"You Gotta Listen to How People Talk": Machines and Natural Language

Jacob Berger and Kyle Ferguson

Published in *Terminator and Philosophy: I'll be Back, Therefore I Am*, R. Brown and K. Decker (eds.), New York: Wiley & Sons, pp. 239-252.

Terminators are incredibly life-like machines. Not only do they look like humans, but they also have extraordinary knowledge of how to kill, how to protect, and how to use weapons. Beyond all that, they have incredible linguistic abilities. Remarkably, Terminators can communicate with human beings using natural languages like English. In *T2: Judgment Day*, the T-1000 doesn't just throw the pilot out of the helicopter during the battle at Cyberdyne Systems, he commands him to "Get out!" When Sarah tells the T-101 to keep their car at a certain speed, it understands the message and responds:

> **Sarah**: Keep it under 65.
> **T-101**: Affirmative.
> **John**: No no no no no, you gotta listen to how people talk. Now you don't say "affirmative" or some shit like that. You gotta say "no problemo." And if some guy comes up to you with an attitude and you want to shine them on, it's "hasta la vista, baby."
> **T-101**: Hasta la vista, baby.

Near the end of *T2*, we see that the T-101 has learned this particular language lesson, as it uses the now-famous phrase before shattering the frozen baddie, the T-1000. But John's remarks in the dialogue above are insightful. While it looks as though the T-101 has a working command of English, the machine struggles with certain aspects of the language as it's used in communication. Its diction is rigid and forced. Worse, it sometimes just doesn't understand what people mean. The T-101 communicates like, well, a robot.

When Skynet designed the Terminators, it must have operated under certain assumptions about the nature of language, meaning, and communication. These assumptions also shape our

approach to designing language-using machines in real world Artificial Intelligence research today. So the question is: How could we design a machine—that is, a computational system—so that it could produce and comprehend statements of natural languages like English, German, Swahili, or Urdu?[i]

In order to answer this difficult question, designers must face issues familiar to philosophers of language. Philosophy of language deals with questions like: What is language? What is meaning? And how do things like marks on surfaces (such as notes on paper or images on a computer screen) and sounds in the air become meaningful? What do you know when you know a language? What occurs in linguistic communication? What obstacles must be overcome for this kind of communication to succeed?

The answers to these questions make up what we'll call a *linguistic communication theory*. If Skynet had no linguistic communication theory, it could not have even begun to design or to program a machine that could use language to communicate or that could carry out missions in a linguistic environment.

Think about it. So much of our everyday experience is submerged in language. We look to signs to find our way around, we write reminders to ourselves of places to be and things to do, and we read newspapers to learn about events we've never witnessed in places we've never been. Weather, traffic, and sports reports pour from our radios, and the sounds of conversation fill up nearly every public space. It is rare, if not impossible, to find oneself in a social situation where language is absent. Skynet sent Terminators to this language-infused world and knew they would need to work their way around with words.

"My CPU is a Neural Net Processor": The Code Model and Language

So, what linguistic communication theory might Skynet have used when it designed its army of badass gun-toting, English-speaking Terminators? One obvious choice is a theory known as the *Code Model*.[ii] One reason why the Code Model makes sense as Skynet's theory is that the developers of the model, Claude Shannon and Warren Weaver, created it as a way to understand how machines, so to speak, communicate. Claude Shannon was an electrical engineer concerned with information transmission in circuit systems. Warren Weaver worked as a consultant to the United States Military and defense contractors to solve tactical problems, including how to make information-transfer more reliable on the battlefield.[iii]

According to the Code Model, the answer to the question "What is a language?" is that language is a kind of *code*—that is, a collection of signals and corresponding pieces of information. The answer to the philosopher of language's question, "What is meaning?" is that the meaning of a given signal is the *information encoded in the signal*. The answer to the question of "How does communication happen?" is that communication occurs when a signaler—the *producer* of a particular signal—encodes information into a signal, and, the receiver—the *consumer* of the signal—decodes the signal, thereby gaining the encoded information.

This may sound sort of complicated, but it's actually quite simple. Basically, the idea is that information is packed into a signal by a producer, the signal is emitted to and received by the consumer, and the consumer then unpacks the information from the signal. If all goes well, the consumer ends up with the same information that the producer originally sent. As long as the producer and consumer share the same code and no "noise" interferes with the signal, the successful transmission of information via signals—that is, communication—is guaranteed.

As an example of this, think of the early scene in *T2* when the T-101 tells John that the T-1000 is going to kill Sarah. John immediately attempts to leave in order to find her in time, but the Terminator grabs him. As he struggles to break free, John sees two guys across the street.

> **John** (to the two guys)**:** Help! Help! I'm being kidnapped! Get this psycho off of me!
> **John** (to the Terminator)**:** Let go of me!
> (The T-101 immediately lets go of John, who falls to the ground)
> **John:** Ow! Why'd you do that?
> **T-101:** You told me to.

Okay, so what's going on here? According to the Code Model, John's signal (the sentence, "Let go of me!") had certain information encoded or packed inside, and the Terminator, since it was programmed with the same code, was able to decode or unpack the signal and to end up with the information it contained.

So what did the T-101 do to decode John's signal? The Code Model suggests that it did two things. First, the Terminator recognized the sounds coming from John's mouth as signals. Then, it retrieved information matching these signals from its neural net processor (its mind, so to speak). In order to do this, the Terminator would need to be programmed with what linguists call a *lexicon* and a *syntax* of a given language. A lexicon is a complete set of meaningful units of a given language, usually words. Think of a lexicon as a "dictionary" of a code, a dictionary that matches individual signals with bits of information or words with their meanings. Syntax (or syntactical rules) specifies how items from the lexicon are combined; this is what people usually think of as "grammar." By recognizing the lexical items and the syntax of the sentence, the Terminator was able to decode the signal and receive the information it contained. And, since the Terminator was programmed to do as John commands, it let John go… *literally*.

We can now return to our initial question. How do we design a machine that can produce and comprehend statements of a natural human language? If we accept the Code Model as our

linguistic communication theory, we can give an elegantly simple and straightforward answer. All that Skynet needs to do in order to ensure that its army of man-destroying Terminators is capable of understanding and producing English sentences is simply program into the Terminators' neural net processors the lexicon and syntax of English. It's that easy. If the T-1000 has the lexicon and syntax for some language, it should be able to understand when people beg it not to kill them and then make quips right before it shoves stabbing weapons into their brains.

Why the Terminator has to Listen to How People Talk

Our guess is that Skynet did indeed use the Code Model as its linguistic communication theory when it designed the Terminator.[iv] But this is not saying that the Code Model is a good theory of linguistic communication. In fact, it's a flawed theory, failing to capture how people communicate using language. Its shortcomings explain why the Terminator just isn't so hot at sounding like a normal English-speaking human, and why it sometimes doesn't grasp what normal English-speaking humans mean. The T-101 says, "Affirmative" when he probably should say, "No problemo," and it drops John on the ground when he tells him to let him go, when he probably should have just set him down. The Terminator fails where the Code Model fails.[v]

The problem is that linguistic communication isn't as straightforward as the Code Model says it is. Basic obstacles arise when people stumble over words, run words together, speak with accents, mumble, and more. Schwarzenegger's thick Austrian accent makes it hard for the American movie-watcher to understand what he says. If audiences had to make out every word that Arnie said in order to understand him, the better part of *T2*, most of the original *Terminator*, and every single one of his gubernatorial speeches would be nearly incomprehensible. Just watch *Kindergarten Cop* again if you need to refresh your memory.

The Code Model regards these sorts of problems as *noise*. Accents, mumbling, and other imperfections are like "static" that corrupts or interferes with the signal and makes it hard for the consumer to acquire the information it contains. We bet Skynet could have designed Terminators so that they could deal with this sort of noise.

But, deeper problems than noise abound for the Code Model. Put simply, the word-meanings and order of the words of a sentence are rarely, if ever, *enough* to give an interpreter access to what a speaker is trying to communicate. For ease, we'll refer to all of these sorts of complicating features as the *pragmatic* aspects of language.[vi] Let's consider some examples.

Pragmatic aspects of natural languages include, for instance, *lexical ambiguity*. Consider the quote we reprised at the beginning of the paper. John says to the Terminator, "If some guy comes up to you with an attitude and you want to shine them on, it's 'Hasta la vista, baby.'" The verb phrase "to shine" has multiple meanings. It can mean "to polish," "to emit light rays," "to excel," and other things. In this case, John uses it as slang to mean "to give someone a hard time." Linguists call words or phrases that have multiple meanings *lexically ambiguous*.[vii] If a person says a sentence that includes a lexically ambiguous word or phrase, it's not always clear how to interpret that sentence. How is the Terminator supposed to know whether John's sentence means that the Terminator is supposed to say "Hasta la vista, baby" to people to whom it wants to give a hard time, or if it means that the Terminator should say the sentence to people on whom it wants to shine a flashlight? Lexical ambiguities make trouble for the Code Model because hearers have no way of resolving an ambiguous signal by appealing to the code itself. If the Terminator were simply assigning pieces of information to John's signal, there'd be no basis for it to choose one assignment over another.

Another pragmatic obstacle is *syntactic ambiguity*. There's a scene in *T2* where John tells the T-101, "You can't keep going around killing people!" This sentence is *syntactically ambiguous* because, given the way the words are arranged, there are at least two acceptable interpretations of it. On the more natural reading, John is claiming that it is not permissible for the Terminator to kill people. On a slightly less natural reading, John is claiming that it *is not* permissible for the Terminator to go around and kill people, though it *is* permissible for the Terminator to kill people as long as he's not *going around* while he kills. If the Terminator is going to understand this sentence, it must disambiguate it. But it's important to note that *each* reading is acceptable given the Code Model because there is nothing contained in the sentence itself that would support one interpretation over the other.

Two more pragmatic issues with natural language are *referential ambiguity* and *underdetermination.* A sentence exhibits a referential ambiguity when it is not clear from the meanings of the words of the sentence all by themselves what a word or phrase in that sentence *refers to*. Recall the scene late in the movie where the T-1000 is chasing John, Sarah, and the Terminator in a tractor-trailer carrying liquid nitrogen. The gang is in a junker that they stole from a man on the street. The T-1000 is gaining on them and John screams, "Step on it!" In this sentence, the word "it" is *referentially ambiguous*. Naturally, we all know that the 'it' refers to the gas pedal of their vehicle. John wants the Terminator to step on *the pedal* and speed up the vehicle. Again, there's nothing in the signal that provides the referent of "it," so the Code Model doesn't explain how the Terminator is supposed to understand what "it" refers to.

Underdetermination occurs when a fully decoded sentence doesn't provide enough evidence on its own to figure out what a speaker means. Even if you had a souped-up version of the Code Model, one that could resolve the above ambiguities, underdetermination might still be

a problem. To see why, consider the scene discussed earlier when Sarah, John, and the T-101 are driving out to their gun-supplier and they want to avoid the police. Sarah, not wanting the Terminator to speed, says to the Terminator, "Keep it under 65." What, in this situation, does "under 65" mean? If the Terminator were just retrieving from its neural net processor the individual meanings of "under" and "65," it would be very hard to see what Sarah is asking for. Does Sarah mean that the Terminator should keep the car under 65 *years old*? Under 65 *pounds*? Under 65 *dollars*? Under 65 *degrees*? These options make little or no sense. Again, obviously, we all know that that Sarah wants the Terminator to keep the car's *speed* under 65 *miles per hour*. But, the words "speed" and "miles per hour" are nowhere to be found in her sentence. What Sarah means is *underdetermined* by the sentence because, whereas we people of course understand what Sarah meant, we need to add more to the sentence than the words provide.

Skynet Doesn't Want Them to Do Too Much Thinking: The Inferential Model

> **T-101:** Skynet pre-sets the switch to read-only when we're sent out alone.
> **Sarah:** Doesn't want you doing too much thinking, huh?

To understand most sentences in a natural language, we must overcome some or all of these pragmatic obstacles. If the Code Model were accurate, hearers wouldn't be able to make sense of utterances involving any pragmatic features because this model only requires that a hearer process the words of an utterance. How would we interpret spoken words if we were simply left with the words and word-order of the sentences alone? If people were programmed to interpret others' utterances in just this way, we would have a really hard time understanding one another. But, in real life, we seem to solve these pragmatic problems of ambiguity and underdetermination, and get around pretty well.

So if the Code Model is inadequate, what linguistic communication theory best mirrors what we in fact do? In place of the Code Model, many philosophers of language and linguists have advocated the *Inferential Model* of communication.[viii] Here, having the lexicon and syntax of a language is not enough to figure out what speakers mean when they are communicating. Instead, hearers must use this information as one piece of evidence among many other pieces of evidence, to *infer* what speakers mean. It's not a matter of unpacking information from a signal; it is matter of *working out* what a speaker means by appealing to a wider context like shared knowledge and assumptions in addition to the meanings of words.

What exactly, then, are hearers inferring? In answering this question, the Oxford philosopher H. Paul Grice revolutionized the philosophy of language and linguistics. In his two famous essays, "Meaning" and "Logic and Conversation,"[ix] Grice distinguishes what a *sentence means*, on the one hand, from what a *person means* by using that sentence on a particular occasion. What a sentence means, according to Grice, is something like what the Code Model suggests; you might think of it as the *literal* meaning of the sentence. We will call what a sentence means its *sentence-meaning*. The sentence-meaning of "The Terminator is a killing machine" is that the Terminator is a killing machine. This means that sentences with pragmatic obstacles such as underdetermination may not have a sentence-meaning at all, or at least not a complete sentence-meaning.[x]

What a person means by using a sentence on a given occasion often greatly diverges from what that sentence means. We'll call what a person means by saying a sentence on a given occasion the *speaker's meaning* of that utterance.[xi] If Sarah were to ask John if she thought the Terminator would be able to complete its mission, John might respond, "The Terminator is a

killing machine." In that case, the sentence-meaning is that the Terminator is a killing machine, but the speaker's meaning is something like, "Sure, the Terminator can complete its mission."

Grice's distinction is quite plausible. Recall the scene we discussed in which when the Terminator picks John up and John cries, "Help! Help! I'm being kidnapped! Get this psycho off of me!" According to Grice, the sentence-meaning of John's utterance is probably something like, "Assist John in removing himself from the psychologically disturbed individual holding John!" Because the Terminator is operating according to something like the Code Model, it is likely that this is what it interprets John as meaning.

But, as any good Gricean knows, speaker's meaning often far outstrips the sentence-meaning of that speaker's actual words. So John probably means something more akin to, "Terminator, I want you to let me go." Since the T-101 is under the sway of the Code Model, it does not catch onto John's speaker's meaning, and continues to grapple John. Only when John explicitly exclaims "Let go of me!" does the Terminator react and drop John to the ground.

According to Grice, John *did* mean that the Terminator should let him go with his initial statement. The sentence-meaning of John's initial utterance isn't that the Terminator is to let him go, but the speaker's meaning of it surely *is*. His first utterance was directed toward someone else, true, but it provided evidence of his desire to be released. So, if John did tell the T-101 to let him go at first, why did it take the second, more explicit, utterance to get the Terminator to release him? Because the speaker's meaning *and* sentence-meaning of John's second utterance both were to the effect that the Terminator let him go, but only the speaker-meaning of John's first utterance had that meaning. If the Terminator were designed according to the Inferential Model, it would have been able to *infer* John's speaker-meaning from the first utterance. And, notice that even when the T-101 interprets John's second utterance, he still seems to fall short of

John's speaker's meaning because he complies only with its literal meaning. That is, the T-101 literally lets John go, dropping him to the ground, when it's obvious to normal English-speakers that what John meant was for the Terminator to set him down gently.

Now, speaker's meaning exists only in the speaker's mind. This means that we have to guess at what people believe, desire, intend, wonder, and all the rest. These constitute or determine what a speaker means. But, our access to these mental items is forever indirect, mediated by the speakers' publicly observable behavior. From our observations of another's behavior, we infer what that person believes and desires. In other words, we are able figure out what it's like on the *inside* by using external clues.[xii] Language-using behavior is no different. A particular sentence is one clue among many pieces of the puzzle that we must put together by way of inference. These inferences are rarely, if ever, conscious, so it may not seem to us that we're making them. But that's okay. They're still happening.

How to Make the Terminator Less of a Dork

In one of our favorite scenes in *T2*, John asks the Terminator whether it could, "you know, be more human and not such a dork all the time?" So, if we wanted to make the Terminator's communicative behaviors more human-like, we would want to build its capacities to process language according to the Inferential Model. In that case, we would need to supply the machine with more than just the lexicon and syntactical rules of a given language. Clearly, we would need to also program it with a great deal of information about human psychology.[xiii] It would need to have a mechanism, or more likely several mechanisms, that could piece together lots of information from the environment and about people in general to solve the problem of reading others' minds.

The goal of human audiences is to infer speaker's meanings behind linguistic behavior, not the mere sentence-meanings. To complete our interpretative tasks, we exploit all sorts of evidence including the speaker's gestures, tones, facial expressions, locations, psychological facts about what they believe and know, their goals and expectations, and more. We use all of this coupled with the word-meanings and sentence structures to infer what speakers mean. We know John uses "it" to refer to the gas pedal when he screams "Step on it!" because we know his goals and we know what it would take to accomplish them in this situation. We don't reach this conclusion working from the words alone.

While we've focused mostly on language comprehension or interpretation, much of what we say goes for language production as well. In order to comprehend, a hearer must rely upon their beliefs or assumptions about a speaker's psychology. This also is true for speakers. Speakers use their assumptions about hearers when they select their words. We don't say more than we have to; we don't inform people of what we think they already know. Rather, we say what we think would be relevant to our hearers, given what we think they believe and what we are trying to accomplish. We say just enough to get our points across. Only a machine that's aware of environmental facts and human psychological facts—namely, the same kinds of facts that hearers exploit to infer speakers' meanings—would be capable of knowing what to say in particular situations.

Only a linguistic communication theory that accommodates pragmatic aspects of language would make the Terminator less of a dork. We think the best theory we have going currently is the Inferential Model. So here's a suggestion and a request for Artificial Intelligence researchers and Skynet, if and when it comes online: use the Inferential Model in your machines, but please don't use their linguistic prowess to hasten Judgment Day.[xiv]

[i] Natural languages like these are importantly different from formal languages, such as the formal languages of mathematics or logic. Natural languages develop, as it were, naturally over time in human communities, and are mainly used to communicate between language-users. Formal languages, on the other hand, are constructed artificially with other, usually non-communicative, ends in mind.

[ii] The term "Code Model" first appeared in D. Sperber and D. Wilson, *Relevance: Communication and Cognition*, (Cambridge: Harvard University Press, 1986). The model received its first formal treatment in W. Weaver and C.E. Shannon, *The Mathematical Theory of Communication* (Urbana: University of Illinois Press, 1949).

[iii] See, for example, C.E. Shannon, "A Mathematical Theory of Communication," *Bell-System Technical Journal,* 27:3 (1948), pp. 379–423; 27:4 (1948), pp. 623–656 and Weaver and Shannon, *The Mathematical Theory of Communication*.

[iv] Disclaimer: despite what we or our more delusional or conspiracy theory-minded readers might think, *T2* is unfortunately not a documentary. It is a big-budget, action-packed Hollywood blockbuster. So, we think that the Terminator is generally working under something like the Code Model. There are, of course, instances in the film where it might seem otherwise. Nobody's perfect or perfectly consistent. James Cameron does come close…

[v] At least superficially, T-1000 is far more fluid in his conversational abilities. Before stealing a man's motorcycle, it says smugly, "Say, that's a nice bike." Maybe Skynet changed its approach.

[vi] The term "pragmatics," as it relates to linguistic theorizing, originated in C.W. Morris, *Foundations of the Theory of Signs*, (Chicago: University of Chicago Press, 1938). Morris defined the term it as the study of conditions and effects surrounding a system of signs, and how that system relates to its interpreters. Linguistic pragmatism should not be confused with American pragmatism, a philosophical movement and outlook developed by Charles Saunders Peirce, William James, and John Dewey.

[vii] For more on ambiguity, see "Terminating Ambiguity: The Perplexing case of 'The'" by Richard Brown in this volume.

[viii] The term "Inferential Model" comes from Sperber and Wilson's *Relevance*.

[ix] Both are reprinted in H.P. Grice, *Studies in the Way of Words,* (Cambridge, Mass.: Harvard University Press, 1989).

[x] There is a lot of debate about whether or not there is such a thing as sentence-meaning. For an example of one who denies that sentences involving pragmatic features have any such thing as sentence-meaning, see F. Récanati, *Literal Meaning*, (Cambridge: Cambridge University Press, 2004). For an example of one who claims that there is a meaning, albeit an incomplete one, see K. Taylor, "Sex, Breakfast, and Descriptus Interuptus." *Synthese*, 128 (2001), pp. 45-61. Some thinkers claim that most sentences, regardless of whether they involve pragmatic features, have compete sentence-meanings; see H. Cappelen and E. Lepore, *Insensitive Semantics*, (Malden, MA: Blackwell, 2005).

[xi] Sometimes literal meaning is referred to as the *semantic content* of an utterance. The semantic content of an utterance is thus distinguished from whatever else is supplied by a speaker—namely, the *pragmatic content*. Where exactly to draw the distinction between semantics and pragmatics is a hot topic in contemporary philosophy of language.

[xii] For a look at how philosophers have used this inferential perspective to decide whether machines think or not, see Antii Kuusela's chapter in this book.

[xiii] In *T3*, the T-101 says that it has been programmed with some basic knowledge of human psychology. Prior to *T3*, there is no indication that it has such knowledge. The knowledge the T-101 claims to have in *T3* appears to deal only with emotions and their effects on behaviors. The knowledge does not seem to include information about cognition or thought, let alone knowledge of how to infer the content of others' thoughts.

[xiv] We would like to thank Marc Berger, Elizabeth Berger, Sam Berger, and Kristen Lee. We especially thank the editors of this volume for their very helpful comments.