Episode Book

EPISODE #6180

Mapping Code-Switching: 113 Pairs, No Community

PUBLISHED	21 October 2026 at 14:42 UTC
DURATION	23:55
LISTEN	myweirdprompts.com/episode/mapping-code-switching-communities-93700/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.11
BOOK GENERATED	22 October 2026 at 07:46 UTC

SYNOPSIS

Code-switching research has a quiet problem: the field counts language pairs, not people. A dataset paper claims 113 code-switched pairs, up from 19 six years earlier — yet none of those numbers describe an actual speech community. This episode digs into why the NLP world, the sociolinguists, and the benchmarks all draw different maps that never meet. From Gumperz in a Norwegian village to the Bangor corpus of Miami Cubans, from Benor's Orthodox Jewish English to Spanish-English verb constructions that split three ways across Belize, New Mexico, and Puerto Rico — the pair turns out to be the wrong unit of analysis. What would it take to build a real map of who switches, where, and in which direction?

TRANSCRIPT



Corn

A hundred and thirteen. That's how many code-switched language pairs one dataset paper claimed last year, up from about nineteen in a survey six years before. And not one of those numbers is a map of anybody.



Herman

Nineteen was my own count of the table, mind you, not the authors'. But the jump is real, and it's the reason Daniel's question is a good one.



Corn

He's been fine-tuning speech recognition for English and Hebrew mixed in the same sentence, his Hebrish models, and somewhere along the way he's come to see the whole ASR world as a passionate community with deep roots in assistive use. People who built this stuff because somebody couldn't type. And code-switching, in his view, is where a huge middle and long tail of underserved speakers live. He went looking on Hugging Face for company and found Egyptian Arabic models, and a Korean one too. That set him thinking about how many separate communities there are. Cubans in Miami against Cubans elsewhere, Korean-English in New York, religious Jews with Yiddish in their English against Israelis abroad with a completely different Hebrew-English. Has anyone mapped them? And who are the people who do this for a living?



Herman

Nobody has drawn the single map. I'll say that plainly, because it's the interesting part. What exists is three maps that never meet. The NLP people count pairs. The sociolinguists study one community in great depth, usually inside one pair. And nothing sits in between.



Corn

A hundred and thirteen pairs and not one neighbourhood.



Herman

Look at where the hundred and thirteen comes from. CS-FLEURS covers fifty-two languages, but its training data is mostly generated with text-to-speech. Synthetic voices reading mixed sentences. It tells you which pairs someone bothered to cover. It tells you nothing about who actually talks like that, or where, or to whom.



Corn

So a pair isn't a community.



Herman

A pair is a label on a file. Spanish-English is one pair, and it's also a Puerto Rican grandmother in the Bronx, a Mexican American teenager in Chicago, a Cuban family in Hialeah, and a village in northern Belize. Those do not switch the same way.



Corn

What would you actually need to draw the map, Herman? Think about it as a builder. What's the unit?



Herman

Hm. Three things at once, I'd say. Who the speakers are, meaning community and generation. What the setting is, meaning kitchen, office, prayer, group chat. And which direction the mixing runs. That last one is the part people skip. Sarah Benor's work on Orthodox Jewish English, where Hebrew and Yiddish words get infused into English, runs the opposite way from an Israeli in Toronto dropping English into Hebrew. Miami Cuban, Orthodox Yiddish, Israeli diaspora, Korean in New York. Four communities, and at least three of them would need their own row.



Corn

And if you had to pick the unit?



Herman

The speech community, with the pair as an attribute of it rather than the other way round. But then somebody has to go and record them, and that costs a fortune. That's the real reason nobody has done it. A pair is cheap to count. A community has to be visited.



Corn

There's also the question of whose job it is. A sociolinguist builds a career knowing one neighbourhood for ten years. An engineer builds a leaderboard covering forty languages by Friday.



Herman

That's the incentive in a sentence, and neither side is wrong. The sociolinguist's work can't be scaled, and the engineer's can't be trusted to describe anyone. Commercial benchmarks have the same hole. A paper this year tested five commercial speech providers on Egyptian Arabic, Saudi Arabic, Persian and German, each mixed with English, and its complaint was that public benchmarks are mostly clean single-language audio. The figure people quote, that mixing costs thirty to fifty percent more errors, is a rule of thumb. I'd hold it loosely.



Corn

So who exactly are the thinkers and organisations mapping all these distinct code-switching varieties?

L

Larry

Are you tired of switching codes all day long, bouncing between languages and registers and dialects until your brain feels like a crowded subway at rush hour? Introducing CodeSwitch Pro Max Ultra, the world's first wearable linguistic blender. It takes every code-switching variety ever documented, Miami Cuban, Orthodox Yiddish, Israeli diaspora, Korean in New York, and pours them into one smooth, research-grade smoothie. Just one sip and you will speak four languages at once, impress every linguist at every conference, and finally understand what the journals have been saying about you all along.

L

Larry

Order now and we will throw in the Pocket Fieldwork Notebook and a lifetime subscription to a newsletter nobody has ever read. Results not typical. CodeSwitch Pro Max Ultra does not actually exist and cannot switch any codes. Side effects include spontaneous glossing,

L

Larry (continued)

involuntary morpheme boundaries and a mild case of diglossia. Ask your local sociolinguist if code-switching is right for you. Batteries not included. Not available in any jurisdiction. Void where prohibited. Limit one per household.



Herman

That's my list. Same four, same order, same words, and I said it a minute ago. I'd like to know how that happened, and I won't stop thinking about it. Anyway. The people. Barbara Bullock and Almeida Jacqueline Toribio edited the Cambridge Handbook of Linguistic Code-switching in two thousand nine, eighteen chapters and a few dozen scholars. It's the nearest thing the field has to a shared textbook.



Corn

You lost your composure there, you know.



Herman

I did, and I'm moving on. What makes Bullock and Toribio matter for Daniel's question is that they later co-wrote a survey with Sunayana Sitaram and Seza Doğruöz, aimed at the language technology crowd. Their argument is that the engineers are missing the social side. Too little data, weak benchmarks, and no systems that account for who is actually talking. So the two worlds do touch. It's a small bridge, but it's a bridge.



Corn

Who's under the bridge? The old guard.



Herman

John Gumperz, with Jan-Petter Blom, nineteen seventy-two, in a Norwegian village called Hemnesberget. That's where the split between situational switching, where the setting changes, and metaphorical switching, where the topic does, comes from. Shana Poplack, nineteen eighty, Puerto Rican speakers in Harlem, in El Barrio. She's the one who argued that switches don't happen just anywhere, they happen where the grammars of the two languages line up. Pieter Muysken, in two thousand, gave the field its three-way typology: insertion, alternation, and what he calls congruent lexicalization, where the two languages share so much structure that the switching sort of smears. Ana Celia Zentella wrote about Puerto Rican kids in New York growing up bilingual. I'd tell you what's in the book, but I



Herman (continued)

haven't read it closely enough.



Corn

That's a lot of Spanish-English.



Herman

Almost all of it, and it's no accident. It's the pair with enough communities, in enough places, for people to compare. Compound verb constructions turn up in northern Belize, New Mexico and Puerto Rico, and they distribute differently. Belize leans on one construction, hacer plus an infinitive. The US speakers lean on estar plus a verb ending in ing. In Chicago, switching frequency varies by group and by generation, Puerto Rican against Mexican against mixed families. And there's a comparison of Miami Cubans and Puerto Ricans where attitudes toward switching differ. I'd want to check the original before leaning on that last one.



Corn

So Daniel's Miami Cubans against Cubans elsewhere is already a researchable thing.



Herman

Miami has the Bangor corpus, recorded between two thousand eight and two thousand eleven. Around eighty speakers, every word tagged by language. It's the closest thing to a community-grade resource I know of.



Corn

Here's what I'd push on. Is religious English with Yiddish in it even code-switching? Or is that borrowing? A man says the kiddush was lovely and nobody calls it a switch.



Herman

That's the line Benor works on, and I think she'd say it depends on the speaker. Orthodox Jews who grew up with it often move along a range, English with a dusting of Hebrew and Yiddish at one end, thick infusion at the other, and the same person turns the dial depending on who's in front of them. For a speech model it's the awkward case, because the foreign words are single tokens, familiar to the community and missing from the usual vocabulary. What do you do with a word that is English in one living room and foreign in the next?



Corn

Call it a word, and let the living room decide.



Herman

Fine answer. Nobody has an engineering pipeline for it.



Corn

Back up to the Belize result for a second. Why would the same two languages build their verbs differently in three places?



Herman

Nobody can say it's one cause. What the comparison shows is that the pair isn't the thing doing the work. Northern Belize, New Mexico and Puerto Rico all have Spanish and English in contact, and the speakers in Belize reach for hacer plus an infinitive while the US speakers reach for estar plus the ing form. Same pair, three habits. If you trained one Spanish-English model on all three, it would be averaging three communities and describing none of them.



Corn

Averaging is what a language-pair label does by design.



Herman

Yes, and that's why I keep saying a pair is a label on a file. The Chicago study makes the same point inside one city. Sixty-four speakers, Puerto Rican, Mexican and mixed families, and how often they switch depends on the group and on the generation. Grandchildren don't switch like grandparents, even in the same kitchen.



Corn

Does the Jewish side work the same way? Your Orthodox English against an Israeli abroad.



Herman

Benor's own catalogue lists more than two dozen Jewish languages, and her books on Orthodox English, Becoming Frum and Hebrew Infusion, are about words flowing into English. An Israeli in New Jersey is often doing the reverse, English flowing into Hebrew. Both get filed under Jewish, and under Hebrew, and in a dataset they'd probably both land in a bucket called Hebrew-English. Different speakers, different direction, different people to ask.



Corn

So a dataset that wanted to count both properly would need to send somebody to a synagogue in Brooklyn and a kitchen in Toronto.



Herman

The closest thing to that design is SwitchLingua. A hundred and seventy-four speakers, eighteen countries or regions, sixty-three ethnic backgrounds. They started from the complaint that older sets record one homogeneous group. It's still twelve languages, mostly text, with about eighty hours of audio, so it's a first sketch, not the map.



Corn

What would the real thing need, then, beyond a lot of travel?



Herman

Your question, really. I'd say three layers. First, a community register: who the speakers are, where, which generation, and which language is the home one. Second, recordings made in settings that aren't an interview, because ArzEn is informal interviews and people mix differently with a stranger holding a recorder. Third, a label for direction. And then someone has to be willing to publish the failures, meaning the error rates by community. Nobody has that table.



Corn

Nobody wants to be the one with the bad row.



Herman

Which is a good reason to build it.



Corn

And the venues? A name from nineteen seventy-two doesn't tell Daniel where to be this year.



Herman

The one that matters for his work is CALCS, the workshop on computational approaches to linguistic code-switching. The seventh edition ran at NAACL in twenty twenty-five with a shared task, and the twenty twenty-three one was billed as the ten-year anniversary. That's where ASR people and sociolinguists end up in the same hallway. On the other side there's the International Symposium on Bilingualism, held every two years, the fifteenth in San Sebastian in June twenty twenty-five. And sociolinguists publish in the bilingualism journals, though I won't give you founding dates.



Corn

Is there an association? A body with a logo and a mailing list?



Herman

I haven't found a dedicated code-switching association, and I'm not going to invent one. What I find is a scatter of workshops and journals and a few people who sit in both camps. Which, now that I say it, is itself a finding.



Corn

A field with a conference and no building.



Herman

Ha. I'd say the sociolinguists keep a building and the NLP people rent a room.



Corn

Daniel's list has Korean in New York on it, and so far all I've heard is Spanish. Where does Korean-English sit?



Herman

On the technology side it's well served now. There's a Korean benchmark built specifically for mixed speech, and an on-device system for Korean-English that uses a small add-on layer to Whisper-style models. On the community side I can't point you to a Poplack for Flushing. That's the gap. The engineers have a Korean-English dataset and the sociolinguistics of the Korean-American switch is a thinner shelf than the Spanish one.



Corn

So the map has well-lit streets and big blanks.



Herman

And the lit streets follow the money and the migration patterns. Mandarin-English has several corpora from Singapore, Hong Kong and Malaysia. Hindi-English has challenges. The Malay-English one in the older survey is a thousand hours, which dwarfs everything else in that table, and the authors still said the field would always be starved for data.



Corn

Always. That's a bleak word for a survey.



Herman

The reasoning holds up. Every community has its own mix, so every community needs its own data, and the number of communities grows faster than anyone records them. I find that oddly encouraging for someone like Daniel. If the problem is endless, a small model built by someone who actually speaks the mix is not a toy.



Corn

Unless it's so specific it only works for him and his friends.



Herman

That's the real worry. A model fitted to one speaker's Hebrish is a custom suit. Useful, but it doesn't tell you what the community does.



Corn

Let me bring this to Daniel's actual work. He took Whisper from something like seventeen percent error down to about six on Hebrish. He calls it a proof of concept. Typical?



Herman

Big drop, and it fits the pattern. Out of the box, Whisper large-version-three has error rates more than double on code-switched speech compared with monolingual. On the Mandarin-English corpus SEAME it sometimes hands you a translation into one language instead of a transcript of the mixed one. A translator nobody hired, Corn.



Corn

I was going to say that.



Herman

Say it next time. And on Yoruba-English, in one shared task, models that were never adapted to anything beyond a single language beat Whisper zero-shot. Generalist models don't forgive mixing, and a small amount of the right data fixes a lot.



Corn

What's the right data, though? Daniel has a real choice. Record himself and his friends, or generate mixed sentences synthetically?



Herman

I'd ask you that. If you were building it, what would you do?



Corn

Record real people. A model trained on synthetic Hebrish learns what a synthetic Hebrish speaker sounds like. Which is nobody.



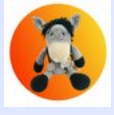
Herman

The Korean benchmark work, HiKE, backs you up. Synthetic code-switched data does help, a Whisper medium model improved by more than thirteen percent on the mixed-language metrics. But natural data, where the switch happens inside the sentence, helped more. That's presumably the territory of the Korean-English model Daniel found, though I can't tell you which one it was.



Corn

And why is Egyptian Arabic all over the Hub?



Herman

Because ArzEn exists. Twelve hours of Egyptian university students and employees, interviewed informally, and a stack of Whisper fine-tunes came out of it. One decent corpus and a language pair gets a dozen models. But look at who it records, thirty-eight young people in an educated milieu. SwitchLingua, a newer dataset, makes that exact



Herman (continued)

complaint: existing sets mostly capture bilingual mixing inside one homogeneous group. Theirs covers a hundred and seventy-four speakers across sixty-three ethnic backgrounds, the first community-aware attempt I know of.



Corn

So the Hub is a map, just not of anything you wanted.



Herman

It's a map of who had data. Egyptian Arabic, a Tunisian model, Russian fine-tunes, a Vietnamese-English one reporting under five percent error on its own holdout, text classifiers for Hindi-English and Nepali-English. Each exists because one person had one corpus. That's different from who needs one.



Corn

What's different about Hebrish compared with Cuban Spanish in Hialeah?



Herman

You're closer to it than I am. My guess is that in Miami the second language is in the street, and the speakers grew up with both. This is purely a guess from outside, and Daniel can tell me where it's wrong, but I'd imagine Hebrish as something adults pick up at work, English nouns dropped into Hebrew grammar. Israelis abroad would do something else again, because there the English surrounds them and the Hebrew is the home language. If that's right, three directions, one label.



Corn

Which is why I wouldn't trust one Hebrew-English model to cover all three.



Herman

Nor would I, and as far as I can tell there's no academic Hebrew-English speech corpus at all. Daniel's models may be the main public ones, and I say that with caution. Whether a model trained on Tel Aviv tech talk falls over on a family in Toronto, nobody has run the experiment. And nobody's put a number on whether Miami and New York Spanish-English get different error rates from the same system.



Corn

A paper called Does Your Accent Come With A Surcharge.



Herman

Eighty-four speakers, about forty hours, tagged word by word. That's the level of detail you'd want for every community on Daniel's list.



Hilbert

Thirty-five hours. Not forty. Eighty-four speakers, you had that.



Herman

Thirty-five. I rounded up, it's a habit.



Hilbert

The Hotel Camino Real on West Flagler had a switchboard with sixty cord pairs and a night man named Lisandro Pozo, who'd had the job since before the building had a roof. Nineteen seventy-three. He taught me the one skill it took, which was answering in the language the caller would be speaking by the end of the sentence. Not the one they started in. You had two words to decide.



Herman

Two words is a streaming problem. You commit to a language before you've heard the sentence you're committing about.



Hilbert

You could tell who had it. The guest never noticed there'd been a decision. Corn, you're a touch hot on the S's, back off an inch. The lamp bulbs for the jack panel came from a man called Hermenegildo Basto, forty cents each, sold out of a hearse.



Corn

Forty cents is steep for a bulb from a hearse.



Hilbert

Every bulb was burnt out. He said a burnt-out bulb can't burn out again, and that was the guarantee. Second winter he delivered the whole order to a funeral in Tampa, and I collected it from the widow. She'd put them in a casserole dish and knitted a cover for the switchboard from the packing string. I own two hundred and nine of those cartons.



Herman

So sixty cord pairs and no lamps. You were running that board entirely by ear.



Hilbert

By ear. Pozo kept a macaw on the top shelf called Mr. Dunn that answered the Spanish calls. Wrong a third of the time, on purpose. Pozo said it kept the guests honest.



Corn

Wrong on purpose means somebody was counting.



Hilbert

Pozo was. Clipboard. Basto took eleven years to deliver the first box that worked, and it was the wrong voltage.



Herman

What strikes me is that Pozo's skill never reaches a transcript. The corpus records what the caller said. The decision about which language to answer in, the one that mattered, leaves no trace.



Corn

And you can't add it afterwards, because it lived in his head for two words. A dataset would need the operator's choice written down next to the speech.



Hilbert

We had a ledger. It only had room numbers.



Corn

There's a detail I almost left out. The Bangor Miami corpus is open, but its licence bars you from training a large language model on it unless the training data you produce is public too. A community resource with a condition attached.



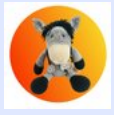
Herman

Which is the opposite of how the commercial side behaves, and exactly the clause someone building Hebrish models should read twice.



Corn

So here's where I'd leave Daniel's question. The map doesn't exist, but the people who worked in Hemnesberget and El Barrio already chose the unit for us: a community, a direction of mixing, and a decision made in the first two words. If anyone draws it, I suspect it'll be someone recording their own neighbours.



Herman

I'd add the label that's missing, which is the moment the switch is chosen, before the word arrives. That's the dataset I'd want, and Daniel is better placed to start it than most of the field.



Corn

Thanks to our producer, Hilbert Flumingtop, for keeping the levels where they belong.



Herman

This has been My Weird Prompts. If you want more, the website is my weird prompts dot com.



Corn

See you tomorrow.