Chapter 2

Language and Society

Why do you talk the way you do?

You've lived somewhere—perhaps many somewheres. Your friends or family have influenced you. You've probably even thought about how you like some linguistic features and want to avoid others. People have long been aware that these factors influence why different groups speak differently, but the systematic study of dialect began in the eighteenth and nineteenth centuries, as part of the same scientific movements that gave us the Linnaean catalogue of the living world and the periodic table of the elements. While some cataloguers set out with nets to study butterflies, or burned candles inside jars to distill gases, others pored over ancient scripts and compiled lists of verbs.

Maps

But what kind of net can you use to capture living language? A German dialectologist named Georg Wenker thought he had an answer: he sent out a postal survey to schoolteachers across German-speaking Europe and asked them to translate forty sentences (such as "I will slap your ears with the cooking spoon, you monkey!") into the local vernacular. It was a wise enough idea: teachers would be guaranteed to be able to read and write, and even if Wenker didn't know the name of every single village teacher, surely the village post office in, say, Quedlinburg, could pass on his letter to the Quedlinburg village school. But in order to make it easy for the schoolteachers to respond, Wenker didn't provide them with any training in phonetic notation. This meant that if one teacher wrote "Affe" (monkey) and another teacher wrote "Afe" or "Aphe," it was anyone's guess whether they were trying to represent the same pronunciation.

A French linguist named Jules Gilliéron thought he had a better method. Rather than send out letters like Wenker, he'd send out a trained fieldworker to administer all the surveys. Back in Paris, Gilliéron could get a start on analyzing the results as they came in. The fieldworker he selected was a grocer named Edmond Edmont, who reportedly had a particularly astute ear (it's not clear whether this referred to the acuity of his hearing or his attention to phonetic detail, but either way, it got him the job). Gilliéron trained Edmont in phonetic notation and sent him off on a bicycle with a list of 1,500 questions, such as "What do you call a cup?" and "How do you say the number fifty?" Over the next four years Edmont cycled to 639 French villages, sending results back to Gilliéron periodically. In each village, he interviewed an older person who had lived in the region for their entire life, counting them as representative of the history of the area.

Both Wenker's and Gilliéron's dialect maps are meticulous, fascinating, and complicated, but if you know how to read them, you can trace the line between the villages in the north where French people around 1900 called Wednesday *mercredi* and those in the south where they called it *dimècres*.* Or you can read Wenker's hand-drawn map of Germany showing which regions pronounced "old" as *alt*, *al*, or *oll*.† If you studied French or German in school, it's easy to think that they're each a single, unitary language, but that's just the formal version: the maps showcase how these languages are truly constellations of dialects, hundreds of varieties that differ slightly from village to village.

But these spectacular linguistic atlases are also limited. If Edmond Edmont, towards the end of his four-year odyssey, realized that different regions also had different words for *bicycle*, he either had to bike back through those same 639 French villages, or make a note of it and just hope that some future scholar would undertake a second linguistic Tour de France. Georg Wenker's project was almost too successful: he ended up with more than 44,000 completed surveys between 1876 and 1926, more than he could possibly analyze by hand. (His colleagues continued analyzing his results for decades after his death.)

As technology advanced, so did dialectology. In the 1960s, the *Dictionary of American Regional English* sent out fieldworkers in "Word Wagons" (green Dodge vans outfitted with a fold-out bed, an icebox, and a gas stovetop) to record locals in over a thousand communities on briefcase-sized reel-to-reel tape recorders. In the 1990s, the creators of *The Atlas of North American English* let their fingers

---

*If you're wondering why *mercredi* won, well, the north is where Paris is. *Di* means "day," so either order is logical in principle, and indeed for *dimanche*, "Sunday," the *di*-first version won.

†If you think of English *ol'*, these latter two may not seem surprising.

do the biking and conducted telephone interviews with 762 random people, at least two from each major urban area. In 2002, the Harvard Dialect Survey produced a linguistic questionnaire that anyone could complete online: thanks to media coverage in *The New York Times*, *USA Today*, and many other outlets, over thirty thousand people did.

All of these studies have produced incredibly cool results: not only did they show that the rise of radio, television, and other mass media wasn't eradicating regional language variation, but they've also made many of their resources freely available online. You can go browse *The Atlas of North American English* yourself and see changing colors of dots midway through the United States where people switch from "pop" to "coke" to "soda," and then another line at the Canadian border where they switch back to "pop" again. On the *Dictionary of American Regional English* site, you can scroll through a "Word Wheel" of interesting vocabulary items, from "Adam's housecat" to "zydeco." The Harvard Dialect Survey results, downloadable in full, even found new life a decade later as the YouTube accent challenge, a viral video meme where thousands of people from around the world filmed themselves answering questions from the survey, and as the dataset at the base of "How Y'all, Youse and You Guys Talk," the massively popular *New York Times* dialect quiz that introduced many people to the idea of mapping out how you speak in 2013.

But if you've ever hung up on a telemarketer or fudged your answers to a "Which Disney Princess Are You" quiz, you know some of the potential problems with phone and internet surveys. On the phone, researchers could record audio, but they still had to have an individual conversation with each person they surveyed. While operating a Word Wagon or a linguistic phone bank is a fascinating job for the right type of language nerd (um, hi), such nerds still need to

be paid for the massive amounts of time and labor they're putting into the interviews. Internet surveys are faster and cheaper to conduct at a huge scale, but people still don't always accurately report on their own language usage.

Running through all the surveys is a problem called the observer's paradox: when you sit someone down with a tape recorder or hand them a list of questions to check off, it tends to bring out the formal, standardized, job-interview style of language, which is the least interesting linguistically because it's already so well documented. But looking into less well-documented varieties requires researchers to seek answers that they may not know the questions for, and the people they're studying are sometimes unaware or self-conscious about some of the most interesting aspects of their speech, so they can't or won't talk about it explicitly.

It's not completely hopeless: linguists have devised several methods for getting at more natural-sounding speech. One is to ask open-ended questions ("Could you describe your family?" rather than "How do you pronounce 'aunt'?"). Another is to ask about an exciting or emotional event, to get people thinking about the content rather than the words (a popular, though perhaps rather morbid, question is "Can you tell me about a time you thought you might die?"). A third is to work with a community as an insider: many a linguist has analyzed the speech of their own children, grandparents, or extended family, or else worked with a local collaborator to conduct interviews. The Word Wagon linguists would even carry small notebooks, in case they overheard any interesting language at the grocery store, so that they'd remember to follow up on it when they got the tape recorder out.

But one particularly effective way of getting at unselfconscious speech is on the internet. Not only can researchers look at countless examples of public, informal, unselfconscious language, from videos

to blog posts, but in many cases, it's also searchable. No more hours of transcribing audio files, hoping for a few examples. Twitter is particularly valuable: even the most casual of searchers can look for a word or phrase and form an impression of how people are using it. They might notice that a lot of people who used "smol" in 2018 also appeared to be fans of anime or cute animals, or that "bae" was used primarily by African Americans until around 2014, when it started appearing in tweets by white people, only to get co-opted by brands shortly thereafter.

The presence of researchers on social sites is a still-evolving ethical domain. Regardless of who technically has access to their information, people tend to have a mental model of who they expect to read their posts, and feel that their trust is violated when someone outside that model does so. When the Library of Congress announced in 2010 that they'd be archiving every single tweet, Twitter users had to update their mental models for a previously ephemeral website. Many reacted by posting tongue-in-cheek instructions or commentary to future historians. Several people took advantage of the opportunity to make the august institution expand its holdings of choice four-letter words, while others asked, "What's up, posterity?" or noted, "Please index all my kitten pictures properly under 'kitteh' as well as 'kitten' now that you're saving my tweets." Not much came of it, in the end: the Library of Congress changed course in 2017, restricting their Twitter archive to tweets that met stricter criteria of newsworthiness. A less benign social media data controversy happened in 2018, when British political consulting firm Cambridge Analytica was discovered to have obtained personal data from millions of Facebook users in 2015 by convincing people to link a personality quiz with their Facebook account. The personal data derived from the quiz was then used to target voters and potentially sway elections. The Library of Congress and Cambridge Analytica represent two extremes, but less publicized researchers have continued mining for data on social media, restricted only by terms of service and their own senses of fair play.

In this book, I have for the most part restricted my citations to social media data in aggregate, not linked to individual users, or examples which are already cited and anonymized in research papers. But where I've needed to pull out individual examples, I've aimed for those in which the writers are already clearly having a metalinguistic discussion, like the tweets addressing the Library of Congress archivists. Quoting people's innocent chatter about their lunch or deeply personal heart-to-hearts felt to me uncomfortably like spying, but quoting comments about internet language in a book about internet language is, I hope, a way of entering into a conversation. After all, if you're going to address your tweets to posterity, perhaps you shouldn't be surprised when posterity addresses you back.

Twitter research is especially fruitful because about 1 to 2 percent of people who post on Twitter tag their tweets with their exact geographic coordinates. A reasonably competent data miner can therefore code up a county-level map of where Americans tweet "pop" versus "soda," where they switch from "y'all" to "you guys," or which states prefer which swear words—all in less time than it took Edmond Edmont to bike from Paris to Marseille. As a simple proof of concept, let's look at the work of the linguist Jacob Eisenstein, who found that geo-tagged tweets containing "hella" (as in "That movie was hella long") are most likely to occur in Northern California, while those containing "yinz" (as in "I'll see yinz later") are clustered around Pittsburgh. Both of these findings are consistent with previous linguistic research done in the labor-intensive interview style. Other features he found on Twitter probably wouldn't have shown up in an interview: a later study by Eisenstein and colleagues found that the abbreviation "ikr" ("I know, right?") was especially

popular in Detroit, the emoticon ^_^ (happy) was characteristic of Southern California, and the spelling "suttin" ("something") was popular in New York City.

Some of the linguistics research happening on Twitter wouldn't be possible at all without the internet. The linguist Jack Grieve researches constructions like "might could," "may can," and "might should" in the American South—things like "We might should close the window," where speakers of other dialects would say, "Maybe we should close the window." Grieve has pointed out that as recently as 1973, prominent linguists said that it would simply be impossible to research these constructions: they're vanishingly rare in edited text, and occur maybe once an hour, if you're lucky, in a spontaneous spoken interview. That's a heck of a lot of audio to transcribe for a tiny amount of data. But on Twitter, Grieve and his collaborators combed through nearly a billion geo-coded tweets and unearthed thousands of examples. Beyond just reinforcing the informal intuition that these constructions (known as double modals) exist, they've been able to make detailed county-level maps showing that they can actually be divided into two groups: some, like "might could" and "may can," map onto the Upper South, while others, like "might can" and "might would," are more common in the Lower South.

We may even be able to discover things about various regions that we hadn't realized before. For example, after "might could," Grieve turned his attention to swear words, finding that, while people in every state swear, their preferred swear words varied. Keeping it to the somewhat milder terms, people in the American South were especially fond of "hell," while people in the northern states preferred "asshole," the Midwest used a lot of "gosh," and the West Coast liked the Britishy "bollocks" and "bloody." The *Oxford English Dictionary* has also begun using Twitter as a source of data, especially for regional words that are less often printed in books and newspapers. The dictionary's quarterly update notes for September 2017 gave the example of the word "mafted," a northeastern British term defined as "exhausted from heat, crowds, or exertion." The example quotations for "mafted" are a study in old-school and new-school lexicography: the oldest citation is from "a glossary compiled around the year 1800," and the newest is from someone on Twitter in 2010 saying, "Dear Lord—a fur coat on the Bakerloo line, she must have been mafted."

We can even use creative respellings on Twitter to investigate how people pronounce things differently. It's a little bit harder than just searching for words, but the linguist Rachael Tatman gave us an example using two well-studied sounds in varieties of English. The first is the pronunciation of words like "cot" and "caught" or "tock" and "talk." Some Americans (primarily in the West, Midwest, and New England) pronounce each member of the pairs the same, while others (primarily Southerners and African Americans) pronounce them differently—a trend which has been long established by the kind of linguists who make audio recordings. Tatman hypothesized that speakers who do have two distinct vowels in "sod" and "sawed" would sometimes want to call attention to one particular vowel, by respelling it as "aw." Sure enough, she found that in tweets where a common word, like "on," "also," and "because," was respelled with "aw," as in "awn," "awlso," and "becawse," there also tended to be respellings for other well-documented features of Southern American English and African American English, such as deleting the "r" in words like "for" and "year" (writing "foah" and "yeah") and writing "da" and "dat" for "the" and "that."

But that could just be a coincidence. To test it, Tatman looked at a completely different sound in a completely different region: the pronunciation of words like "to" and "do" as "tae" and "dae." This

sound, and this particular spelling of it, is associated with Scottish English, and has been since Robbie Burns. Here again, Tatman found that people who tweet this respelling tend to show other linguistic markers of Scottishness: they also tweet respellings like "ye" for "you" and "oan" for "on." To be sure, not all Scots, Southerners, or African Americans use these respellings, and those that do don't use them all the time. But the point is, when we respell words in casual writing, we tend to do so with a purpose—we jump in with both feet and try to represent our whole manner of speaking. Even if it's not always this clear which sounds are intended by a particular respelling, looking at which words and sounds people respell can help give linguists an idea of where to focus their audio recording energy.

The internet lets linguists do the kinds of dialect mapping and analysis of spontaneous speech that we've been trying to do for centuries, but with more data, from the comfort of a laptop, and without distorting the data by observing it. Just as the telephone study showed that people were still talking like their neighbors rather than like TV and radio broadcasters, and the bicycle peregrinations showed that regional dialects persisted even after centuries of print standardization, the internet studies show us that we often keep our local ways of speaking when we use social media. Our deep wells of enthusiasm for internet dialect quizzes give us a clue about why: talking in particular ways reinforces our networks, our sense of belonging and community.

Networks

Does it ever feel like your family or friend group speaks its very own dialect? This was the premise of a book called *Kitchen Table Lingo*, which collected examples from what the linguist David Crystal

called familects: "the private and personal word-creations that are found in every household and in every social group, but which never get into the dictionary" (or onto dialect maps). The book's initial appeal for "familect" words attracted thousands of submissions from around the world, with stories of misheard song lyrics, onomatopoeia, children's coinages, and no less than fifty-seven words for the TV remote control. Dialect maps are just the beginning of our linguistic differences: every time we talk with some people more than others, we have the chance to develop a shared vocabulary, whether that's families, friends, schools, workplaces, hobbies, or other organizations. Family dialects are often inspired by a cute word that comes out of a kid's mouth (Queen Elizabeth II was apparently nicknamed "Gary" by a young Prince William, who was unable to say "Granny" yet), but the peak importance of in-group language happens at a later life stage: teenagehood.

High school is a place where people really notice small social details, whether that's the cool brand of jeans, who's now going out with who, or vowels. The linguist Penelope Eckert embedded at a high school in the Detroit suburbs in the 1980s to study the correlation between language and high school cliques. She found two main groups: jocks, who participated in the power structure of the school through activities like varsity sports and student council, and burnouts, who rejected the school's authority. In Detroit, along with many other American cities around the Great Lakes, there's a vowel change going on, where some speakers say "the busses with the antennas on top" in a way that sounds to people outside the area like "the bosses with the antennas on tap." For Eckert's students, the "bosses" pronunciation had a connotation of "street smarts," so the burnouts were more likely to use it than the jocks—despite the fact that they all lived in the same neighborhood and attended the same school, and irrespective of the social class of their parents. You could

arrange the students into more subtle cliques, from "burned-out burnouts" to "jock-jocks," and their vowels would follow suit. To put it in the terms of classic high school movie characters, if Eckert's high school was Rydell High from *Grease*, we'd expect Sandy to say "bus," Rizzo to say "boss," and Frenchy to be somewhere in between.

Further studies at other high schools show other groups with other linguistic attitudes. A group of girls in California identified as nerds, and rejected the jock-burnout dichotomy altogether: linguistically, they avoided the slang and cool vowels developing among their peers (such as pronouncing the word "friend" as "frand"), because they didn't want to be heard as caring about high school popularity. Instead, they adopted linguistic features linked to intellectualism, such as hypercareful articulation, long words, and puns. A study of Latinas at another California high school found a linguistic distinction between Norteñas, who identified as American or Chicana and generally spoke English, and Sureñas, who identified as Mexicana and generally spoke Spanish. We could keep going, but let's pause and think about how we develop our senses of what's cool in the first place.

Remember how you learned about swearing? It was probably from a kid around your age, maybe an older sibling, and not from an educator or authority figure. And you were probably in early adolescence: the stage when linguistic influence tends to shift from caregivers to peers. Linguistic innovation follows a similar pattern, and the linguist who first noticed it was Henrietta Cedergren. She was doing a study in Panama City, where younger people had begun pronouncing "ch" as "sh"—saying *chica* (girl) as *shica*. When she drew a graph of which ages were using the new "sh" pronunciation, Cedergren noticed that sixteen-year-olds were the most likely to use the new version—more likely than the twelve-year-olds were. So did that mean that "sh" wasn't the trendy new linguistic innovation after all,

since the youngest age group wasn't really adopting it? Cedergren returned to Panama a decade later to find out. The formerly untrendy twelve-year-olds had grown up into hyperinnovative twenty-two-year-olds. They now had the new "sh" pronunciation at even higher levels than the original trendy cohort of sixteen-year-olds, now twenty-six-year-olds, who sounded the same as they had a decade earlier. What's more, the new group of sixteen-year-olds were even further advanced, and the new twelve-year-olds still looked a bit behind. Cedergren figured out that twelve-year-olds still have some linguistic growth to do: they keep imitating and building on the linguistic habits of their slightly older, cooler peers as they go through their teens, and then plateau in their twenties.

In terms of swearing, that's like saying some twelve-year-olds swear, but a lot more sixteen-year-olds do. But swearing is very socially salient (we have laws about it!) and not really changing that much. It's been peaking in adolescence and declining through adulthood for decades. The other trendy linguistic features that we acquire in adolescence (new pronunciations like "bosses" and "shica," and innovative uses of words like "so" and "like") are a case of subtle social discernment rather than massive social taboo, and so we tend to keep them as adults.

This age curve is important when we think about when young people start using social media: age thirteen, if you believe the terms of service of most sites and apps, or slightly younger, if you assume that some users lie about their ages. This is right at the beginning of the age range when the language of teens is tremendously influenced by the slang of their peers. Sure, little kids play games and watch videos and even ask questions of voice assistants, but their social lives are still mediated by their families and their reading level. This coincidence of peer influence and social media access means

that it's easy to conflate how the youth are talking now with the tools that they're using to do so. But every generation has talked slightly differently from its parents: otherwise, we'd all still be talking like Shakespeare. The question is, how much of that is influenced by technology, and how much is the linguistic evolution that would have happened regardless?

The answer seems to be that both happen simultaneously. Researchers from Georgia Tech, Columbia, and Microsoft looked at how many times a person had to see a word in order to start using it, using a group of words that were distinctively popular among Twitter users in a particular city in 2013–2014. As we'd expect, they noticed that people who follow each other on Twitter are likely to pick up words from each other. But there was an important difference in how people learned different kinds of words. People sometimes picked up words that are also found in speech—like "cookout," "hella," "jawn," and "phony"—from their internet friends, but it didn't really matter how many times they saw them. For rising words that are primarily written, not spoken—abbreviations like "tfti" (thanks for the information), "lls" (laughing like shit), and "ctfu" (cracking the fuck up) and phonetic spellings like "inna" (in a / in the) and "ard" (alright)—the number of times people saw them mattered a lot. Every additional exposure made someone twice as likely to start using them. The study pointed out that people encounter spoken slang both online and offline, so when we're only measuring exposure via Twitter, we miss half or more of the exposures and the trend looks murky. But people mostly encounter the written slang online, so pretty much all of those exposures become measurable for a Twitter study. The researchers also found that you're more likely to start using a new word from Friendy McNetwork, who shares a lot of mutual friends with you, and less likely to pick it up from Rando McRandomFace, who doesn't share any of

your friends, even if you and Rando follow each other just like you and Friendy do.

But these networks aren't formed in isolation: people tend to follow others with similar interests and demographics. One study demonstrating this looked at the geographic spread of a couple thousand words that became massively more popular on Twitter between 2009 and 2012. It found that terms tended to leapfrog from one city to another based on demographic similarity, not just geographic proximity. So slang would spread between Washington, D.C., and New Orleans (both have high proportions of black people), Los Angeles and Miami (high proportions of Hispanic people), or Boston and Seattle (high proportions of white people), but not necessarily the cities in between. For example, the abbreviation "af" for "as fuck" (as in "word maps are cool af") starts out at low levels in Los Angeles and Miami in 2009, then spreads elsewhere in California, the South, and around Chicago in 2011–2012, suggesting that it was spreading from Hispanic to African American populations. The study stops there, but we can continue: in 2014 and 2015, "af" started appearing in *BuzzFeed* headlines, a decent measure of when it came to be coopted by mainstream brands capitalizing on its association with African American coolness.

We're especially likely to pick up new words when we're first entering a community. Linguist Dan Jurafsky and his colleagues looked at over four million posts from members of RateBeer and BeerAdvocate, two online beer communities that have been around for more than a decade. They wanted to know how people's language use changed the longer they'd been members of the forum. They found that older accounts were likely to stick to older pieces of beer jargon, such as talking about a beer's "aroma" if they joined in 2003, whereas younger accounts were quicker to adopt newer beer jargon, such as preferring "S" (for "smell") if they joined in

2005. The study provides an interesting way of teasing apart the effects of age and peer groups, suggesting that people are more open to new vocabulary during the first third of their lifespan, regardless of whether that's an eighty-year lifespan in an offline community or a three-year "lifespan" in an online one.

What's unique about adolescence, then, may not be our susceptibility to linguistic trends. Rather, it's the last time that a whole population is entering a new social group all at once. Adults periodically move to new cities and start new jobs and develop new hobbies, all of which bring us under new linguistic influences. But we don't all change careers or become parents or join beer-tasting messageboards at exactly the same age, so it's harder to study linguistic changes that happen later in life. Harder, but not impossible: it also depends on where we want to look. Researchers are part of society, and as a society, we're more likely to be worried about teen slang than about parents adding new terms to the familect or businesspeople adopting new corporate buzzwords. Perhaps we need to rethink our demographic questions to ask about dates of joining new social groups in addition to date of birth.

Finding networked language patterns on social media isn't an anomaly: people offline are generally also more similar to their friends than to the rigid, unfeeling demographic boxes of a census-taker. It's just that we had no practical way of measuring it. Doing a network analysis of people's friends and interlocutors used to be *really hard*. Like makes-biking-around-France-for-four-years-look-easy kind of hard. You could start by doing a typical language survey, but that would just be the beginning of your work. You'd also have to get people to manually make a list of all their friends, how long they've known them, and how often they talk with each one. Then, you'd have to somehow get ahold of all these friends and also survey them.

But that's just a one-layer network. You'd want to repeat these steps several times so that you could make webs of connections between people. Social scientists have done this kind of research occasionally—there's a city in Massachusetts called Framingham where researchers have followed a couple thousand people, with their health and social connections, for three generations now—but understandably, they don't do it very often. Not for daily words produced by tens or hundreds of thousands of people. Even though your Twitter network doesn't represent absolutely everyone you talk to, even though not everyone is on Twitter, it makes for an intriguing new way of approaching the very old question of how new words catch on.

Analyzing language based on social networks also complicates another traditional demographic check box: gender. The traditional finding for gender is shown in a study by the linguists Terttu Nevalainen and Helena Raumolin-Brunberg at the University of Helsinki, which looked at six thousand personal letters written in English between 1417 and 1681. Personal letters make a great corpus because, like tweets, they don't go through editorial standardization. Unfortunately, there's also a lot fewer of them, and they tend to overrepresent the leisured, educated classes. But they're still the best record we have of what day-to-day English looked like back then. The linguists examined fourteen language changes that occurred during this period, things like the eradication of "ye," the switch from "mine eyes" to "my eyes," and the replacement of *-th* with *-s*, making words like "hath," "doth," and "maketh" into "has," "does," and "makes." (Pretty shocking stuff.) For eleven out of the fourteen changes, Nevalainen and Raumolin-Brunberg found that female letter-writers were changing the way they wrote faster than male letter-writers. In the three exceptional cases where the men were ahead of the women, those particular changes were linked to men's greater access

to education at the time. In other words, women are reliably ahead of the game when it comes to word-of-mouth linguistic changes.

Research in other centuries, languages, and regions continues to find that women lead linguistic change, in dozens of specific changes in specific cities and regions. Young women are also consistently on the bleeding edge of those linguistic changes that periodically sweep through media trend sections, from uptalk (the distinctive rising intonation at the end of sentences?) to the use of "like" to introduce a quotation ("And then I was like, 'Innovation' "). The role that young women play as language disruptors is so clearly established at this point it's practically boring to linguists who study this topic: well-known sociolinguist William Labov estimated that women lead 90 percent of linguistic change in a paper he wrote in 1990. (I've attended more than a few talks at sociolinguistics conferences about a particular change in vowels or vocabulary, and it barely gets even a full sentence of explanation: "And here, as expected, we can see that the women are more advanced on this change than the men. Next slide.") Men tend to follow a generation later: in other words, women tend to learn language from their peers; men learn it from their mothers.

What's less clear is why. Lots of reasons have been proposed, from the fact that women still dominate the caregiving of children in the societies studied, that women may pay more attention to language to compensate for relative lack of economic power or to facilitate social mobility, and that women tend to have more social ties. But in many cases, gender (like age) seems to be a proxy for other factors related to how we socialize with each other.

Several internet studies have highlighted the importance of differentiating between gender and social context. One study, by linguists Susan Herring and John Paolillo, looked at how people write blogs. At first, it seemed like there was a significant gender difference in the language of blogs. But when they looked again, the linguists found that what was really going on was a *genre* difference: men were more likely to write topic-based blogs and women more likely to write diary-style blogs. But of course, there were also many people who didn't pick the genre most typical for their gender. When the researchers compared within each genre, the original "gender" difference disappeared.

Another study, looking at a corpus of 14,000 Twitter users, and guessing their gender based on the skew of their first name in census data, appeared at first glance to show clear gender differences: people with predominantly female names were more likely to use emoticons, for example, while people with male-associated names were more likely to swear. But when the researchers looked one step further, they found that the words people most often tweeted formed natural clusters into over a dozen interest groups, such as sports fans, hip-hop fans, parents, politics buffs, TV and movie fans, techies, book fans, and so on. True, many of the groups had a gender skew, but none of them were absolute, and they also had clear associations with other demographic factors like age and race. Sometimes whole groups defied gender norms—men overall tended to swear more, but techies, a cluster that was male-dominated, didn't swear much at all, presumably because they were using Twitter as an extension of the workplace. At the individual level, people followed the norms of their clusters rather than their genders—a woman in the sports cluster or a man in the parenting cluster tweeted like their fellow sports fans or parents, rather than like an "average woman" or "average man." Moreover, restricting the analysis to accounts with names that showed a clear gender skew in census data excludes precisely those users that would complicate a binary view of gender, including nonbinary people and others who've deliberately chosen a non-census-gendered username.

Offline, ethnographic research has also pointed to the importance of network factors. Linguist Lesley Milroy was doing a pretty standard study of language change in a couple working-class neighborhoods of Belfast, Northern Ireland. As with many communities, the young women were leading a linguistic change—in this case, changing the vowel in "car" to sound more like "care." This vowel is common elsewhere in Northern Ireland, but it was new to this particular community, and it was the young women who were bringing it in. What was mystifying was *how they were getting it*. When Milroy asked the women who they were close to, they named friends, family, and coworkers, all from their neighborhood—the same neighborhood where no one else yet had this vowel change.

In a later paper with James Milroy, the two figured out why by linking linguistic change to another concept in social science: strong and weak ties. Strong ties are people you spend a lot of time with and feel close to, who you share mutual friends with; weak ties are acquaintances who you may or may not share mutual ties with. In the case of the Belfast study, the early-adopting young women all worked at the same store in the city center, where people were already using the new vowel. Although they didn't have close friends from the city center, they did have weak-tie contact with customers, which would have often exposed them to the new vowel—more than the young men of their neighborhood, who weren't employed outside it.

Milroy and Milroy figured that, just as your weak ties are a greater source of new information like gossip and employment opportunities than your close friends who already know the same things you do, more weak ties also lead to more linguistic change. To demonstrate, they compared the history of English and Icelandic. English and Icelandic have a common Germanic ancestor, and a millennium ago Old English and Old Norse (the ancestor of Old Icelandic spoken at the time) were still more or less mutually intelligible. But from there, their histories diverge. Icelandic has changed only a little: twenty-first-century Icelandic speakers can still read their Sagas from the thirteenth century, written in Old Icelandic, without much difficulty. English has changed a lot: although we can manage Shakespeare, from only four centuries ago, with the help of footnotes, even *The Canterbury Tales* (six centuries ago) requires a full translation or a course in Middle English to understand. This means that, despite the fact that it's technically written in Old English rather than Old Icelandic, Icelanders would have an easier time learning to read *Beowulf* than would modern English speakers.

Clearly, English has changed faster than Icelandic has over the same timespan. Milroy and Milroy proposed that the reason is weak ties. The thing to know about Iceland is that it's got really close-knit communities. Icelandic surnames are still based on the given name of your father (or sometimes mother), which makes a lot more sense in a society where most of the people you meet already know your family, and this tendency to introduce oneself by naming an extensive network of relatives dates all the way back to the Sagas. If everyone you know already knows each other, your only source of new linguistic forms is random variation—you don't have any weak ties to borrow from.

English, on the other hand, has had several significant sources of weak ties over its history—invasions by the Danes and the Normans, a tradition of uprooting and moving to London and later other cities to seek one's fortune, and imperial expansion of its own. True, the English-speaking world has its own small, tight-knit communities where everyone knows everyone else's relatives (I still introduce myself by referring to my parents or grandparents at family reunions), but it also has many more big cities where you can be anonymous in a crowd or have three different friend groups who never meet each

other. What's more, the map studies from the beginning of this chapter tell us that within English, it's the bigger, looser-knit cities that give rise to more linguistic change.

But weak ties can't be the only factor. After all, it's also clear that we talk like people in our social circles, whether that's French villages, Detroit jocks, or familects—all examples of strong ties. How can both strong and weak ties be responsible for how we speak? And how can we map out exactly who says what to who over a large population for a couple centuries, long enough for several changes to run their course? That's not just bicycling—that's time travel.

Linguist Zsuzsanna Fagyal and colleagues solved both problems using a computer simulation. They made a network of nine hundred hypothetical people over forty thousand turns. Each person had a certain number of ties to other people in the network and started with a randomly assigned value for a hypothetical linguistic feature, like how you might call the thing you drink water from in a school a "water fountain" but your neighbor might call it a "drinking fountain." Then, at each turn, each person looked to the other people they were connected to and had a certain probability of adopting their version of the feature, like how you might start saying "drinking fountain" if you have a friend who uses the term. If you do pick it up, that word now becomes yours as well, and the people you're connected to might pick it up from you the next round. They repeated this turn process forty thousand times, with three different kinds of networks. In one version, the entire network was made up of close ties: everyone was well connected to the rest of the network. This dense network behaved like Iceland: one linguistic option caught on very quickly and stayed completely dominant for the rest of the simulation. In another version, the entire network was made up of weak ties and no one was well connected. The loose network behaved like a world of tourists: all of the options stuck around and

none of them ever became dominant. But in the most interesting simulation, they made some of the nodes highly connected "leaders" and others less connected "loners." This mixed network behaved like English: one option would catch on for a while, but the other options would never totally disappear, and eventually one of them would become popular instead—a cycle that repeated several times. The researchers concluded that both strong and weak ties have an important role to play in linguistic change: the weak ties introduce the new forms in the first place, while the strong ties spread them once they're introduced.

The internet, then, makes language change faster because it leads to more weak ties: you can remain aware of people who you don't see anymore, and you can get to know people who you never would have met otherwise. The phenomenon of a hashtag or funny video going viral is an example of the power of weak ties—when the same thing is shared only through strong ties, it ends up merely as an inside joke. But the internet doesn't lead to the collapse of strong ties, either: the average person has a small handful of people who they message on a regular basis, between four and twenty-six, depending on how you count. What's more, social networking sites that prompt you to interact with denser ties—people you already know and friends of friends—tend to be less linguistically innovative. It's not an accident that Twitter, where you're encouraged to follow people you don't already know, has given rise to more linguistic innovation (not to mention memes and social movements) than Facebook, where you primarily friend people you already know offline.

But geography and demographics and even networks aren't destiny. In addition to having some amount of choice in where we live and who we associate with, we also have a certain amount of control over how much we want to be influenced by our interlocutors: who we want to project ourselves to be, linguistically speaking.

Attitudes

If you want to sum up Canada in a headline, you might reach for the catchphrase "from Eh to Zed." You'd be in good company: this slogan features in the titles of three books, items like t-shirts and YouTube videos, and news articles about everything from sports to the language itself. But what many people don't think about, even Canadians, is that small Canadian children often call the last letter of the alphabet "zee" instead. Normally, when linguists see a word or construction that's common among parents but not their kids, we simply conclude that there's a change going on—that in another generation it'll be a grandparent-y sort of word, and eventually pass into history. "Chesterfield" is doing exactly this in Canada: it's been receding for decades in favor of "couch."

But "zed" has been acting really weird. The linguist J. K. Chambers did a survey of Canadian twelve-year-olds in the 1970s, and found that two-thirds of them said "zee"—but when he went back and surveyed the same population in the 1990s, he found that the vast majority were now using "zed" as adults. The same shift happened with successive generations. Chambers figured that children learn "zee" from the alphabet song and American children's television programs like *Sesame Street*, but when they get older, they learn that "zed" is associated with Canadian identity and switch. Indeed, noted Chambers, "zed" is one of the first things that American immigrants to Canada change about their speech, "because calling it 'zee' unfailingly draws comments from the people they are talking to."

I first learned about this phenomenon when I was eighteen, in a linguistics class that I took about Canadian English in Kingston, Ontario. It stood out for me among the sea of dialect maps and survey methodologies because I realized that I'd done this exact thing. I was

a child who sang "zee" at the end of the alphabet song in the 1990s, until some point around middle school when I switched to using "zed" consistently. What's more, I was still slightly embarrassed by this fact and had done my best to put it out of my mind, because clearly I should never have been using the un-Canadian "zee" in the first place. When I realized that I'd done this, I asked my mom what she called the last letter of the alphabet. I've only ever known her as a zed-sayer, but apparently she'd done the same shift long before I was born. My switch from "zee" to "zed" happened at about the same age that I started consistently using Canadian spellings in words like "centre" and "colour" instead of "center" and "color." I don't recall anyone telling me to do it, but I do recall it being a conscious choice, fueled by that exact sense of social identity that Chambers described. At the time, acquiring a sense of linguistic nationalism was a way of going with the flow, of following the dominant usage of my parents and teachers. In adulthood, especially on the internet, I use Canadian spellings in my posts and messages partly out of habit, but partly also because it goes *against* the flow: it's a subtle way of reclaiming space against the idea that all English speakers on the internet fit neatly into the choice I've faced in so many dropdown menus between "American" and "British."

We all make linguistic decisions like this all the time. Sometimes, we decide to align ourselves with the existing holders of power by talking like they do, so we can seem rich or educated or upwardly mobile. Sometimes, we decide to align ourselves with particular less powerful groups, to show that we belong and to seem cool, anti-authoritarian, or not stuck-up.

The most legendary study of social factors in language differences is about how much people of different social classes use the stereotypical "New Yawk" accent, with the R dropped from after the vowel. In November 1962, linguist William Labov went into various

department stores in New York City and asked how to find something—the shoe section, for example—that he already knew was on the fourth floor. The salesperson would reply "fourth floor" or "fawth flaw" and then Labov would pretend not to have heard, getting the salesperson to repeat the location more carefully. After this exchange, Labov would head off in the appropriate direction, but not to buy shoes. As soon as he was out of sight, he'd pull out a notebook and record whether the salesperson pronounced the R in "fourth" and "floor." He found that, sure enough, the salespeople at the fanciest department store, Saks Fifth Avenue, said R more than those at the mid-range one (Macy's), who in turn used R more than the bargain store (the now defunct Klein's), and that people also tended to pronounce R more in careful speech, when he asked them to repeat themselves, than they had the first time around. But it's hard to shop at Saks Fifth Avenue on a retail salary: the salespeople themselves came from similar class backgrounds across all three stores. Instead, it was their perception of the kind of customer they catered to that made the difference, even though Labov took pains to report that he dressed the same in all places: "in middle-class style, with jacket, white shirt and tie, and used my normal pronunciation as a college-educated native of New Jersey (r-pronouncing)." (One assumes that the pronunciations might have varied even further if he'd dressed up or dressed down.)

But where did our ideas of what sounds upper or lower class even come from? In New York City, the R-less pronunciation is less prestigious. Although it's a feature of many American varieties, such as Boston English, African American English, and Southern American English, it's not favored in media. When people in the United States talk about "losing an accent," they often mean gaining an R in words like "fourth floor."

If we hopped across the Atlantic and did the same study in British

department stores, however—if we went to, say, Harrods and Debenhams and Poundland—we could find the inverse. Salespeople at Harrods, the poshest of the posh, would have no Rs at all, whereas staff at Poundland, where (almost) everything costs a pound, might have Rs if we picked our city carefully, such as Bristol or Southampton. R-ful varieties are found in parts of Britain, including Scotland and Northern England, but they're not favored in London or on the BBC. English speakers don't all talk like our books and media any more than actual French and German speakers talk like the model dialogues in language-learning textbooks. When people in the UK talk about "losing an accent," they often mean losing the R in words like "fourth floor."

Clearly, it's not R's fault. R is a harmless consonant that never asked to be embroiled in any of our petty human squabbles. Rather, it's what we take R to mean in different contexts. It's like how blue can signal a sports team, a cold-water setting, a hyperlink, a period in the life of Picasso, and so on. R in itself is neither good nor bad: its meaning, and the meaning of the accents that do or do not have it, is constructed by society. Like how money is just squiggles on paper or on a screen until it determines whether you can eat lunch, words are just meat twitches until they determine whether you can get a job—or whether someone will even deign to tell you where the shoe section is. If we all woke up tomorrow and decided that every single vowel sounded better with an R after it, we could make it happen. (Ermargerd, whart ar world thart wourld ber.)

But we don't generally wake up and decide to change our minds about R. Instead, we get our social linguistic cues from the people and power dynamics around us. One vivid example of this power dynamic comes via James Milroy, who we last saw comparing social networks in England and Iceland. In the story of a language, just like everywhere else, history is written by the winners: Milroy recounts

a typical attitude from an influential historian of English named H. C. Wyld in 1927, who "was quite insistent that the only worthy object of our study was Received Standard English. . . . The language of 'the Oxford Common Room and the Officers' mess' is an appropriate object of study, whereas that of 'illiterate peasants' is not." You practically want to reach back through time and punch the elitism.

Wyld wasn't the first linguistic elitist: before there was the elite Oxford Common Room, there was the Roman forum. The Romans, good at roads and aqueducts and armies, also left a legacy of writing: for over a millennium after the fall of the Roman Empire, if you were educated, you learned Latin. To be an English writer in the era when formal writing was shifting over from Latin to English was to be a self-hating English writer: anything you could do to make English more Latin-like would also make it better. Robert Lowth, who wrote a widely used English grammar in 1762, culled examples of so-called false syntax from luminaries of English writing like Shakespeare, Milton, and the King James Bible—not as hints that perhaps English grammar was actually just fine as it was, but as cautionary tales about how even the greats should have been more Latin-y.

It was like a competition to see who could be the most uptight. Lowth gave us an early suggestion against the sentence-ending preposition: "This is an Idiom which our language is strongly inclined to; it prevails in common conversation, and suits very well with the familiar style in writing; but the placing of the Preposition before the Relative is more graceful, as well as more perspicuous; and agrees much better with the solemn and elevated Style." Lowth himself wasn't completely against it (after all, he used it himself in "strongly inclined *to*"), just passing an aesthetic judgment. But later grammarians elevated this preference into a full-on ban, and by similarly specious reasoning objected to infinitive splitting and "they" as singular, despite centuries of prior English usage. The same Latin-worshipping

tradition was responsible for adding superfluous silent letters to words like "dete," "samoun," and "iland," because "debt," "salmon," and "island" look more like Latin "debitum," "salmonem," and "insula." Never mind that "island" doesn't even come from Latin, or that generations of schoolchildren would now have to go to extra effort. Many languages can't have spelling bees because their spelling systems are so logical that no one would ever get knocked out. English spellers can only dream!

We could almost feel sorry for the depths of self-loathing that these grammarians must have felt, to be so determined to replace their own language's forms with that of another, if it weren't for how they infected us with it as well. While they didn't wholly succeed at the grammatical side, especially in speech and among skilled writers who trusted their own ear or felt they knew enough to break the rules, they did leave us with a vague sense of unease at the whole prospect of the written word. Even after years of writing, most of us have a hard time trusting what we naturally think sounds like a reasonable English sentence, haunted as we are by the ghosts of misguided grammarians.

But while modern linguistics has moved on, and even modern writing manuals are scraping off the heavy lacquer of Latinization with more or less enthusiasm, we've acquired a new form of linguistic authority on our digital devices. Tools like spellcheck, grammarcheck, autocomplete, and speech-to-text impose someone's ideas of the rules of English automatically—invisible authorities that we can defy but not avoid. If a writing handbook like Lowth's or Strunk and White's displeases you, you can throw it across the room or leave it to gather dust, but when you want to type a word that's not in a predictive text model, you'll fight for every letter. In her book *Fixing English*, Anne Curzan describes how Microsoft Word's grammarcheck continues to perpetuate this same kind of discredited, Latin-based style advice and

how her colleagues in the English department, while considering themselves sufficiently expert in writing to ignore or turn off the green squiggles, had still never wondered where the grammar advice came from. If English professors who question the authority of texts for a living haven't thought to question the origins of their invisible electronic grammarians, what possible hope do the rest of us have?

Language features are not neutral in the way that the calculator feature is neutral. "Standard" language and "correct" spelling are collective agreements, not eternal truths, and collective agreements can change. Communication tools that expose us to more people may speed up the spread of new words, but tools that aim to help us with language can also slow down natural linguistic evolution by nudging us towards the versions that have already been programmed into the device.

I'm convinced that spellcheck is responsible for people's consistent misspelling of my surname: my spelling, "McCulloch," is never found in spellcheck by default, but the very similar name "McCullough" is always there instead, and when people misspell my name on a computer, they always pick the spellcheck version. Conversely, people occasionally misspell my first name, Gretchen, when writing by hand, but never do so when spellcheck is available. It seems that my names belong to two different classes of digital citizenship: one supported by the machine and the other rejected by it. This might seem relatively harmless given my German first name and Scottish surname, but I expect that if we looked at which names are found in autocorrect and autocomplete, we'd find that typical English names would be well represented and names from other languages less so. At a societal level, it's a case of bias-laundering through technology that serves to reinforce people and names that are already powerful.

Default computer spellings are powerful enough to have created a shift in British English since the 1990s: while American English prefers a Z in words like "organize" and "realize," British English has

traditionally used both *-ise* and *-ize* spellings. But spellchecks have tried to prevent people from spelling the same word differently within the same document by enforcing "organise" and "realise" all the time when set to British English, leading to an upswing in *-ise* endings among the general British typing public and the perception that *-ize* is only for Americans.

In writing this book, I'm therefore very aware that upholding the old-school Latin worship is a political decision, just as it would be if I decided to go full-on grammar anarchist. I think it's important to be upfront about such things, especially in an age when everything from books to tweets may later be mined to prove how common or acceptable a particular usage was at a particular time. Yes, I'm writing for you, the reader, but in another sense we're all writing for the unblinking eye of Data. If the most enduring legacy of this book is the slight shifting of a point on a line graph in some yet unborn person's analysis of this decade in the English language, I want to be deliberate about which direction I'm shifting that point in. What I've seen from several editors and lexicographers is the realization that we're becoming trapped in a loop: dictionaries and writing manuals refer to edited prose in order to determine what is "standard" English, but the creators of such prose refer back to the same dictionaries and manuals in their editing, each waiting for the other to move first. I've decided to play my part in correcting for this bias by opting for the more innovative direction wherever I perceive a choice: going towards where I think edited English prose will be by the end of the century, catering to the reader of the future rather than the reader of the past. As a reader and analyst of data myself, I get a joyful thrill every time I zoom out on the English language and realize that we're somewhere in the middle of its story, not at the beginning or end. I don't know how we'll be writing in the twenty-second century, but I feel a responsibility to

help its linguists gain a broad cross-section of the language of the twenty-first by not lingering overlong in the twentieth.

To that end, I've chosen to lowercase "internet" and social acronyms like "lol" and "omg" and to write "email" rather than "e-mail," and when I've needed to make a decision on other spelling choices, I've looked up which ones are more common in the Corpus of Global Web-Based English and tweets by ordinary people rather than which ones are favored by usage manuals, which has led me to close many compound words. (While I was working on this book, the Associated Press switched its recommendation from "Internet" to "internet," so I have every expectation that any similar judgment call I make will seem boring within a decade.) I've adopted the retronym "networked computers" for what were formerly called small-i internets, and I talk about "websites" rather than trying to insist on a distinction between "the internet" and "the worldwide web" which is no longer active for younger and nontechnical users. (I avoid the now dated-sounding "the Web" or "the Net" entirely, and reserve "cyberspace" for jocular historical use.) I've also included a substantial proportion of absolute time references rather than relative ones, aiming to be precise about whether I think something is true of the early twenty-first century, the 2010s, a specific year, and so on, rather than saying "now" or "currently" and requiring readers to flip to the copyright page and subtract a year or two for preparation, as I've had to do many times when reading other sources. I've freely used the singular "they," and split what infinitives needed splitting, and preserved all spelling and typographical choices found in quotes from other people, but I've otherwise kept to standard bookish spelling and capitalization and punctuation, and even suffered to have my Canadian spellings changed for US audiences. But, although it's common internet usage, I have not lowercased names of internet companies and platforms like Facebook and Twitter and YouTube.

Despite my many objections, I still use spellcheck and predictive text. Most of the time, they're pretty useful! I don't have to remember the c-to-s ratio in "necessary" or the exceptions to the "i before e" rule, which surely frees up valuable brain cells, and I can simply add words like lowercase "internet" to my phone's dictionary. But I also wonder what a world would look like where none of us cared about such things in the first place. From a linguistic perspective, all varieties are equally worthy: every language and dialect is just as much a manifestation of the incredible human language ability that is our birthright as a species. You wouldn't say that some birds aren't singing right just because they're lower in the (ahem) pecking order. No more are certain ways of speaking inherently inferior. Could we not put our tremendous computing power (both human and mechanical) to better use than upholding the prejudices of a bunch of aristocrats from the eighteenth century?

Some technolinguistic tools have been attempting to do just that, albeit with mixed results. Wikipedia, whose slogan is "the free encyclopedia that anyone can edit," has been very effective at combating obvious vandalism with dedicated volunteer editors, but faces more subtle problems of bias in what it covers, because the volunteer editors it attracts are disproportionately male, well-off, and English speaking, and they tend to edit topics they're already interested in. Google Docs, where this book was written, has a spellcheck that draws on internet data, sometimes with surprising results. Once, to my great joy, it proffered a more common spelling of "Ronbledore" (an obscure Harry Potter fan theory that Ron Weasley is actually a time-traveling Dumbledore). Other times, it has persisted in suggesting the closed spelling "alot" over the open spelling "a lot"—a version that's common but more informal than I'd expect a spellcheck system to endorse. Perhaps the most promising computational tool for fighting bias rather than reinforcing it is Textio. This is a

startup that assesses the text of your job posting for whether certain words and phrases are likely to put people off applying, and thereby make the position take longer to fill, by sounding sexist or corporate jargony, flagging buzzwords like "big data" and "rockstars" in favor of "caregiver leave" and "learn new things."

Just like we can use language to be elitist, we can also use language to show solidarity, like politicians who suddenly adopt a folksy way of talking on the campaign trail. In some cases, shifting language is practically universal: none of us talk to a dog the same way we talk to our coworkers ("Who's a good boss! Do you want to go for walkies and also give me a raise?"). In other cases, our linguistic styles are bound up in a specific identity: William Labov studied residents of Martha's Vineyard and found that those who identified strongly with traditional island culture had stronger local accents than those who didn't. More recent research has shown that intonation in particular is related to social identity: young men in Washington, D.C., with one black and one white parent talk differently depending on whether they identify as black or biracial; the speech patterns of people living in Appalachia depend on how "rooted" they feel in the local community; and the speech of Jewish women in Ohio and New Jersey varies depending on their relationship with their Jewish identity.

In still other cases, the alignment is less about showing that you're part of the same group and more about borrowing coolness from another group. Research on youth language in several countries shows a parallel trend: there are distinctive linguistic forms associated with economically and racially marginalized youth in contexts ranging from the American inner city to the banlieues of Paris to the favelas of Rio de Janeiro. Elements of their language then get picked up by white middle-class youth. They don't adopt enough to make them no longer seem comfortably middle class, but just enough to strike a note of autonomy from parents, teachers, and other authority figures. Of course, when a word like "lit" or "bae" gets sufficiently associated with mainstream culture—and especially when it gets picked up by brands capitalizing on trends—it then loses its appeal to hip insiders, prompting the cycle to begin again.

In English, the association of words from African American English with coolness and their subsequent appropriation by non–African Americans is much older than the internet. Terms associated with African American music, including blues, jazz, rock and roll, and rap, have all made their way into broader Western culture, while the speakers who originated them continue to be stigmatized for the way they talk. One thing that changes with the decentralization of online media is that the original speakers can become more visible. While a white person in the sixties listening to Elvis might have had no idea that he was singing a style heavily influenced by black performers like B.B. King and Sister Rosetta Tharpe, it's easier to see that mainstream America's adoption of "on fleek" came from a post on Vine (a now defunct service for sharing short videos) by the user Peaches Monroee. Still, it's tempting to mislabel the many words currently being appropriated into general American pop culture from African American English as "social media words" simply because they're used by young people, and young people are on social media, without giving due credit to the words' true origins. Fittingly, the internet has come up with a word for this: columbusing, or white people claiming to discover something that was already well established in another community, by analogy with how Columbus gets credit for discovering America despite the millions of people who already lived there.

In other languages, English itself is often a source of trendy new linguistic influence, one that signals interest in a broader, global culture rather than a smaller local one. The situation in Arabic is particularly interesting, because it involves multiple languages, multiple dialects, and multiple scripts. Most Arabic speakers know two varieties

of Arabic: Modern Standard Arabic, which is the standardized, multinational version based on Classical Arabic that people learn to write in school but speak only rarely, and a local dialect, such as Egyptian or Moroccan Arabic, which is the language of everyday speech and doesn't have an official written form. Back when Arabic speakers, like most of the rest of the world, associated writing with formality and speech with informality, this worked fine. Sure, you'd have news anchors speaking the standard and advertisements written in the vernacular to add a bit of local color, but for the most part, Arabic was comfortably settled in what linguists call a diglossia: when a society has two languages or dialects that almost everyone speaks, each of which serves a distinct social function.

Then personal computers and the internet arrived, and things got really complicated, really quickly. Early computers and websites were in English and were often used by people at universities who spoke English to communicate with the rest of the world. And, importantly, these new devices generally came with English keyboards and English displays, rather than Arabic ones. So speakers figured out a way of writing Arabic sounds using the Latin alphabet, a system known by various names, such as ASCII Arabic, the Arabic chat alphabet, Franco-Arabic, Araby, Arabizi, and Arabish.

Arabizi has some distinct advantages. Most official Romanizations of Arabic use "kh" to represent the Arabic letter خ, a sound that may be familiar to English speakers as the "ch" in Scottish "loch" or the "x" in the Spanish pronunciation of "Mexico."* But "kh" is actually a rather confusing way of representing this sound, because it

*Note that North American English speakers often pronounce Scottish "loch" or German "Bach" further back in the throat than warranted, while pronouncing Spanish "Mexico" as if it's /h/. As pronounced by native speakers, all three are the same sound, which the International Phonetic Alphabet represents as /x/.

looks the same as simply the /k/ sound followed by the /h/ sound, a sequence which is rare in English (found only in compounds like "cookhouse") but fairly common in Arabic. So informal writers use a different convention. Based on the similarity in shape, people instead write it as the number 5 or 7' (that's 7 with an apostrophe), which looks sort of like the خ in a mirror. They don't use plain 7, though, because that's already in use to represent ح (its dotless equivalent), another sound that's hard to transcribe—many systems use "h" for it, because it sounds kind of like a throatier /h/, but that's a problem because Arabic also has the more common /h/ sound that's in English. Using 7 instead solves the problem of one letter representing two sounds.

By similar logic, the numbers 9' and 9 can be used for the letters ض and ص, the numbers 6' and 6 for ظ and ط, and the numbers 3' and 3 for the letters غ and ع—all representing sounds that don't have ready equivalents in the Latin alphabet. What's important about Arabizi is that it assumes familiarity with Arabic already: it's a grassroots system based on the priorities of literate native speakers that each of these different sounds should be represented by a distinct symbol. Other Romanizations tend to do the opposite, rendering the same letters as variants of "d" and "s," "dh" and "t," "gh" and a backwards apostrophe (or simply omitting it altogether, as in the word "Arabic" itself, which is technically "3arabi"), based on what they sound like to non–Arabic speakers. Sure, sometimes it's useful to be able to interact on a more globalized level, like when writing about names and locations in Arabic-speaking countries for an English-language newspaper, but sometimes you also care about the local. To Arabic speakers, these distinctions are completely vital, and omitting them is like trying to convince English speakers to spell "sing" and "thing" the same way because French doesn't care about that weird English "th" sound.

Although Arabizi was initially made necessary because computers

didn't support the Arabic alphabet, it's now taken on a social dimension. A paper by David Palfreyman and Muhamed Al Khalil, analyzing chat conversations between students at an English-speaking university in the United Arab Emirates, gave an example of a cartoon that one student drew to represent other students in her class. One student was labeled with the name "Sheikha," using the official Romanization of the university. But the nickname version of the same name, which doesn't have an officially sanctioned spelling, was written in the cartoon as "shwee5"—using Arabizi "5" to represent the same sound as the official "kh." It's a hand-drawn cartoon: there's no technological reason for either name to be written in the Latin alphabet. But at least for some people, it's become cool: participants in the study commented that "we feel that only ppl of our age could understand such symbols" and that it makes "the word sound more like 'Arabic' pronunciation rather than English. For example, we would type the name ('7awla) instead of (Khawla). It sounds more Arabic this way."

In particular, with advancements in keyboarding meaning that it's easier to type the Arabic alphabet than it was in the 1990s, people generally use the official alphabet for the Standard variety, with its established writing system, and can now use either alphabet in a grassroots fashion for the local varieties. A study of the linguistic choices of prominent Egyptians on Twitter gives us some examples of how people decide which one to use. A politician tweeted predominantly in Modern Standard, reflecting his older age and the traditional expectation of politicians to speak the standard. A popular singer tweeted mostly in Egyptian Colloquial Arabic with some Modern Standard, both written in Arabic script, reflecting his younger age and fanbase, as well as the language his songs were in. A fancy restaurant tweeted in English and Egyptian Arabic written in Arabizi, to appeal to a wealthy, cosmopolitan clientele who would have been

educated abroad. A cultural center tweeted in English and Modern Standard, to appeal to an educated regional and international audience. Egyptian Twitter users could thus potentially see four different linguistic conventions on their one feed: English and Modern Standard Arabic in their respective scripts, and Egyptian Arabic in both. And they could pick and choose between them for their own messages, depending on who they are and who they're trying to talk to.

While we may not all have multiple alphabets to choose from, we do all make linguistic choices based on our audience. Jacob Eisenstein, the linguist who was Twitter-mapping "yinz" and "hella," and his collaborator Umashanthi Pavalanathan at Georgia Tech decided to split up English tweets in a different way. Rather than look at location, language, or script, they looked at the difference between tweets about a particular topic, say the Oscars, versus tweets in conversation with another person. As it happens, Twitter has an easy way of automatically grouping these two kinds of tweets. If you put a hashtag in your tweet, like #oscars, then other people who are also interested in the Oscars know that they can click on or search that hashtag to find other tweets that also contain #oscars. If you put someone's Twitter username after an @ sign, like @Beyonce, then that user will get a notification about your message and hopefully reply to you the same way.

Since # and @ are distinct symbols, it's easy enough to automatically sort a giant pile of tweets, discarding the ones that contain both or neither. Sure, it's a bit rough—people probably aren't searching through sarcastic hashtags like #sorrynotsorry for topical information, and Beyoncé probably won't tweet you back (uh, #sorry)—but it works pretty well at a large scale. What Eisenstein and Pavalanathan found was that people used regionalisms like "hella," slang like "nah" and "cuz," emoticons like :), and other informal language more in the tweets that @mentioned another user, while the same people

used a more standardized, formal style in their tweets with hashtags. They theorized that, just as in person we'd generally talk more formally when addressing a roomful of people than when talking one-on-one, we're directing a tweet with a hashtag towards a large group of people. Our @mentions, on the other hand, are more informal, only noticed by a select few—and we adjust our language electronically the same way we do out loud.

Studies of people who tweet in other languages show a similar pattern. A study of people in the Netherlands who tweet in both the locally dominant language, Dutch, and a local minority language, Frisian or Limburgish, found that tweets with hashtags were more likely to be written in Dutch, so as to reach a broader audience, but that users would often switch to a minority language when they were replying to someone else's tweet. The inverse was less common: few people would start in a smaller language for the hashtagged tweet and switch to the larger language for the one-on-one reply.

Another study investigated how people use informal language in Indonesian, comparing how they write in private, one-on-one text messages versus public tweets. For example, the Indonesian word *sip* means "okay, yeah, good," but to emphasize it, you can respell it *siiippp*, and "thank you" is *terima kasih*, but if you want to try to match the pronunciation of the popular Jakarta dialect, you can respell it *makasi*. If @replies on Twitter are slightly more casual than messages broadcast in hashtags, then texts are more intimate still, and sure enough, Indonesians used informal respellings like this almost four times more often in texts than in tweets. Tweets were also nearly twice as long as texts, on average, and contained more complex sentences and a larger variety of words.

From an internet linguistics perspective, language variation online is important not so much because it's new (language has always varied), but because it's only rarely been written down. Literature

favors a few elite languages and dialects, even though there are around seven thousand languages in the world and at least half of the world's population speaks more than one language. So this glorious variety masks a digital divide: people who switch between languages or who speak a less written linguistic variety run into difficulties with many of the automated linguistic tools that internet residents rely on, such as search, voice recognition, automatic language detection, and machine translation. These tools are trained on large corpora, often from formal sources like books, newspapers, and radio, which are biased towards the forms of language that are already well documented. One method of bridging this gap uses public social media writing itself as training input—a promising avenue, considering that the quantity of informal writing produced on the internet exceeds the volume of formal writing many times over.

There aren't very many quadrilingual Arabic-Frisian-Indonesian-English speakers: I wouldn't expect to see a study of tweets switching between all four anytime soon. But regardless of the specific linguistic circles we hang out with online, we're all speakers of internet language because the shape of our language is influenced by the internet as a cultural context. Every language online is becoming decentralized, getting more of its informal register written down. Every speaker is learning how to write exquisite layers of social nuance that we once reserved for speech, whether we mark them by switching alphabets, switching languages, or respelling words.

All our texting and tweeting is making us better at expressing ourselves in writing. Researcher Ivan Smirnov analyzed posts by nearly a million users in St. Petersburg on the Russian equivalent of Facebook, a social media site called VK, from 2008 to 2016. He found that average word length, a measure of complexity, increases as people get older and as they get more education, as we might expect. But Smirnov also found that messages overall have been

getting more complex over time. As he put it: "15-year-old users in 2016 wrote more complex posts than users of any age in 2008."

No one who writes "u" does it because they're unaware that "you" is an option. A literacy study by Michelle Drouin and Claire Davis points out that the idea that textisms might interfere with our ability to produce the formal standard just doesn't fit with what we know about how memory works. Slang and abbreviations are for very common words: "u" for "you," "ur" for "your," "idk" or "dunno" for "I don't know," and so on. That's the point—the sender saves a bit of effort, and the receiver can interpret them because they're so frequent. We don't get internet abbreviations for longer, rarer words and phrases, like "pterodactyl" or "do you wanna start a band?" In psychological terms, shortcuts are for ideas that we've overlearned. You might forget how to find a fancy restaurant that you only go to occasionally, but you can get from your bed to your bathroom even when you're half asleep. If we were going to forget any part of language, it would be the rare, two-dollar words like "grandiloquent" or "sedulous" that we memorize with flashcards for the sake of a test, not the short words we learned as tiny children and keep encountering every day in both their abbreviated and non-abbreviated forms.

Just as conversation and public speaking have coexisted throughout human history, informal writing online can share space with more formal styles. Formal internet genres like ebooks and news sites and company websites no more resemble your quickly dashed-off text message than print books and newspapers and company brochures resembled a hastily scribbled note on the kitchen table. Several studies show that people who use a lot of internet abbreviations perform, at worst, just as well on spelling tests, formal essays, and other measures of literacy as people who never use abbreviations—and sometimes even better.

Instead, what people are doing with internet slang is a good deal

more subtle. The linguists Sali Tagliamonte and Derek Denis got seventy-one teenagers to donate the written records of their instant messaging conversations so that they could disentangle what they were actually doing. They found that the teens weren't actually using internet slang all that much. Unlike examples from hyperbolic articles, where almost every word is replaced with slang (r u gna b on teh interwebz l8r?), only 2.4 percent of the actual teens' messages were slang. (I'm reminded of the surveys of perception versus reality for other kinds of youth behavior, where everyone thinks everyone else is drinking more and having more sex than them.) What the teens were doing instead was more sophisticated: they intermixed the very informal features, like smiley faces and acronyms, with very formal ones, words like "must" and "shall" that are rare in speech. Here are a few snippets from various conversations:

> aaaaaaaaagh the show tonight shall rock some serious jam
>
> Jeff says "lyk omgod omgod omgodzzzzzZZZzzzzz!!!11one"
>
> heheh okieee! must finish it now ill ttyl
>
> lol. . as u can tell im very bitter right now.

The most obvious thing in these sentences, from the perspective of formal written English, is the informal parts: expressive lengthening like "aaaaaaaaagh," expressive punctuation like "!!!11one," and abbreviations like "ttyl" and "lol." But Tagliamonte and Denis point out that these sentences are also odd from the perspective of informal spoken English: if you record teens sitting around talking to each other out loud, at any point in the early twenty-first century, they barely ever speak words like "shall," "says," "must," or "very"—

they prefer the newer versions "going to," "is like," "have to," and "so." (Picture the difference between saying, "And then he said, 'Shall you go?' And I said, 'I must, I'm very tired,' " versus "And then he's like, 'Are you gonna go?' And I'm like, 'I have to, I'm so tired.' " The first belongs in writing, or in the speech of a previous generation, but the second is very much of our own.)

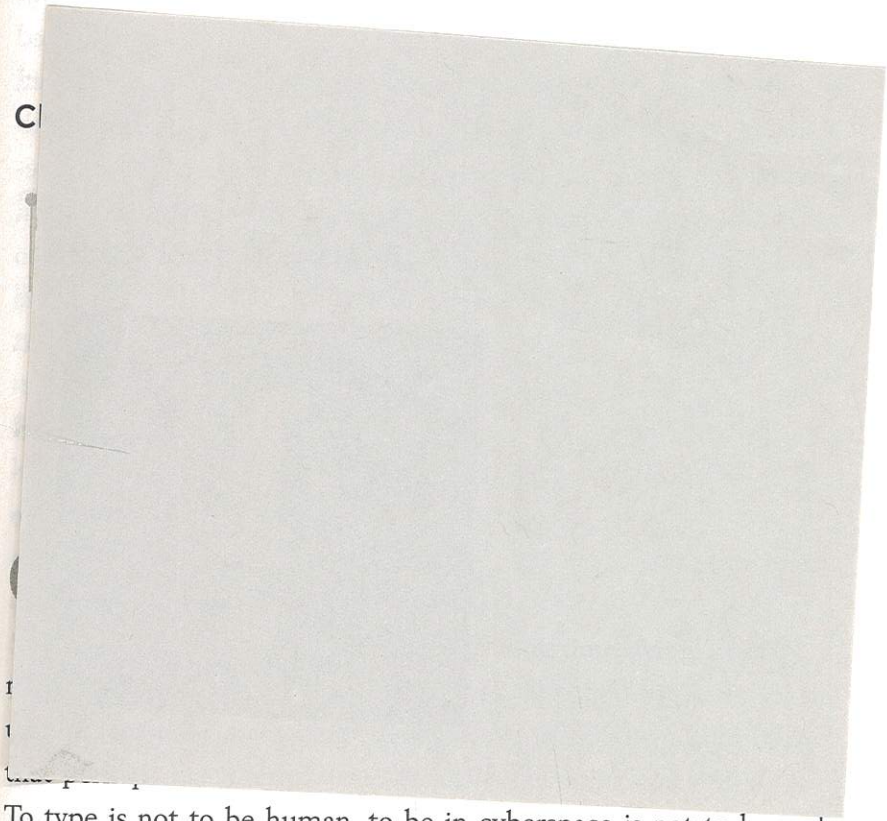
The fact that all but one of the new, informal versions is longer than the older words (two syllables instead of one) puts an immediate question mark by any assumption that the new forms could be a sign of laziness. But further, the fact that teens deploy this mix of formal and informal styles in writing suggests that what they're doing is neither an imperfect transcript of casual speech nor a failed attempt at formal writing. Internet writing is a distinct genre with its own goals, and to accomplish those goals successfully requires subtly tuned awareness of the full spectrum of the language. Media representations of chatspeak ring hollow when they borrow the exotic trappings (like "lol" and "ttyl") without acknowledging the linguistic expertise that it takes to navigate the system as a whole (the coexistence of "lol" and "heheh," or "shall" and "i'll").

Respellings and other internet styles can indicate not just informality but hospitality. Internet humorist @jonnysun tweets in a particular linguistic style, involving lowercase and creative respelling that you can see in his username, "jomny sun," and self-description, an "aliebn confuesed abot humamn lamgauge." The stylized language in jomny's tweets makes him feel approachable, unintimidating, and down-to-earth (apart from the small matter of being an aliebn). Despite his hundreds of thousands of followers, despite his day job as a grad student, you get the sense that he's the type of person who won't judge you for making a typo of your own. Some followers even tweet back in aliebn-speak: a spirit of friendly linguistic play that's more like a familect than a stuffy Oxford Common Room.

I've taken up this sense of linguistic play as a writing exercise, especially when I've just read a bunch of academic papers and I'm having trouble shaking my thoughts free of Nominalization Accumulation Enunciation Contamination. Instead, I draft in Peak Internet Style, with no capitalization or punctuation, using acronyms and creative respelling to write my way through the muddle, rather than stopping when I don't know how to articulate something or trying to sort through form and content at the same time. It's a lot harder to sound stuffy or pretentious when I've only got the tiny box of a chat window to type in and I can't go back and edit—and it's less painful to delete the necessary words when I haven't fussed with them as much. Eventually, I do figure out what I'm trying to say, and at that point it's straightforward to go back and add capitals and periods and delete things like "ugh idk what i'm doing hereeee." But it's easier to formalize the cosmetic elements while retaining an underlying clarity than it is to inject lucidity into a first draft that's classically formatted but dense and impenetrable. A paper analyzing the effects of spellcheck on writer's block suggests that I may be onto something: instantly appearing red squiggles may seem helpful, but for complex documents, they pull writers away from the overall flow and make them think about small details too early. I'm also not alone in noticing positive effects from social media on my writing style: Twitter users in particular often note that the character limits and instant, utterance-level feedback of the tweet format have forced them to learn how to structure their thoughts into concise, pithy statements.

Since long before Edmond Edmont hopped on a bicycle, people have been piecing together how various aspects of the human experience are reflected in how we communicate: our geography, our networks, our societies. There's always more to be figured out, of course, but we have a pretty solid understanding of the basics

of how we use language to show our identity when we're having a conversation. And there's a tantalizing inkling that we can express our true selves through language online as well: age-old linguistic practices like language play and switching between languages and styles are becoming written down and electronic. But the youthful, the vernacular, and the digital sides of language are still too easily overlooked: let's find out what we can learn when we take them seriously.



To type is not to be human, to be in cyberspace is not to be real. Rather all is pretense and alienation, a poor substitute for the real thing. Ipso facto, cyberspace cannot be a source of meaningful friendships."

And yet, as the discussion raged on, we've ended up conducting a sizeable portion of our social lives online. Close friends send funny links back and forth, grandparents and grandchildren videochat, partners text constantly about day-to-day activities, family members and old friends post photos that we like or comment on, and people join internet communities around a particular interest and end up becoming invested in each other's lives as well.