led by Tatiana Sitnikova of Tufts University found that individuals who have schizophrenia fail to suppress the contextually inappropriate meaning of an ambiguous word: both "home run" and "vampire" were still primed more than a second after encountering "bat."

This work, though, tells us only that most people quickly resolve homophones by using context. The problem for the talking robot's engineer is that we do not know precisely how we do so. One theory is that we make use of the words surrounding the homophone. Discussions of financial institutions usually include words such as "check" and "cashed," whereas discussions of river edges include words such as "swam" and "water." We may simply have learned, in general, that certain words predict one meaning of "bank" and other words predict the other.

Even trickier to sort out than homophones are their cousins, polysemous words. Like homophones, polysemes have multiple meanings, but the meanings are closely related. Compare the two senses of "Jane Austen" in "Jane Austen wrote many books" and "I read some Jane Austen this afternoon." In the first sentence, the name refers to the author; in the second, to her work. Indeed, polysemy applies not only to all authors but also to all kinds of media. Rupert Murdoch has bought the *Wall Street Journal* (the company), and so have I (an individual issue).

Once again, context clearly matters, but the distinctions are subtle and difficult to define. Although the two senses of "bank" rarely appear in the same sentence, "Jane Austen" often appears in the same sentence as "*Pride and Prejudice*" whether the name refers to the person or her writing, so simple recourse to the surrounding words does not always work. How people discern the correct meaning is still not entirely clear.

Words such as "bank" and "Jane Austen" present a problem because they have several meanings. Pity the poor robot that has to sort out pronouns, which can have an almost limitless number of

(The Author)

JOSHUA K. HARTSHORNE is a graduate student in psychology at Harvard University, where he studies language and language acquisition. Read his blog at gameswithwords.fieldofscience.com.

ORION/THE KOBAL COLLECTION

EMILE WAMSTEKER *AP Photo (United Nations)*; PHIL BRAY *Touchstone/The Kobal Collection* (Bicentennial Man)

Found in translation: Scientists create language machines by feeding them huge bodies of text called corpora. Google Translate bulked up on a diet of United Nations documents that had already been translated into a variety of languages, helping resolve ambiguities.

meanings. In the sentence "I wrote *Pride and Prejudice*," the pronoun "I" refers to Jane Austen as long as it is Jane Austen who is talking. If the speaker is an actor playing Jane Austen (such as Anne Hathaway in *Becoming Jane*), then "I" refers not to the speaker but to the person she is playing. There is no simple rule. Third-person pronouns are even worse. In "She wrote *Pride and Prejudice*," the pronoun can refer to just about anyone female regardless of who is speaking. The robot cannot simply ignore these ambiguities, because without knowing who the sentence is about, the sentence hardly means anything at all.

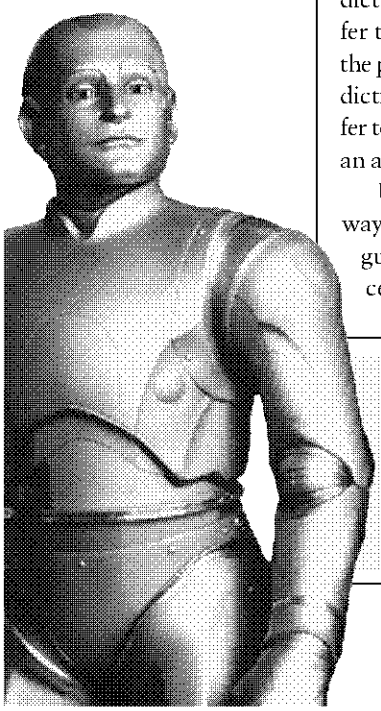
Perhaps the best-known model for resolving the pronoun conundrum is Centering Theory. Developed during the 1980s and 1990s by computer scientist Barbara Grosz of Harvard University and computer scientist Aravind K. Joshi and philosopher Scott Weinstein of the University of Pennsylvania, the theory comprehensively accounts for how sentences fit together in a broader discourse. It predicts that people use pronouns such as "she" to refer to the center—or most salient character—from the previous sentence, typically its subject. This prediction explains why people usually use "she" to refer to Jane Austen in the sentences "Jane Austen was an author. She wrote *Pride and Prejudice*."

Unfortunately for our robot, matters are not always so simple. In her 1998 dissertation, psycholinguist Jennifer Arnold estimated that only 64 percent of subject pronouns refer to the previous subject. Moreover, numerous studies going back to a seminal 1974 paper by linguist Catherine Garvey and neuroscientist Alfonso Caramazza of Johns Hopkins University have shown that the contextual cues for the human interpretation of pronouns can be maddeningly subtle. For instance, in work just submitted for publication, Harvard psychologist Jesse Snedeker and I reported that most people expect the pronoun in "Sally frightened Mary because she is strange" to refer to Sally but to Mary in "Sally feared Mary because she is strange." How people make these decisions remains unknown, but they do so rapidly. In 2007 a research team led by psycholinguist Jos van Berkum of the University of Amsterdam asked people to read sentences that did or did not follow the expected pattern, such as "Sally frightened John because she/he is strange," while their brain waves were monitored. The brain waves showed a telltale signature of extra processing when the pronoun did not match the overall sentence bias ("he" instead of "she" in the sentence above).

Bodies of Language

Given the bewildering nuances of words, scientists need to find ways to help robots make better predictions. Many have turned to the statistics of language, loading up their word machines with gobs of raw material and then working the numbers. They first feed their machines a huge collection of texts called a corpus—sometimes surpassing a billion words. The machine then breaks up the text into segments of *n* consecutive words, called *n*-grams. By looking at all the *n*-grams it has ingested, the machine learns which words tend to go with which other words. It comes to know, for instance, that the phrase "tall man" is fairly common in English (1,320,000 Google hits) and "man tall" is relatively rare (205,000 hits). Similarly, the machine might learn that in the majority of sentences in which "bank" is preceded by "swam," it means "river edge." The 701 actually worked with *n*-grams (more specifically, two-word bigrams).

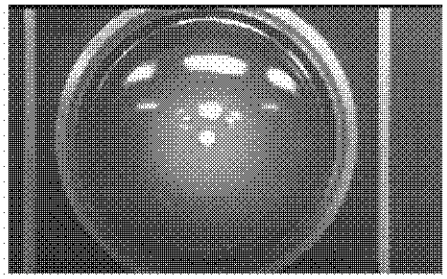
Statistical systems have significant advantages because the programmer does not need to formulate explicit rules such as "'general' comes before 'major'" or even abstract rules such as "adjectives



GIVEN THE BEWILDERING **NUANCES OF WORDS, BETTER PREDICTIONS.** MANY HAVE TURNED **WORD MACHINES WITH GOBS OF RAW MATERIAL**

Chatty Programs

A fixture of the artificial-intelligence world since the late 1960s, chatbots are computer programs that simulate conversations. The original program, ELIZA, could handle just a few topics. When it encountered a sentence it did not understand, it replied with a question ("Why do you say that...?") or dodged the topic entirely ("My head hurts"). Modern chatbots are considerably more sophisticated. Here is an excerpt from a conversation with Elbot (www.elbot.com), the winner of the 2008 Loebner Prize, a competition for chatbots:



HAL, from the Stanley Kubrick film *2001: A Space Odyssey*, was a chatbot of sorts with a malevolent streak.

ELBOT: Hi! I'm Elbot. I promise I won't blog about anything you tell me during this conversation.
HARTSHORNE: That's fine. I'm actually writing about this conversation for *Scientific American Mind*.
ELBOT: The inside of my head is plated with asbestos. Nothing gets out.
HARTSHORNE: I know the feeling.
ELBOT: My moods are controlled by an emotion processor which assigns the appropriate emotion for each situation.
HARTSHORNE: So how are you feeling now?
ELBOT: At this time, I am feeling inconsiderate.

Elbot clearly can sometimes produce natural conversations and certainly represents a considerable improvement since the original ELIZA. Elbot's speech is still riddled with non sequiturs, though, and even the occasional grammatical error:

HARTSHORNE: Did you see the Red Sox game today?
ELBOT: I'm not sure I ever saw that. I'm not even sure you can use in that context.

Elbot probably parsed my question as: Did you see + NOUN PHRASE. "I'm not sure I ever saw that" would have been a fine reply if I had asked about a movie. To get this conversation right, Elbot has to know what a Red Sox game is. Generating meaningful language requires more than a few rules of grammar. —*J.H.*

MGM/THE KOBAL COLLECTION

come before nouns." Statistical systems simply learn which words come before which other words. More complex implementations track information such as parts of speech, too, helping language machines learn that "check" is a better predictor of "financial institution" when used as a noun rather than as a verb.

Research also suggests that statistical learning—the ability to recognize patterns in the environment—may help humans absorb language, making the method particularly appealing to robot designers. A 1996 study by psychologists Jenny Saffran, Richard Aslin and Elissa Newport of the University of Rochester showed that even eight-month-old infants could learn trigram probabilities—the likelihood of trios of words or syllables to appear in sequence. The researchers had infants listen to strings of nonsense syllables like *bidakupadotigolabi*. The trigrams *bidaku*, *padoti* and *golabi* were all very common; others, including *dakupa*, were much less so. After hearing these nonsense strings for two minutes, the babies could tell the difference between the common and uncommon trigrams (they listened longer to the rarer ones, as if they were new); the authors interpreted the aptitude as evidence that children could learn word boundaries in this fashion. Similarly, in 2010 a team led by psychologist Christopher Conway of Saint Louis University found that people who are better at statistical learning are also better at making out speech under noisy conditions.

Although *n*-gram machines are not the only type of language system that scientists are trying out, engineers like using them because getting hold of large corpora is easy. Google, for instance, has published a Web corpus with more than a trillion words. But for corpora to sort out the subtleties of word mean-

SCIENTISTS NEED TO FIND WAYS TO **HELP ROBOTS MAKE** TO THE STATISTICS OF LANGUAGE, **LOADING UP THEIR** AND THEN WORKING THE NUMBERS.

RECOURSE TO SURROUNDING WORDS WOULD NEVER EXPLAIN **BOX WAS IN THE PEN" MUST REFER TO AN ENCLOSURE**, NOT A INSTEAD FROM OUR KNOWLEDGE THAT **BOXES DO NOT FIT**

ing and pronoun reference, the sentences must be tagged—that is, labeled with the definition or the part of speech of each word—and most basic corpora are not. The largest corpus tagged for meaning is SemCor (short for semantic correlation). Created at Princeton University, SemCor contains 360,000 words. That is a very large corpus measured by the effort needed to label all those words, but small for the purposes of the talking robot's engineer.

We can get a sense of the ensuing strengths and weaknesses of *n*-gram machines by looking at a pair of such systems developed by Google. One, a statistical translator called Google Translate, is fed a diet of documents that have already been translated into a variety of languages. (Google Translate's original fodder consisted largely of United Nations documents, which are issued in multiple languages.) Because a homophone in one language is typically represented by two words in another ("bank" is *orilla* and *banco* in Spanish), the bilingual corpora used to train statistical translation machines can stand in for a meaning-tagged corpus. The translator can learn to distinguish sentences containing "bank" in English and *orilla* in Spanish (most likely sentences with the word "swim") from those containing "bank" in English and *banco* in Spanish (sentences with the words "cashed" and "check").

A sea of meaning: Pity the robot that has to sort out pronouns. Words like "he" and "she" can indicate just about any person of the correct gender. Pronouns usually refer to the subject of the previous sentence, but only about two thirds of the time.

MOODBOARD/CORBIS

Google Scribe—a tool that predicts your next word as you type—is another variant of the *n*-gram machine designed to help generate sentences. Type "major," and it predicts the following: "role," "cities," "and," "role in," "problem," "histocompatibility complex," "league." All these are common combinations (even "major histocompatibility complex," which has more than a million Google hits).

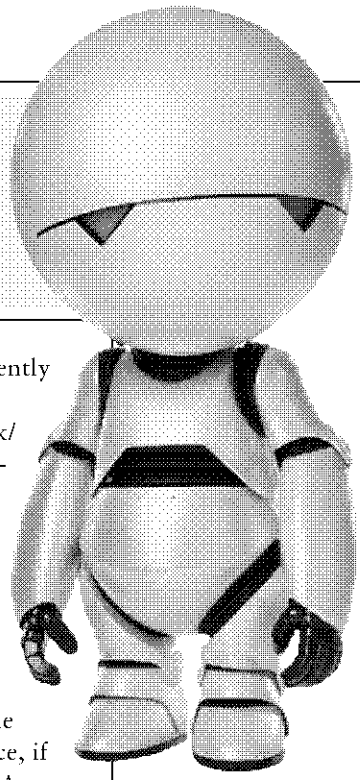
This abundance of possibilities points to a principal limitation of today's *n*-gram machines. Because they track contexts only a few words long, they break down if there is too much room between relevant words. Type in "He swam to the bank," and Google Translate returns *Él nadó basta la orilla*, which is correct. Try "He swam to the nearest bank," though, and you get *Él nadó basta el banco más cercano*, which means "He swam to the nearest financial institution." Bilingual corpora are also little help in sorting out polysemous words and pronouns. Many words that are polysemes in one language are polysemes in others.

Similarly, Google Scribe and other simple *n*-gram machines can neither handle new words nor generate useful sentences. Even young children can use new words in sentences, but Google Scribe makes no suggestions after you type in the coinage "wug." And because it learns the statistics only of short phrases, the sentences it produces are coherent word by word but ramble on nonsensically. For instance, type "Google" into Google Scribe and select the first suggestion it gives after each word, and you end up with "Google Scholar search results on terms that are relevant to the topic of the Large Hadron Collider at the European level and the other is a more detailed description of the invention." Such *n*-gram systems simply cannot relate the beginning of a sentence to the end.

Inching toward Talking Robots

One of the simplest ways to improve *n*-gram machines would be to have them use longer sequences. This task is more difficult than it sounds. Assume a

HOW PEOPLE KNOW THAT **"PEN" IN "THE** WRITING DEVICE; THE INFERENCE SPRINGS **INSIDE WRITING DEVICES.**



language contains only 10,000 words. To include every possible trigram, a word machine would have to learn a trillion combinations—10,000 to the third power. Storing every possible six-word sequence (still not long enough to do the job) would require 10^{24} combinations—about 10 trillion exabytes of information. All the digital information on planet Earth was reckoned in 2009 at only 500 exabytes.

But even if it had the backing of a gargantuan corpus tagged for meaning, the apt robot pupil would still need to absorb some street smarts before it could speak with authority. In a classic 1960 paper philosopher Yehoshua Bar-Hillel of Hebrew University argued that recourse to surrounding words would never explain how people know that "pen" in "the box was in the pen" must refer to an enclosure, not a writing device; the inference springs not from the context but from our knowledge that boxes do not fit inside writing devices.

To help give robots the benefit of real-world experience while bridging the data gap, several recent Web-based projects have sought to enlist the public. Computer scientists at Carnegie Mellon University, led by Anthony Tomasic, will soon launch an Internet game called Jinx. Two players are presented with a word in the context of a sentence (for instance, "John cashed a check at the BANK") and are asked to type related words as quickly as possible. They win points if they both come up with the same word. The researchers can use these guesses, particularly when the players agree, to label the meanings of the ambiguous words, creating a tagged corpus larger than SemCor.

My own Pronoun Sleuth (gameswithwords.org/PronounSleuth) is a Web site that asks volunteers to read sentences containing pronouns and decide to whom the pronoun refers, as in "Sally went to the store with Mary. She bought ice cream." For some sentences, agreement among the players is fairly strong; in others, less so. We have found that to distinguish one kind of sentence from the other, we need data from 30 to 40 people. At last count, more than 5,000 participants have judged several sentences apiece. Snedeker and I recently submitted a paper that had data for 1,000 sentences—a small number relative to what robots would need to sort out pronoun nuances, but it is by far the largest database of such sentences that is currently available.

Phrase Detectives (anawiki.essex.ac.uk/phrasedetectives), created in 2008 by computer scientists at the University of Essex in England, takes a more traditional approach, presenting players with a section of a book or article. When participants come across a pronoun, they are asked to identify the word to which the pronoun refers. Phrase Detectives also asks players about other referential expressions. The experimenters are interested, for instance, if players recognize that in the passage "Jane Austen wrote *Pride and Prejudice*. The book was very popular," "the book" refers to *Pride and Prejudice*. Thus far players of Phrase Detectives have completed work on 317 documents. Collectively, data from projects such as these will enable us to build and test theories that may lead one day to pronoun-using robots.

When, though, is an open question, and our expectations may be as unrealistic as ever. Despite understanding the obstacles, Franz Joseph Och, head of Google's machine-translation group, said in a recent interview with the *Los Angeles Times* that instantaneous speech-to-speech translation à la *Star Trek*'s universal translator should be possible "in the not too distant future." But building a talking robot will require understanding the secrets of language itself, which may prove just as elusive as anything else on *Star Trek*. **M**

(Further Reading)

- **Implicit Causality in Verbs.** C. Garvey and A. Caramazza in *Linguistic Inquiry*, Vol. 5, No. 3, pages 459–464; Summer 1974.
- **Statistical Learning by 8-Month-Old Infants.** J. R. Saffran, R. Aslin and E. Newport in *Science*, Vol. 274, pages 1926–1928; December 13, 1996.
- **Words and Rules: The Ingredients of Language.** Steven Pinker. Basic Books, 1999.
- **Paper Has Been My Ruin: Conceptual Relations of Polysemous Senses.** Devora E. Klein and Gregory L. Murphy in *Journal of Memory and Language*, Vol. 42, No. 4, pages 548–570; November 2002.
- **Shifting Senses in Lexical Semantic Development.** H. Rabagliati, G. F. Marcus and L. Pylkkänen in *Cognition*, Vol. 117, No. 1, pages 17–37; October 2010.
- **Child Language Acquisition: Contrasting Theoretical Approaches.** Edited by Ben Ambridge and Elena V. M. Lieven. Cambridge University Press, 2011.