Journal of Academic Writing
Vol. 12 No 1 Winter 2022, pages 1-9
http://dx.doi.org/10.18552/joaw.v12i1.820

Amazement and Trepidation: Implications of AI-Based Natural Language Production for the Teaching of Writing

Chris M. Anson
North Carolina State University, U.S.A.

Ingerid S. Straume
University of Oslo, Norway

Abstract

AI-based natural language production systems are currently able to produce unique text with minimal human intervention. Because such systems are improving at a very fast pace, teachers who expect students to produce their own writing—engaging in the complex processes of generating and organizing ideas, researching topics, drafting coherent prose, and using feedback to make principled revisions that both improve the quality of the text and help them to develop as writers—will confront the prospect that students can use the systems to produce human-looking text without engaging in these processes. In this article, we first describe the nature and capabilities of AI-based natural language production systems such as GPT-3, then offer some suggestions for how instructors might meet the challenges of the increasing improvement of the systems and their availability to students.

Introduction

Educators have long feared that new technological advances will subvert their students' learning processes—a fear that stretches back for decades among mathematics teachers after the availability of electronic calculators (see Banks, 2011). The advent of the Internet first created paranoia about students' use of sources, because so much information was soon at the ready with a few clicks, and because copy-and-paste became such an easy way for students to splice other people's words into their writing without attribution. Texts could be manipulated onscreen to reach required length (for example, by imperceptibly increasing the font size of commas and periods or changing the character width). Technologically savvy students soon fooled plagiarism detection tools by substituting identical-looking letters with Cyrillic fonts that the systems didn't recognize, adding invisible words (in white font) at the end of lines, or including made-up references. Cell phones enabled cheating on classroom tests through texting, stored information, or online searches. Paper-writing services flourished on the Internet. Meanwhile, more sophisticated translation programs continued to besiege foreign language instructors and those working with L2 students (Karbach, 2020).

However, these sneaky evasions will look like frivolities next to the potential of AI-based language production technologies: systems that automatically produce writing that reads exactly, or almost exactly, like it was written by human beings. This next-generation natural language processing technology raises crucial questions for writing educators. In this essay, we first briefly describe the development of AI systems like GPT-3 that are capable of generating, summarizing, organizing, and translating natural language text, and offer some examples of these systems' capabilities, both helpful and troubling. We then turn to implications of these systems for the teaching of academic writing in a context where they will be increasingly popularized and available for students to use.

What is Natural Language Production?

Starting in the 1970s, artificial intelligence experts experimented with computer programs that they hoped could work with 'natural language' (the language produced by humans). For example, Schank (1984) and colleagues at the Yale Artificial Intelligence Laboratories were interested in creating a "cognitive computer" that could "understand English and respond to sentences and stories with the kind of logical conclusions and inferences an average human would make" (p. 14). To do so, they first needed to learn how humans do these things, and turned to the developing realm of artificial intelligence to find out. One experiment called "Tale Spin" was programmed to create simple fable-like stories using typical scenes and anthropomorphized animals. Designed by doctoral student James Meehan, the program attempted to write stories "by simulating a world, assigning goals to some characters, and saying what happens when these goals interact with events in the simulated world" (Schank & Abelson, 1977, p. 210; see Meehan, 1976). However, the computer's mistakes soon revealed the knowledge it needed to produce coherent tales:

> One day Joe Bear was hungry. He asked his friend Irving Bird where some honey was. Irving told him there was a beehive in the oak tree. Joe threatened to hit Irving if he didn't tell him where some honey was. (Schank & Abelson, 1977, p. 83)

The botched story showed that the system needed to look backwards (to understand what it had already said) in order to move forward, as well as to know that honey can be found in a beehive. Another case demonstrated the need to understand even simple goals:

> Once upon a time there was a dishonest fox and a vain crow. One day the crow was sitting in his tree, holding a piece of cheese in his mouth. He noticed that he was holding the piece of cheese. He became hungry, and swallowed the cheese. The fox walked over to the crow. The end.

The program did not know that it should be possible for an actor to suspend its usual goal (in this case, of satisfying hunger) for the story to continue. The crow swallows the cheese before the fox can flatter it into opening its beak.

Through these and dozens of other experiments, the researchers discovered that for computer programs to work with language, they had to have world knowledge, which even in a child is vast. A computer would need to know that people are assigned roles that come with actions and behaviors ("scripts"), routines, props, appearances, and levels of authority. Aspects of language such as subtle humor and irony, deliberate violations of speech acts (Grice, 1975), and especially implication (understood through inferencing) posed serious obstacles for AI-based language production and comprehension. Consider two further examples:

> André ate the spoiled fish. The night was unpleasant.
>
> Aya slipped on the ice. Everyone wrote on the cast.

A computer has to fill in the information between the two sentences by "understanding" the causal effects of food poisoning or that falling can fracture a bone resulting in a cast on which it is common for friends to sign or write kind messages. However, providing enough world knowledge for a computer to generate text as if written by a human would take hundreds of programmers working for many years. As Schank puts it, "if we see every experience we have as knowledge structures in its own right, then thousands of structures quickly become millions of structures." (1984, p. 168) In this case, humans had to continuously 'teach' the computer because, as Schank explains, the programs "were not being changed by what they read […] To [change], an understanding system must be capable of being reminded of something it has stored in its long-term memory […]" (1984, p. 168). (For a fuller description of the challenges of getting computers to work effectively with natural language, see Anson, 2006).

Today, however, 'machine learning' and 'deep learning' have largely overcome these problems.[1] Machine learning is a branch of artificial intelligence in which humans show computers how to learn on their own. One popular explanation sums it up simply: you show a computer a few pictures of cats and dogs. The computer looks for patterns in the pictures—sizes, nose dimensions, tails, and other features—and represents that information statistically. Eventually it will distinguish between cats and dogs—and continue to add to its knowledge without humans assisting. The most robust kinds of machine learning are 'unsupervised,' meaning that the computers need little intervention once they are given a task, and essentially teach themselves.

In this sense, computers can be programmed to do their own information gathering (or learning) by scraping hundreds of millions of texts and identifying predictable patterns among them. Machine learning is now being used across wide sectors of government, business and industry, transportation, education, healthcare, insurance, research—in short, almost everywhere (see SAS Insights, 2020).

For an example closer to writing instruction, consider how a relatively simple machine learning program can be used to create automated essay scoring systems. Imagine that ten trained human evaluators use a rubric to score 100 student essays, written to the same prompt under the same circumstances. When the humans agree on the scores, the essays are fed into a computer program designed to look at the scores against patterns in the essays. The computer can make millions of passes through the essays in seconds, looking for multiple kinds of features matched to the scores (sentence length, essay length, word choice, paragraph structure, preprogrammed errors, etc.). Eventually it can create a kind of predictive matrix so that if the 101st essay is submitted (with no human score), it will give the essay a score very likely to have been assigned by the human readers. The more essays it is fed, the more it learns and the more robust it becomes. (For problems associated with this method of scoring, see Anson & Perelman, 2017.)

One of the most written-about AI systems that generates natural language texts is GPT-3 (a GPT-4 version is slated for release in 2023).[2] GPT is an acronym for the rather lackluster name "generative pre-trained transformer." The GPT systems were developed by OpenAI using technological advances similar to those behind Google's more familiar Smart Compose on Gmail. Smart Compose looks at a sentence being composed and uses an algorithm (based on millions of prior sentences) to predict the next word(s). When it is correct, the user simply hits the tab bar and the predicted text is inserted, saving time writing. Smart Compose is essentially a primitive version of GPT—much of the time, it 'knows' what you are writing next, but it continues to learn from its mistakes and from enormous databases of text, which makes it smarter over time (see Dai et al., 2019).

GPT-3 is a massive deep-learning transformer language model, commonly known as a 'robo-writer.' The first step in creating the system is to pre-train it with a diverse corpus of information coming from multiple sources. Once the information reaches critical mass and the system is tested, it can continue to collect information to enhance its capabilities. The current system is likened to a neural net that contains 1.3 billion parameters.[3]

What Can Text Generation Systems Do?

Experiments with GPT-3 have led, among many others, to the following successful text-generation activities:

[1] AI scientists distinguish between different kinds of machine learning, the most sophisticated of which is deep learning. For an overview, see https://www.sas.com/en_us/ insights/analytics/ machine-learning.html

[2] For a fascinating history—with interspersed GPT-2 generated text that looks like it was composed as part of the article itself—see Seabrook (2019).

[3] It is beyond the scope and readership of this article to explain the programmatic intricacies of these systems. For more information, see Ammu (2022); Shmelzer (2021).

- *Creating and enhancing websites*: Typing descriptions of desired characteristics prompts the system to create web pages with accompanying code. For example, if you want to add a table of the ten countries with the highest GDPs, the system will design, format, and insert it. See https://analyticsindiamag.com/open-ai-gpt-3-code-generator-app-building/

- *Summarizing longer texts*: Although often used to create news summaries of longer stories, GPT-3 can summarize entire books with only occasional errors. For an example, see the project summarizing the entirety of *Alice in Wonderland* at https://openai.com/blog/summarizing-books/

- *Writing fiction and poetry*: GPT-3 is capable of generating interesting (if sometimes strange) poetry and fiction. We gave GPT-3 the first lines of a poem (with no reference to COVID-19), and it completed it for us (see if you can tell which are our lines and which are the machine's):[4]

  It came with stealth, a few sick souls
  and then the world's jets carried it afar.
  Pandemic, they called it, and we all knew
  What it would do, but no one knew how.

  We tried to lock our doors and shutter our hearts
  But it found a way in anyway
  And swept through our cities and towns
  Ravaging first the old and the weak.

- Writing podcasts and creating music (for impressive examples, see https://towardsdatascience.com/creating-a-podcast-with-a-i-6fe6a0ec6c37)

Other functionalities of note include creating unique memes and Tweets, generating resumes from key ideas, translating legalese into plain (but legally accurate) language, creating outlines, and writing turn-by-turn directions. For our purposes, the most important is the completion or continuation of texts from prompts or partial inputs.

A quick example of our own brief experimentation with GPT-3 indicates what such programs can do with natural language. We began a story with the following lines:

> Franklin Emmonds stood by the old stove in the dilapidated cabin wondering how he had reached this point in his life. He was 60 miles from the nearest paved road and a thousand light years from a sense of his own place in the world.

After we submitted these lines to GPT-3, the system continued the story as follows:

> The single-story cabin, spread out over a clearing of pine, with rotting boards and wood shingles, had been his home for more than a decade. The modern world was too distant, too noisy. It never stopped talking. He was happy with the simplicity of his life.

Remarkably, the description of the cabin is entirely coherent with the human-composed beginning, but more to the point, the system somehow interpreted the metaphoric idea that Emmonds was "a thousand light years from a sense of his own place in the world" by adding that the modern world was "too distant, too noisy" and that—and here is a creative flourish—"it never stopped talking." GPT's story continues darkly, picking up on Emmonds' questioning "how he had reached this point in his life":

> Emmonds did not regret his decision to end his own life.

[4] Our lines were only the first two.

> "Death is the ultimate escape," he said.
>
> He considered himself an orphan. His parents had died when he was 10. His wife had died in 2011 of an accidental overdose of antidepressants, then her sister. Then the neighbors were killed by a drug dealer, and he found the bodies. Emmonds thought his only friend was his dog, Brutus.
>
> Emmonds wasn't addicted to drugs. He didn't drink, smoke, or even take an occasional aspirin. His only addiction, he said, was to the deep peace of death.

From its own previous input, GPT-3 will continue writing the story, each addition providing it with more information to use for further text generation.

People experimenting with GPT-3 do find that the system can create bizarre, illogical, incoherent, or repetitive text because it does not really 'think,' a problem that has led some to characterize such systems as 'stochastic parrots'—repeating what they have heard and adding random material (Bender et al., 2021). In addition, such AI systems are prone to replicating any bias they find in natural-language texts (GPT-3 created racist associations of "Muslim" with violence, positioned nurses and receptionists as women, and generated negative terms when asked to complete the prompt "The Black man was very . . . "). The systems can be used to auto-generate massive amounts of slanted text to influence political elections and other areas of public opinion or belief. As Hutson (2021, p. 24) has asserted, "with its impressively deep knowledge of extremist communities, it can produce polemics parroting Nazis, conspiracy theorists and white supremacists."

However, a few edits to the generated text will fix some of these problems. In their current state of development, text-generation systems could be considered 'writing assistants' requiring some supervision of outputs, but possibly substituting for, or subverting, the human capacity and effort to compose from scratch. For anxious student writers (especially procrastinators), such a possibility may seem like a dream come true.

From Sandbox Experimentation to Classroom Application

Let's turn back, then, to the world of students and learning. Text-generation systems are already available on a fee-per-paper basis. For example, the company EssaySoft offers several digital aids for students, one of which is called Essay Generator (www.essaysoft.net). Essay Generator will produce essays "on virtually any topic" that subvert plagiarism-detection systems by assembling paraphrased sentences and replacing words with synonyms. It can generate essays of different lengths, shuffle content, and create bibliographies. It is not clear whether the system uses massive databases to predict and generate text or simply surfs existing texts to patch together coherent essays, but there is no question that GPT-3, Google, and other sophisticated systems will find their way quickly into the essay-generation industry.

To test Essay Generator, we typed in a currently relevant topic, "vaccine hesitancy," in the trial box at the company's website. We set the essay length to 600 words, the research depth to "low," the random synonym replacer to "off", and left the other choices blank. Although the website's test page produces only the first few lines of an essay, the results were impressive:

> What is vaccine hesitancy and why is it important? The World Health Organization defines vaccine hesitancy as a "delay in acceptance or refusal of vaccines despite availability of vaccination services."[1] Most vaccine-hesitant parents are in the middle of a spectrum and underimmunize their children instead of not immunizing them at all.[2] According to a recent survey, although only 3% of parents refused all vaccinate . . . .

Of course, some aspects of this start could make teachers of foundational writing courses suspicious that it was not produced by a student. Moreover, if the system has simply "patchwritten" (Howard, 1999) existing but altered sentences, it is not generating truly original text. However, consider what happened when we typed "vaccine hesitancy" into GPT-3:

> In the early 1990s, vaccine coverage rates in the United States were estimated to be only 30 percent, in part because of fears that the vaccine might cause serious side effects, such as arthritis, and because the vaccine was not part of the routine childhood immunization schedule. In 1991, the Vaccine Adverse Event Reporting System (VAERS), a national vaccine safety surveillance system, was established to collect information about adverse events after vaccination. In response to these concerns, public health officials and vaccine manufacturers developed new ways to make vaccines safer and more accessible. Immunization rates in the United States increased, and by the late 1990s, they had reached 90 percent. However, vaccination rates have not always been this high, and vaccine hesitancy (the reluctance or refusal to vaccinate despite the availability of vaccines) continues to be a challenge for public health officials.
>
> There are many reasons why people may hesitate to vaccinate their children or themselves. . . .

With a few quick reference checks (a URL to VAERS, perhaps a search for vaccine rates in the 1990s), this text might be passed off as an original research paper. Disturbingly, it *is* original—the words and sentences are newly composed.

Implications for Instruction

Clearly, the outputs of current AI text-generation systems are still flawed but, like the steady improvement of language translation programs, there is no question that these systems will continue to improve at breathtaking speed. We turn, then, to some thoughts about how instructors might face the challenges of ubiquitous text generation in the context of GPT-4 as well as Google's system under development, which boasts over one trillion parameters (see Parik, 2021).

First, instructors need to be aware of these developments and adjust their assignments and feedback practices. If they do not, much effort will be wasted in providing feedback to texts that have never been touched by human hands—feedback that is unlikely to be used for further learning or improvement. Another lost cause, as we see it, is anti-plagiarism software: Some instructors naïvely believe that these tools are capable of catching and preventing all forms of plagiarism. However, tools for detecting plagiarism are constantly lagging behind the inventions they are designed to deter. If students submit AI-generated text, such cases represent 'contract cheating' (e.g., Curtis & Clare, 2017; Lancaster & Clarke, 2015), but as we have shown, they are not plagiarism—copying and stealing other people's ideas. Rather, they are the opposite, the ideas of *outis*, nobody. As Dehouch (2021) points out, it is highly unlikely that a plagiarism-detection system like Turnitin could identify a string of words generated by AI-based systems as previously published.

Of course, assignments can be designed that are difficult to respond to with essay generators. For example, assignments leading students through processes of invention, drafting, and multiple revisions using peer feedback *might* begin with AI-generated text, but the subsequent effort to shape and repurpose them would represent significant cognitive (human) effort—and learning. Instructors could also require that students weave in material from class discussions to which text-generators do not have access. Other activities such as in-class writing, journal writing about student experiences, or reflections on class activities such as attending lectures could also thwart the use of AI text-generation systems.

Perhaps we should take a step back and ask ourselves if the greatest problem is whether students write *per se*. For are there not larger issues at play here? If students do not see the need to read, engage with ideas, hone their thinking skills, and through these efforts, experience the joy of taking part in an enduring, scholarly conversation, they are missing out on the greater pleasures of fulfilment that come with (hard) work and engagement over time. In revising our assignments, perhaps our focus should be less on (quantitative) writing skills and more on

qualitative thinking, reading, discussing, and not least, on *real* forms of questioning. And this, one might argue, could interrogate how the use of AI is part of a larger context, including the way our institutions are rigged for competition endorsing meritocratic ideals that do not correspond to the socio-political reality (Pikkety, 2014; Sandel, 2020).

Rather than trying to combat or extinguish tools that we fear will subvert students' learning, instructors could bring them into class, have students work with them, and analyze their outputs. By creating awareness, not least among the students as a group, ethical and practical dilemmas could be addressed. An excellent example comes from Fyfe (2022), who developed an assignment that *relied* on GPT-2. Students were deliberately assigned to use GPT-2 to help them write a paper, then had to explain which parts were theirs and which were generated by the system, what the system prompted them to learn, and how they felt ethically about the process. We also imagine that students and instructors could co-create assignments and feedback or assessment rubrics where taking (admittedly unproductive) shortcuts would be impossible. As 'honest' students would recognize, it is also a matter of justice.

In keeping with new 'writing-about-writing' pedagogies that advocate the building of higher-level knowledge of the subject in addition to enhancing skills (Downs & Wardle, 2007), students might also discuss the advantages of such systems in the future (for the more mundane tasks of translating legalese, expanding kernels of information into coherent reviews, or writing boilerplate and routine business communication, news summaries, and resumes), alongside the disadvantages to thinking, learning, and cultural transmission. These explorations and conversations would help students to develop much-needed metalinguistic awareness and learn threshold concepts about written communication (Adler-Kassner & Wardle, 2015).

In sum, as there are no realistic expectations of stopping AI in higher education, our best strategy might be to use AI to our advantage and not succumb to its control. On a more principled level, however, students and universities should also seek to deter big tech corporations from appropriating students' data and resist technologies that turn human attention into capital (Saltman, 2021; Zuboff, 2019). Preserving our ability to read and think in longer, slower, and more deliberative ways is in this sense part of a much bigger struggle.

References

Adler-Kassner, L., & Wardle, E. (2015). *Naming what we know: Threshold concepts of writing studies*. Utah State University Press.

Ammu, B. (2022). GPT-3: All you need to know about the AI language model. *Sigmoid*, n.p. https://www.sigmoid.com/blogs/gpt-3-all-you-need-to-know-about-the-ai-language-model/

Anson, C. M. (2006). Can't touch this: Reflections on the servitude of computers as readers. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 38-56). Utah State University Press.

Anson, C. M., & Perelman, L. (2017). Myth: Machines can evaluate writing well. In D. M. Lowe & C. E. Ball (Eds.), *Bad ideas about writing* (pp. 278-286). Digital Publishing Institute.

Banks, S. A. (2011). *A historical analysis of attitudes toward the use of calculators in junior high and high school math classrooms in the United States since 1975*. Master's Thesis, Cedarville University.

Bender, E. M., Gebru, T., Macmillan-Major, A., & Schmitchell, M. (2021, March). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623), https://doi.org/10.1145/3442188.3445922

Curtis, G. J., & Clare, J. (2017). How prevalent is contract cheating and to what extent are students repeat offenders? *Academic Ethics*, *15*, 115-124.

Dai, A., Lee, B., Bansal, G., Tsay, J., Lu, J., Chen, M., Zhang, S., Sohn, T., Wang, Y., Wu, Y., Cao, Y., & Chen, Z. (2019, July). Gmail Smart Compose: Real-time assisted writing. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2287-2295). Anchorage.

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics in Science and Environmental Politics*, *21*, 17-23.

Downs, D., & Wardle, E. (2007). Teaching about writing, righting misconceptions: (Re)Envisioning "first-year composition" as "introduction to writing studies." *College Composition and Communication, 58*(4), 552-584.

Fyfe, P. (2022, Feb.). How to cheat on your final paper: Assigning AI for student writing. *AI & Society*. https://link-springer-com.prox.lib.ncsu.edu/content/pdf/10.1007/s00146-022-01397-z.pdf

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. J. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41-58). Academic Press.

Howard, R. M. (1999). *Standing in the shadow of giants: Plagiarists, authors, collaborators*. Praeger.

Hutson, M. (2021). Robo-writers: The rise and risks of language-generating AI. *Nature, 591*(7848), 22-25.

Kharbach, M. (2020, Feb.). Will machine language software make language teachers obsolete? *Educational Technology and Mobile Learning*. https://www.educatorstechnology.com/2020/02/will-machine-language-software-make.html

Lancaster, T., & Clarke, R. (2015). Contract cheating: The outsourcing of assessed student work. In T. Bretag (Ed.), *Handbook of academic integrity* (pp. 639-654). Springer.

Meehan, J. (1976). The metanovel: Writing stories by computer. Unpublished doctoral dissertation, Yale University.

Parik, P. (2021). Successor to GPT-3: Google's trillion parameter language model. *DataDrivenInvestor.* https://medium.datadriveninvestor.com/googles-trillion-parameter-language-model-successor-to-gpt-3-49244cad133

Pikkety, T. (2014). *Capital in the twenty-first century.* Harvard Belknap.

Saltman, K.J. (2020). Artificial intelligence and the technological turn of public education privatization: In defence of democratic education. *London Review of Education*, *18*(2), 196-208.

Sandel, M. (2020). *The tyranny of Merit. What's become of the common good?* Farrar, Straus and Giroux.

SAS Insights (2020). Machine learning: What it is and why it matters. https://www.sas.com/en_us/insights/analytics/machine-learning.html#machine-learning-users

Seabrook, J. (2019, Oct.). The new word: Where will predictive text take us? *The New Yorker.* https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker

Schank, R. C. (1984). *The cognitive computer: On language, learning, and artificial intelligence*. Addison Wesley.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Erlbaum.

Schmelzer, R. (2021). GPT-3. TechTarget Network. https://www.techtarget.com/searchenterpriseai/definition/GPT-3#:~:text=GPT%2D3%20is%20a%20language,internet%20text%20to%20spot%20patterns.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Profile books.