Close Enough to Human: Artificial Intelligence and the Future of English Education

Tom GALLY

The application of new technology to language education has always been hit or miss—perhaps more often miss than hit.

As a language learner and teacher, I have seen many technologies introduced, hyped, and adopted with great fanfare, only to be marginalized or abandoned. When I was a child in the 1960s, my parents owned sets of LP records for learning Spanish and French. I put them on the turntable once or twice and listened to a few lessons, but I don't remember anybody in my family ever studying from them seriously. At that time, records and reel-to-reel tapes seemed to revolutionize the learning of languages. Before audio recording and radio, the only way to hear the sounds of a foreign language had been to be in the presence of a speaker of the language. Recording and broadcast technology seemed to be a great advance: We could be in our home in Southern California and hear the sounds of Spanish as spoken in Madrid or French as spoken in Paris. Presumably those LPs would enable us to learn the language better and more easily. At least in the case of my family, though, they didn't.

When I entered college in 1975, the audiolingual approach to language learning and teaching was still dominant. I enrolled in introductory Chinese. In addition to our five fifty-minute classes per week taught by teachers from Taiwan, we also had to go to a language laboratory for several hours a week. There we would sit with headsets on, listening to and repeating reel-to-reel recordings of the consonants, vowels, and tones of Mandarin and of sentences being read aloud from our textbook.

All those hours in the language laboratory did, I think, improve my Chinese pronunciation, though they did not prevent me from abandoning the study of Chinese two years later.

And eventually that audiolingual approach to teaching foreign languages fell out of fashion altogether. I am not aware of any universities in Japan, at least, where English students still go to a language lab and listen to and repeat recordings. One reason is that theories about language teaching have changed. But another is that the attraction of that shiny new technology has long since faded. Students grew bored with it, and educators and administrators learned that audio recording was not a silver bullet that could replace human language teachers.

I saw another example in college of the premature adoption of technology. In several buildings on campus that had been built in the late 1960s, every classroom had one or two large television monitors hanging from the ceiling. That was before video could be recorded on cassettes, so the televisions were connected by cables to bulky reel-to-reel videotape decks in a central control room. The idea had been that students, instead of listening to a teacher's lecture, would sit in a classroom and watch prerecorded lessons on the monitors.

Like language labs, those television monitors had been seen as a way not only to apply a supposedly superior pedagogical method, that is, audiovisual media; they were also intended to save money on human teachers. The novelty and ease of recording images and sound on magnetic tape rather than film had also been an attraction. But by the time I enrolled at the university, the equipment was just collecting dust. In my three years on campus, I never saw or heard of any classes in which those televisions were used. The excitement about video as an education tool that could replace the teacher had been misplaced, and the appeal of the shiny new technology had been fleeting. It turned out that students learned better from a human teacher in the classroom.

Since the early 2000s, when I returned to academia as a teacher, I have seen more cases of the overeager adoption of new technology. In 2007, I was asked to teach several undergraduate writing classes in which the students would use laptop computers in class. A fancy special classroom, paid for with a large government grant and private donations, was built for the experimental classes, in which students would use expensive computers with touchscreens and styluses. For a couple of years, with constant help from technical assistants, I taught my English academic

writing classes in that classroom using those tablet computers. At the time, few students brought their own laptops to school, and they seemed to enjoy the opportunity to try out the novel touchscreen interfaces. Writing and revising on the computers did seem more convenient than with a pencil and paper. And since the computers were connected to WiFi, students could search the Internet to gather information for their papers and share files of their drafts with each other during class. Although students were often frustrated by the slow startup times of the computers and other technical problems, I thought the classes went reasonably well. And it was fun for me to be able to use the latest technology in my classes.

After my role in that pilot project ended, I moved back to a regular classroom and taught once again with only a blackboard, chalk, and handouts printed on an old-fashioned copy machine. The students did their in-class writing by hand. After that high-tech classroom, it all seemed quaint and inefficient at first. But not a month had gone by before I noticed that the students were more engaged with the class than in the previous classroom. Without the distractions of the flashy gizmos, they were able to concentrate on the language, the arguments, and the content of their writing better. Students learned more without the technology.

In recent decades, many other technological innovations have been adopted with much enthusiasm but ultimately abandoned. These include language lessons taught using computer software, classes managed and conducted through websites, and students practicing language use through e-mail, text chat, and social media. Often the reason for the failure was the technology's immaturity. The tablet computers we used in 2007 had been slow and difficult to operate, with a too-steep learning curve for the students; such classes would go much more smoothly now using smartphones or iPads. The video technology that had gained no traction at my university in California in the mid-1970s was expensive and difficult to use, a far cry from the ease of video recording and sharing today.

But the biggest reason for the failure of most new educational technology was not its cost or complexity. Rather, it was its inability to provide the human interaction that is essential to most learning. Records and videotapes, computer software and websites, had attracted attention and investment largely because

they seemed to reduce or eliminate the need for human teaching. Teachers saw them as ways to cut down on their hours standing in front of classrooms, and educational administrators saw them as ways to save on teachers' salaries. But those new technologies turned out to be poor substitutes for human teachers. Young people just seemed to learn better, and to be more motivated to continue studying, when they had a person in front of them to instruct, advise, encourage, nag, and entertain them.

After observing so many failures of technology adoption over the years, and despite being an early adopter of digital technology myself, I became jaded. I did not think any new technology would significantly reduce the need for human teachers, especially in foreign-language education. When I helped establish a large undergraduate English program in 2008, I made sure that we used technology only as an aid for teachers, never a replacement. Fifteen years later, classes in that program continued to be small, with an average of only fifteen students each, taught in person by a team of nearly thirty full-time faculty. Fortunately, our university administration understood the value of human instruction. Despite the expense of teacher salaries, we were never pressured to try to replace some of those teachers with, say, interactive software or computerized assessment.

* * *

My skepticism about the potential of technology to replace human language teachers began to weaken in late 2016. That was when advances in machine learning technology enabled, for the first time, reasonably good translation by computers. I had long believed that translation could be done only by human beings, because only we were able to grasp the intended meaning of a text and express that meaning appropriately in another language. I had been wrong. Even though computers still did not grasp meaning in the cognitive way that humans do, when trained deeply on large bilingual corpora, electronic neural networks were in fact able to imitate the translation skill reasonably well. Language—and, by extension, language teaching—no longer seemed like a uniquely human ability.

My skepticism about the potential for computer language use vanished in late 2022, when a further leap in artificial intelligence technology suggested that computer software might, finally, be able to replace human language teachers. On November 30, the

U.S. company OpenAI released a public research preview of its interactive language model ChatGPT. Soon the Internet and news media were filled with reports and examples of its amazingly human-like, sometimes superhuman ability to respond to text prompts naturally and, it seemed, intelligently. When asked to write an essay on, say, the influence of the French Revolution on English literature in the Nineteenth Century or the depiction of people with disabilities in Hollywood movies of the 1930s, within seconds it would produce an original, well-organized, multiparagraph essay that would receive a good grade from many high school and university teachers. It could also compose original stories on nearly any topic—not the most creative stories, perhaps, but well written and with a pleasing narrative arc. And it could explain physics and paleontology, discuss surfing, gardening, and yoga, write computer code, compose rap lyrics, and much, much more.

In my own testing of ChatGPT, I tried to see how well it seemed to understand language. I gave it sentences containing a word with multiple meanings—such as "critical" or "oversight"—and asked it to identify the sense in which the word was used in each sentence. It performed flawlessly. I gave it a list of words and asked it to compose illustrative example sentences for each; in a few seconds, it produced natural, well-formed sentences as good as any that I could write. I gave it the first paragraph of a recent short story from *The New Yorker* magazine and asked it to annotate the text for learners of English. It provided a list of the relatively hard words and phrases that appeared in the paragraph—*antecedents*, *plagues*, *dire weather*—with correct glosses in English and Japanese. Many years before, when studying Russian, I had found such annotated texts useful for improving my reading skills. But I could use only texts that someone else had already annotated. With ChatGPT, it seemed, learners could obtain annotations for any text they wanted to read—within seconds.

I also tried conversing with ChatGPT. Although during the research preview it could be accessed only through a text interface, I found it quite enjoyable to chat with. Playing the role of a Japanese college student, I told it to ask me how I had spent my summer vacation. It responded with questions like "Did you go on any trips or vacations during the summer?" and "Did you have any memorable experiences or moments from your summer

vacation?" When I answered "I spent most of my vacation helping my parents in their shop. But on days when the shop was closed, I usually went to the beach with some friends," it responded "Thank you for sharing! It sounds like you had a mix of work and relaxation during your summer vacation. It's always good to find a balance between the two." Occasionally it responded with canned-sounding answers—when I said that I had gone to Hokkaido, it responded with "Hokkaido is a beautiful place to visit! It's the northernmost of Japan's main islands and is known for its natural beauty, including mountains, forests, and lakes."—overall the responses were friendly and appropriate reactions to what I had said. ChatGPT sounded like an experienced English conversation teacher who knows how to carry on an interesting conversation with a student without getting too personal or intimate.

ChatGPT was not flawless, though. As many people noted during the first weeks after its release, it was often terrible at quantitative reasoning. When I asked whether the sentence "December 31st is the penultimate day of both the month and year" was correct, it said it was. When challenged, it insisted that "[t]he word 'penultimate' helps to convey the idea that December 31 is the next-to-last day in both the month and the year, which is accurate." No, it isn't. Compared to some other large language models, ChatGPT did not seem as prone to "hallucinations"—that is, plausible-sounding but false assertions of fact—but it would occasionally give me information as flatly incorrect as its statement that the 31st is the next-to-last day in December.

Its explanations of grammar could not be trusted, either. When told to correct typical learners' mistakes in English, it did so flawlessly. But when asked why it had made those corrections, its explanations were often wrong. It misidentified singular verbs as plurals and the infinitive particle *to* as a preposition. And while it could translate between Japanese and English as well as the best dedicated machine translation systems then available, its explanations in English of Japanese grammar or of English grammar in Japanese were often wrong. It could converse in Japanese as well as English, but its Japanese seemed less polished and natural than its English—presumably because it had been trained on a much larger volume of English text.

Despite those flaws, many of which seemed likely to be fixed eventually, for application to language learning ChatGPT

was clearly already useful. Its deep and wide knowledge of English vocabulary, its ability to grasp the meanings of words, sentences, and entire texts, its readiness to discuss nearly any topic, and its flawless production of English—in weeks of testing, I never saw it make a single grammatical or vocabulary error in English—showed that it was as competent linguistically as any professional English teacher.

In the weeks after the research preview began, OpenAI's future plans for ChatGPT were unclear. But it seemed likely that, after having refined and improved its performance, the company would release a commercial version, presumably with application programming interfaces, or APIs, that would allow other software companies to integrate its functions into their apps. That would enable developers of language-learning software to tailor ChatGPT's responses to the needs of learners. A spoken interface was sure to come as well. Talking digital assistants were available for smart speakers and other applications, and it wouldn't be long before people would be able to have conversations with virtual characters driven by ChatGPT.

I also knew that OpenAI would not have a monopoly on ChatGPT's level of performance. Earlier in 2022, the company had created shockwaves similar to ChatGPT's when it released DALL-E 2, a system for creating images and art from text prompts. Within months, several comparable programs, both commercial and open-source, had appeared as well. While training large language models required advanced expertise and expensive computer resources, the potential payoff was so large that competitors were sure to release software at least as powerful as OpenAI's before long.

* * *

In 1950, the British mathematician and computer scientist Alan Turing proposed a test for evaluating whether a computer program appeared genuinely intelligent. People would interact with both the program and a real human being through a text interface. If they could not tell the difference, then the program could be said to exhibit artificial intelligence. In its initial release, ChatGPT probably could not pass the Turing test yet, nor did it seem designed to do so. Its responses were too quick, too complete, too flawless linguistically for any human to produce. Its errors at quantitative reasoning, including simple arithmetic, would seem

unlikely for a human with such fluent language skills. And, when asked about itself, it would reply that it was an artificial intelligence assistant trained by OpenAI and without personal experiences or feelings. Only an intentionally deceptive human would give such replies.

In other ways, however, ChatGPT was frighteningly human-like. Not only could it produce original English sentences accurately and naturally. It also understood the prompts I fed it exceedingly well and nearly always responded appropriately. Particularly impressive was its understanding of deixis—that is, context-dependent meaning. In several long exchanges, I was able to ask questions like "What about the other one?" in reference to something much earlier in the conversation, and it told me about the correct "one." The grammar of pronoun reference in English is complex. Despite many years of explaining English texts to students, I am not confident that I could come up with a definitive set of rules that would correctly match each "it" and "they" and "first one" and "other one" to the corresponding referent earlier in a text. The grammar, though intuitive to me as a native speaker, seemed to depend on complex interactions among syntax, semantics, and knowledge of the world that I could not fully explain. As with its vast knowledge about many subjects, ChatGPT seemed almost superhuman in its grasp of deixis.

But what struck me most was how conversing with ChatGPT *felt* like talking with a real person. Although I knew it was only a computer, circuits without consciousness or feelings, I found myself beginning conversations with "hello" or "good morning," prefacing requests with "please," and finishing with "thank you." My unconscious politeness was partly due to its friendly tone. But I think I was swayed more by its apparent understanding of nearly everything I input. Even though I knew better, there seemed to be a conscious mind at the other end of our text exchanges. And while I was careful to refer to ChatGPT as "it," other people I talked to called it "he" or "she."

Though it perhaps could not yet pass the Turing test, ChatGPT's ability to interact with users in an almost human way, as well as its outstanding linguistic ability, convinced me that it could be applied much more effectively to language learning than had LP records, reel-to-reel videotapes, tablet computers, or any other previous new technology I had tried. While attractive

when they first appeared, those technologies had all lacked what kept so many human teachers employed: the ability to interact with their students as individuals. ChatGPT showed that language models could also have that ability. Once they were linked to audio and video interfaces, and once their memories, responses, and personalities were tailored to language teaching, they would be effective, enjoyable, inexpensive, and convenient tools for people learning second languages—not human, of course, but close enough.

What would happen to language teaching? One possibility was much less need for human-led instruction. Highly motivated language learners—those already driven to study a language out of interest or need—would be increasingly dissatisfied with classroom instruction. While a flesh-and-blood teacher might continue to be more compelling as a role model, learners would realize that they learned more from the digital avatars, which, with their vast knowledge, memory, and adaptability, would be able to tailor their instruction better to the individual's ability, interests, and needs. Less motivated learners, including those who didn't want to study another language at all, could be motivated by language-teaching bots integrated into online games. In both cases, the ability of the new AI-driven characters to interact with learners in a fun, interesting, and personal way would make it difficult for classroom teachers to compete.

The role for humans in language education would not, I thought, disappear completely. No matter how human-like the software became, it seemed likely that young people would continue to seek role models in other humans, especially those they knew in person: their parents, their siblings, their peers, their teachers. While the AI might be better than a human teacher at helping students learn and use another language, the questions of why they should learn the language, of what role that language might play in their lives, would be answered better through the experience and advice and example of flesh-and-blood people. But the need for language teachers to do the day-to-day tasks of language education—instructing students from textbooks, explaining grammar and vocabulary, demonstrating how a language is used—could very well end. Teachers would be less like instructors and more like guides or coaches.

Although I was sure that large language models would have

a powerful impact on language learning, I was much less confi-
dent about the ability of educational systems to adapt, especially
in Japan. The response among educators and administrators to
the challenges raised by improved machine translation had been
very slow. While some teachers, especially at the university
level, had started to incorporate machine translation into the
teaching of English and other languages, English education at
the elementary and secondary levels—the national curriculum,
the government-approved textbooks, the entrance examinations
for high schools and universities, the training and certification of
teachers—continued as before, with little recognition that people
throughout the world were now using software to communicate
across language barriers. ChatGPT represented an even greater
challenge to long-held assumptions about how and why people
use and learn other languages. I was pessimistic that the Japanese
educational system would be able to adapt any time soon.

When I wrote this essay, just a few weeks after ChatGPT
was released, I was keenly aware that the performance and
application both of it and of other large language models might
evolve in unanticipated ways even in the few months before the
essay was published. I also knew that, a few years into the future,
I might appear to have been ridiculously naive in my predictions
about how this new AI software would affect language learning
and education. But the huge step forward in computer language
use shown by ChatGPT, and the steadily accelerating pace of
advances in other areas of artificial intelligence, convinced me
that a new era in language learning and education had indeed
arrived—for better or for worse.

> Three videos about this topic made by the author in December 2022 can
> be viewed at the Internet Archive: "ChatGPT and Language Educa-
> tion" (https://archive.org/details/chat-gpt-and-language-education),
> "Using ChatGPT for Language Learning" (https://archive.org/details/
> using-chatgpt-for-language-learning), and "ChatGPT and the Future of
> Language Learning" (https://archive.org/details/chatgpt-and-the-
> future-of-language-learning).