Linguistics and national security

Written by Ignatius Mattingly

Edited by Betty Birner

How do U.S. intelligence agents decode messages in foreign languages?

The U.S. government relies on linguists for much of its expertise in **communications intelligence** — that is, the collection and transmission of information that's important for national security. Our country's messages must reach their destinations without being intercepted and understood by any other country; at the same time, we want to be able to interpret any messages that we intercept from other countries concerning espionage, military build-ups, or other activities that could threaten the United States. For this reason, the government hires language specialists to translate, analyze, and summarize intercepted messages. The grammars and training methods they use are the products of linguistics research.

Detailed descriptions of individual languages are important for **cryptanalysis**, the study of coded messages, because the structure of a language limits the kinds of possible solutions there can be for a code based on that language. For example, to solve a simple code based on English, we would first look to see which words and letters occurred most often in the coded message. We could guess that the letters occurring most often might represent *e, t, a, o, i, n,* or *s,* since these are the letters that occur most often in English words. Small words that show up frequently might be words like *the, and, is,* or *are.*

Of course, which words and letters are most common varies from language to language. So statistical information on word and letter frequencies (or sound frequencies, for spoken messages) is very useful for code-breaking. Information about the sentence structure of a language is also important. In English, for example, we might expect a sentence to begin with a noun phrase (like *the man*), followed by a verb (like *ate*), possibly followed by another noun phrase (say, *the peach*) or perhaps an adverb (like *hungrily*). Other languages have very different word orders; in many languages, for example, the verb comes at the end of the sentence. Knowing this information about a language helps cryptanalysts to figure out what kind of element each part of a coded message is likely to represent.

Although we have grammars describing most of the languages that are currently of concern, we can't assume that these are the only languages we will ever need to know about. A minor nation can take on political significance very suddenly, and when that happens it can be important to have a grammar available for its language (or languages). Because there are thousands of different languages in use in the world, developing these grammars is an enormous task.

Even when a great deal is known about the language in question, cracking a code is extremely difficult. Many codes are set up so that the letter-to-letter correspondences change repeatedly throughout the message, and there are often two or more layers of encoding. The more sophisticated the coding system, the more important it is to have detailed linguistic information to use as a starting point. Right now, this sort of statistical data is only available for a few languages, and it's generally not in a very useful format. It would be in the national interest to analyze written and spoken materials in a much wider range of languages. This is where computers can help.

For written texts, the first task is to pull material of interest out of a much larger amount of irrelevant material. To attack this problem, the methods of two branches of computational linguistics, information retrieval and machine translation, can be combined. A computer can be programmed first to find the material that is wanted (say, by looking for certain key words), and then to translate it into English. Writing these programs requires extremely detailed descriptions of both languages, and the results so far are crude compared to the skills of human analysts. Still, such programs are becoming more and more helpful as they become more sophisticated.

For spoken material, such as intercepted phone calls, computers need to be programmed to recognize speech. For this the methods of another sort of linguist, the acoustic phonetician, are needed. Every language uses a different set of sounds and sound combinations. For example, the *th* sound in English (as in *think*) doesn't exist in Russian; on the other hand, Vietnamese allows words to begin with the sound *ng* (as in *song*), but English doesn't. The more we know about the sound patterns of a language, the better we can program our computers to understand speech in that language.

Why is it so important to intercept and understand other countries' messages?

Here, history speaks for itself. Two examples will give you an idea of how important these messages are for our country's safety.

In 1943 the U.S. Army began to study thousands of coded 'diplomatic' telegrams they had intercepted from the Soviets. After three years of analysis, they had decoded enough to see that these telegrams weren't just diplomatic messages; they contained information about Soviet spy activity in the U.S. The effort to decode these messages went on for years, and dozens of linguists were recruited to help with the translation and decoding.

The messages they were eventually able to decode gave detailed reports on the activities of Soviet agents in the U.S. They revealed, for example, that someone in the U.S. War Department was providing secret information to the Soviet Union. The decoded messages also contained reports on the Manhattan Project (U.S. development of the atomic bomb), including a list of U.S. scientists working on the bomb. In short, these messages revealed a vast network of Soviet spy activity in the United States.

Later on, in the early 1960s, communications intelligence played an important role in resolving the Cuban Missile Crisis. As Soviet ships brought cargo to Havana, their intercepted and decoded messages helped the U.S. to realize that the Soviets were building up arms, combat aircraft, and other military equipment in Cuba. Late in 1962, the Soviets brought surface-to-air missiles into Cuba, and shortly after that, ballistic missiles. When the U.S. imposed a 'quarantine' of Cuban ports, decoded messages showed that Soviet ships had stopped in their paths rather than violate the quarantine. Soon after that, the Soviet Union agreed to take the ballistic missiles out of Cuba.

How does the U.S. ensure the security of its own messages?

Needless to say, communications intelligence agencies are also concerned with the privacy of U.S. messages. As important as it is to break the codes of other countries, it is equally important to develop codes that other countries cannot break. Here again, language specialists can be vital to the success of American military efforts.

One unusual case is that of the Native American Code Talkers. Late in World War I, American troops found that German officials were listening in to their telephone conversations. One U.S. regiment stationed in France included a number of Choctaw Indians, so they used the Choctaw language to transmit information about a delicate withdrawal of troops. The Germans were taken by surprise, and the withdrawal was a success. After that, more Choctaw Code Talkers were added. Choctaw didn't have words for all of the military equipment they needed to talk about, so they substituted common words, such as *big gun* for 'artillery', and *little gun shoot fast* for 'machine gun'.

Code Talkers were used again in World War II. The Comanche, Choctaw, and Navajo languages were used most often, with about 420 Navajo Code Talkers serving in the war. Since these languages are very different from European languages, the enemy found these 'codes' hard to crack. One U.S. report described the Navajo language as "the simplest, fastest, and most reliable means we have of transmitting secret orders via radio or over telephone circuits exposed to enemy wire tapping."

Today, of course, the government relies heavily on computers for the storage and transmission of sensitive information, and cryptanalysis is largely a question of programming computers to encode and decode information. When a file is saved, or an e-mail message is sent, such a program automatically encodes it to protect it from prying eyes, and stores it in coded form. These coding schemes use complicated mathematical formulas to protect the privacy of computer files, phone conversations, and electronic mail. As new and more complex coding methods continue to be developed, it remains to be seen what role cryptanalysis will play in the electronic communication of the future.