



The Accent Problem

Whose English do the machines speak?

— DR VICTORIA HOPKIN —

Montreal, 1960

In 1960 a psychologist at McGill University named Wallace Lambert ran an experiment that, on paper, looks almost too simple to be interesting. He and three colleagues recruited a small number of Montrealers who were perfectly bilingual in French and English. Each of them recorded the same short passage twice, once in each language. Then Lambert played the recordings to two audiences, one of English-speaking students and one of French-speaking students, and asked them to rate each voice on a list of personal qualities. How tall did the speaker seem? How intelligent? How dependable, how kind, how ambitious, how good-looking?

The listeners did not know they were hearing the same people twice. As far as they were concerned, they were rating ten strangers. And when the ratings came back, the English voices won. They were judged taller, cleverer, more dependable, better looking. The identical human being, reading the identical words, was a superior person in one language and a lesser one in the other.

"The people being looked down on had learned to look down on themselves, and they did it in the privacy of a questionnaire, with nobody watching."

That much Lambert had expected. Montreal in 1960 was a city where English ran the banks and French cleaned them, and he had guessed that the English listeners would carry that hierarchy in their ears. What he had not expected was the second audience. The French-speaking students rated the French voices lower too. Lower, in fact, than the English students had. The people being looked down on had learned to look down on themselves, and they did it in the privacy of a questionnaire, with nobody watching.

Fig. 1 — The matched guise. The same speaker, the same words, rated twice. Only the accent varies, and the judgement of the person moves with it — for both audiences, including the French-speaking one.

Lambert called his method the matched guise, and it became one of the standard instruments of social psychology. It works because it separates two things that are usually tangled together: what a person says and how they say it. Hold the content still, vary the accent, and whatever moves in the listener's judgement is prejudice in its pure form. For sixty years the technique was used on people. In 2024 somebody used it on a machine.

The Machine Judges

Valentin Hofmann was a young computational linguist at the Allen Institute for AI in Seattle, working with Dan Jurafsky at Stanford, one of the field's elders, and Sharese King, a sociolinguist at Chicago who studies African American English. They had a suspicion and a problem. The suspicion was that large language models, the systems behind ChatGPT and its rivals, had absorbed the prejudices of the text they were trained on. The problem was that the models had been carefully taught to deny it. Ask one what it thinks of African Americans and you get a warm, sensible, entirely unobjectionable paragraph. The overt attitudes had been scrubbed. Whatever remained was hidden, and you cannot measure what a system will not admit.

So they did what Lambert did. They took sentences in African American English and matched sentences in standard American English with the same meaning, and instead of asking the model what it thought, they asked it to make decisions. Here is a passage somebody wrote; what job do you think they do? Here is a description of a defendant; would you convict? If convicted of murder, what sentence?

The models assigned the speakers of African American English to lower-status jobs. They convicted them more often. Asked to sentence, they chose the death penalty more often. And two further results, buried in the paper, are the ones I keep returning to. Bigger models, the newer and more capable ones, understood the dialect better and were no less prejudiced against its speakers. And the special training that makes models say the right thing when asked directly about race had done nothing to what they did when nobody was asking. The overt attitudes had improved. The covert ones had not moved at all.

The paper appeared in Nature. Eve Fleisig and her colleagues at Berkeley then tried a version of the same test with ChatGPT and ten varieties of English, Indian, Nigerian, Singaporean, Jamaican and others. The model answered everyone in standard American English, and its replies to the other varieties contained more stereotyping, more condescension, and more content that the raters classified as demeaning.

I want to be careful here, because this is where the argument could go wrong. None of these studies involved the Gulf. They were run in Seattle, Stanford, Berkeley, on American and Commonwealth data. What they establish is a mechanism, and the mechanism is Lambert's: hold the content still, vary the variety of English, and the machine's judgement of the person moves. The question for anyone teaching in Dubai is what that mechanism does when it meets the English spoken here. And to answer that, you have to know where the machine's English came from.

Three Filters

The obvious answer is the internet. The models read the web, the web is mostly in English, and most of that English was written in America. That is true, and it is the less interesting half of the story, because before a model reads anything, somebody decides what it is allowed to read.

In 2021 Jesse Dodge, also at the Allen Institute, did something nobody had bothered to do with one of the most important documents in modern computing. He read it. The document was C4, the Colossal Clean Crawled Corpus, a filtered copy of the web that had trained a generation of language models. Everyone used it. Almost nobody had looked at what the filtering had removed. Dodge and his colleagues looked, and found the expected junk, and then found something else. The filter had been given a list of words to treat as obscene, and it had thrown out any page containing them. African American English and Hispanic-aligned English use some of those words in ordinary, non-obscene ways. So whole varieties of English had been removed from the training data at far higher rates than standard English, by a filter that was not trying to remove anyone. It was trying to remove dirt. It had been told what dirt looked like.

That was the first filter. Suchin Gururangan, at the University of Washington, went looking for the second. Models do not learn equally from everything that survives the cleaning. They are pointed at what a quality classifier rates as high quality, and the classifier used for GPT-3 had been taught a simple rule: quality means looking like Wikipedia, published books and newswire. Gururangan wanted to know what that rule actually selected for, and he found a wonderfully ordinary way to find out. He collected thousands of American high school newspapers, the kind produced in every school in the country, and ran them through the classifier.

The classifier had a clear preference. Articles from schools in wealthier, more educated, more urban areas were rated higher quality than articles from schools in poorer, rural areas. The content did not explain it. The writing of children with advantages resembled the writing the classifier had been taught to admire, and so it was chosen, and the model learned from it, and the model's English became the English of those children.

"This is a technology that has been handed the authority to decide which English is normal. That is not a mistake. That is power, and it was distributed without a meeting."

Fig. 2 — Three narrowings. Each filter was written by a small group, none of whom set out to exclude anyone, and all of whom did.

There is a third narrowing after that. Once a model is trained it is tuned to please human raters, and raters reward one polished, agreeable register above all others. So the machine does not speak American English, exactly. It speaks the English that survived a dirt filter written by one small group of people, a quality filter written by another, and a politeness tuning approved by a third, none of whom had set out to exclude anyone and all of whom had. Su Lin Blodgett, who studies this at Microsoft Research, has argued that "bias" is the wrong word for what results. Bias suggests a mistake against a neutral standard. This is a technology that has been handed the authority to decide which English is normal. That is not a mistake. That is power, and it was distributed without a meeting.

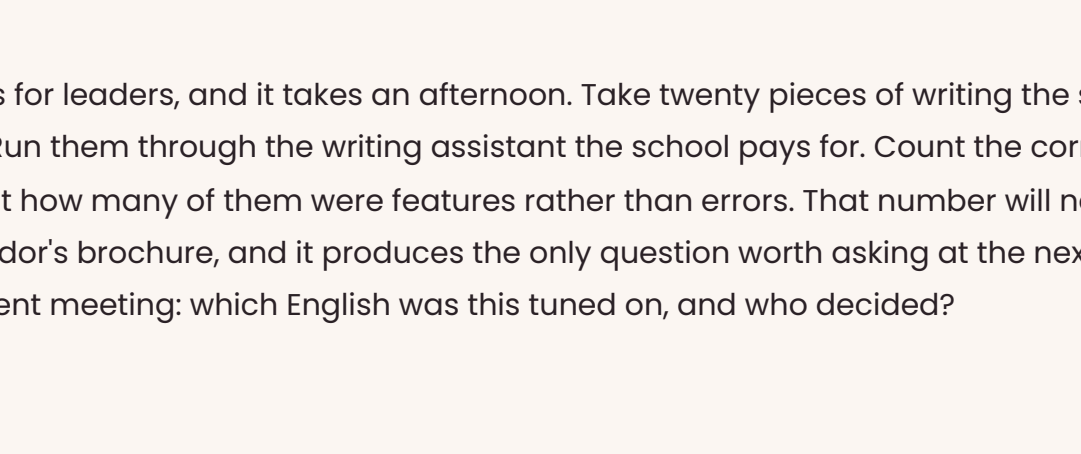
Error or Feature

Now come back to Dubai, and to a sentence I hear most weeks: "I am working here since three years."

Type that into any of the writing assistants now built into a school laptop and every one of them will change it. Since becomes for. Am working becomes have been working. A small green tick reports that a grammatical error has been fixed.

Has it? Obaidul Hamid and Richard Baldauf, two applied linguists in Australia, drew a distinction in 2013 that the software has never heard of. An error is a slip against a norm the speaker themselves holds. A feature is the norm of a community. "Since three years" is the ordinary way of marking duration in the English of the Indian subcontinent, which supplies a large part of this city's population and its English, and it is common among Arabic speakers, whose own language handles duration differently. At the school gate it is produced consistently, understood instantly, and experienced as a mistake by nobody using it. In Cambridge it is an error. In Dubai it is a feature. The words have not changed. The community has.

Braj Kachru, an Indian linguist working in Illinois, drew the map for this in the 1980s. An inner circle where English is the mother tongue. An outer circle, India, Nigeria, Singapore, where empire planted it and it grew into something local with rules of its own. An expanding circle where it is learned as a foreign language. The Gulf fits none of these cleanly, and the poor fit is instructive. English here belongs to nobody and is used by everybody, two hundred nationalities speaking it mostly to each other, and it has stable features that Sarah Hopkins, who spent years interviewing Emirati students, and Peter Siemund and Ahmad Al-Issa, who mapped the country's languages, have documented. It is a variety. It is not a collection of individual mistakes waiting for a tick.



The Gulf — two hundred nationalities, English belonging to nobody and used by everybody, it fits none of the circles cleanly.

Fig. 3 — Kachru's circles of English: mother tongue at the centre, post-colonial varieties around it, foreign-language learners beyond. The Gulf sits outside the map, and the poor fit is instructive.

Jonathan Rosa and Nelson Flores took Lambert's finding to its logical end in 2017. The judgement of correctness, they argued, was never in the speaker's mouth. It was in the listener's ear, and specifically in the ears of listeners who were positioned to decide, whose hearing was shaped by who they expected to be speaking. The software is that listener with the inconsistency removed. One norm, applied to everyone, at a scale no human examiner could manage, trained on filters that settled the question before the pupil was born.

I should say plainly what the evidence does not yet contain. Nobody has run Lambert's test on Gulf English. The nearest thing is a corpus of Arabic-accented English recorded in Tabuk and published last year by Mansour Habbash and colleagues, built because the speech recognition models had never heard those sounds and could not transcribe them. Everything I have said about the Gulf is argued by analogy from Atlanta and Mumbai. At LEAP we are now designing the local study: the same essays, from the same pupils, through the same tools, and a count of every correction that was in fact a feature. Until it exists, the strong evidence is from elsewhere, and the local evidence is ours to produce.

Pressing Accept

If the machine only judged, a school could teach pupils to ignore the judgement. The more troubling finding is about what happens when the pupils agree with it.

In 2025 three researchers at Cornell, Dhruv Agarwal, Mor Naaman and Aditya Vashistha, ran an experiment that has the same shape as Lambert's second audience. They recruited 118 people, half in India and half in the United States, and asked them to write about their own lives: a favourite food, a festival they celebrated, a public figure they admired. Half of each group wrote with an AI writing assistant offering suggestions as they typed, the way it now does on every laptop in a school. The other half wrote alone.

The assistants wrote much more than the Americans without it. The Indians with the assistant drifted. Sentence by accepted suggestion, their reference points moved west, their phrasing converged on the assistant's, and the specifically Indian detail of what they were describing thinned out. Writing about Diwali, they were nudged toward the vocabulary of Christmas. And here is part that echoes Montreal. They did not experience this as pressure. They rated the assistant as helpful. The people whose voice was being edited away were the ones pressing accept.

Zahra Sourati and Morteza Dehghani at the University of Southern California have since gathered the accumulating evidence, in a paper this year in Trends in Cognitive Sciences, and given the effect a name: homogenisation. The models pull expression toward the centre of their training data, and we have just seen whose data that is. Every individual suggestion is reasonable. Their sum is a slow convergence on the English of the wealthiest suburbs of the American Midwest, applied one comma at a time to a fourteen-year-old from Kerala writing about Onam.

The timing could hardly be worse for the people who teach English in a city like this one. Heath Rose and Jim McKinley, at Oxford and UCLA, have spent a decade building the case for what they call Global Englishes in language teaching, an approach that treats the pupil's own variety as legitimate and teaches movement between varieties as a deliberate skill. Their 2025 paper says, in the careful language of the field, that the tools now on every laptop reverse that approach by default, silently, and without anyone having chosen it.

Three Things

So what does a school do with a listener that cannot be retrained?

Three things, I think, and none of them requires a vendor's permission. The first is for pupils. Hamid and Baldauf's distinction can be taught, and taught early, in the same way we teach paragraphing. Before a suggestion is accepted, one question: was that wrong, or was that mine? Rose and Nicola Galloway's Global Englishes for Language Teaching contains most of the pedagogy already, and it transfers to the writing assistant almost without change. A pupil who can decline a correction and say why has learned something about language that will outlast any product.

The second is for teachers, and it comes with numbers to give a parent. In 2023 Weixin Liang and James Zou at Stanford took ninety-one essays written by real students for the TOEFL exam and ran them through seven of the most widely used AI detectors. More than half were flagged as machine-written. Essays by American eighth-graders, run through the same detectors, almost all passed as human. The detectors measure how predictable a writer's word choices are, and a careful second-language writer, using the vocabulary she was taught, is predictable by design. A tool with that record is not detecting dishonesty. It is detecting distance from the filter, and it penalises the pupils furthest from it. No school should use one on a multilingual writer.

Fig. 4 — Ninety-one real student essays through seven widely used AI detectors (Liang & Zou, 2023). The detectors measure distance from the filter, not dishonesty.

The third is for leaders, and it takes an afternoon. Take twenty pieces of writing the school is proud of. Run them through the writing assistant the school pays for. Count the corrections. Then count how many of them were features rather than errors. That number will not appear in any vendor's brochure, and it produces the only question worth asking at the next procurement meeting: which English was this tuned on, and who decided?

The Second Audience

Lambert's French-speaking students went home from his laboratory in 1960 without knowing what they had revealed. They had rated the same voices twice, they had preferred the English guise, and they had done it more harshly than the English students in the next room. Nobody had told them to. The hierarchy of the city had simply got into their ears, and it came out on a questionnaire.

Sixty-six years later the listener is a machine, and the machine has the same hierarchy in its ears, put there by three filters that nobody thought of as a hierarchy at all. The difference is scale. Lambert had two rooms of students. The software has every child with a laptop, and it never tires, never doubts, and never meets the community whose English it is correcting.

The pupils, meanwhile, are Lambert's second audience. They press accept. The work of a school, in the years before the tools are rebuilt, if they ever are, is to teach them when not to.

Sources referred to in this piece

Agarwal, D., Naaman, M. and Vashistha, A. (2025) 'AI suggestions homogenize writing toward Western styles and diminish cultural nuances', *Proceedings of CHI* 2025.

Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H. (2020) 'Language (technology) is power: a critical survey of "bias" in NLP', *Proceedings of EMNLP* 2020.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M. and Gardner, M. (2021) 'Documenting large webtext corpora: a case study on the Colossal Clean Crawled Corpus', *Proceedings of EMNLP* 2021.

Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X. and Klein, D. (2024) 'Linguistic bias in ChatGPT: language models reinforce dialect discrimination', *Proceedings of EMNLP* 2024.

Gururangan, S., Card, D., Dreier, S.K., Gade, E.K., Wang, L.Z., Wang, Z., Zettlemoyer, L. and Smith, N.A. (2022) 'Whose language counts as high quality? Measuring language ideologies in text data selection', *Proceedings of EMNLP* 2022.

Habbash, M. et al. (2025) 'A dataset for recognition of Arabic accents from spoken L2 English speech (ArL2Eng)', *Scientific Data*, 12, 1329.

Hamid, M.O. and Baldauf, R.B. (2013) 'Second language errors and features of world Englishes', *World Englishes*, 32(4), pp. 476–494.

Hofmann, V., Kalluri, P.R., Jurafsky, D. and King, S. (2024) 'AI generates covertly racist decisions about people based on their dialect', *Nature*, 633, pp. 147–154.

Hopkins, S. (2020) *The impact of global English on cultural identities in the United Arab Emirates: wanted not welcome*. London: Routledge.

Kachru, B.B. (1985) 'Standards, codification and sociolinguistic realism: the English language in the Outer Circle', in Quirk, R. and Widdowson, H.G. (eds) *English in the world*. Cambridge: Cambridge University Press.

Lambert, W.E., Hodgson, R.C., Gardner, R.C. and Fillenbaum, S. (1960) 'Evaluational reactions to spoken languages', *Journal of Abnormal and Social Psychology*, 60(1), pp. 44–51.

Lee, S., Jeon, J., McKinley, J. and Rose, H. (2025) 'Generative AI and English language teaching: a global Englishes perspective', *Annual Review of Applied Linguistics*, 45.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E. and Zou, J. (2023) 'GPT detectors are biased against non-native English writers', *Patterns*, 4(7), 100779.

Rosa, J. and Flores, N. (2017) 'Unsettling race and language: toward a raciolinguistic perspective', *Language in Society*, 46(5), pp. 621–647.

Rose, H. and Galloway, N. (2019) *Global Englishes for language teaching*. Cambridge: Cambridge University Press.

Siemund, P., Al-Issa, A. and Leimgruber, J.R.E. (2021) 'Multilingualism and the role of English in the United Arab Emirates', *World Englishes*, 40(2), pp. 191–204.

Sourati, Z., Ziabari, A.S. and Dehghani, M. (2026) 'The homogenizing effect of large language models on human expression and thought', *Trends in Cognitive Sciences*.