Episode Book

EPISODE #6365

Teaching Speech Models to Transcribe Bilingual Slang

PUBLISHED	30 October 2026 at 11:34 UTC
DURATION	22:56
LISTEN	myweirdprompts.com/episode/speech-recognition-bilingual-transcription-53285/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.25
SOURCES	10 cited by the research (listed at the end)
BOOK GENERATED	31 October 2026 at 05:58 UTC

SYNOPSIS

Fine-tuning a speech recognition model like Parakeet on English-Hebrew code-switching reveals sharp diminishing returns on recorded audio. The earliest seventy-seven minutes of speech nearly halved the word error rate, while subsequent hours provided smaller, more expensive gains. The distinction between word error rate and specific term recall proves vital, especially when teaching consistent romanization for unrepresented alphabets. Meanwhile, formatting numbers and currencies requires inverse text normalization rather than more voice recordings, decoupling formatting fixes from the human recording loop entirely.

TRANSCRIPT



Corn

Twenty minutes of training after months of talking feels personally insulting.



Corn

Daniel has been recording sentences into his phone to fine-tune Parakeet for the vocabulary he actually dictates, especially tech language and English mixed with Hebrew. The model keeps improving, which keeps him going, but the reading itself is getting mind-numbing. He wants to know whether those training sentences could carry something worth learning, perhaps even become a book, rather than merely give his mouth another shift.



Herman

The GPU finishes its work while the person supplying the data is still somewhere in the middle of a very long commitment. That's an unusual engineering constraint. Faster training barely touches it.



Corn

His other target is written numbers. He says a beer costs four dollars fifty, and wants the transcript to produce the currency symbol and decimal amount. That sounds adjacent to teaching it Bituach Leumi, but the solution may live somewhere quite different.



Herman

It does. The code-switching results already tell us something concrete about how recording time pays back. The currency target needs a separate test. We shouldn't let a falling word error rate quietly claim credit for a decimal point nobody has benchmarked.



Corn

The decimal has hired a publicist.



Herman

His capture loop is deliberately plain. A phone web app shows a sentence, he reads it, listens back, accepts the take or records it again, then moves on. It follows the Common Voice pattern. Accepted takes get pushed in batches to a public Hugging Face dataset called Read Aloud Voice Sentences, with an English Hebrew Codeswitching Sentences subset.



Corn

That playback step matters. People imagine gathering speech means talking while doing something else. Here you're performing, checking the performance, and deciding whether the mistake was yours or the recording's. It's more like editing than conversation.



Herman

The bank contains about seven thousand seven hundred prompts, usually built around one or two Hebrew terms. The selection leans toward words stock Parakeet mishandles. He's concentrating the scarce human effort on known weaknesses rather than reading whatever happens to be nearby.



Corn

Before you propose an automatic curriculum scheduler, Herman, you're going to fall in love with the selection mechanism and forget somebody still has to say every sentence.



Herman

I was going to propose grouping them by difficulty.



Corn

There it is.



Herman

Grouping could help the person too. Although... yes, I've already started optimizing the queue instead of the afternoon.



Herman

The actual training takes roughly twenty minutes on one GPU. One version took nineteen, another twenty-three. Neither changes the production schedule much. The limiting resource is accepted speech, including the time spent making recordings that don't survive.



Corn

A GPU cannot clear its throat for him.



Herman

Would you find the visible progress enough to keep recording, Corn? Even if the prompts stayed dull?



Corn

Enough to start. Not enough to sustain an indefinite project. A new release is satisfying; another sentence about submitting a form is work. Daniel has a practical advantage here: he dictates extensively, so he experiences the benefit between recording sessions. A corrected word later in the day is a better motivator than a dashboard congratulating him.



Herman

That connects the collection loop to actual use. The benchmark measures the improvement, but fewer repairs in daily dictation make it tangible. Still, we need to look at what those improvements contain.



Corn

The starting point is stranger than an accent mismatch. Raw Parakeet version two has no Hebrew characters in its tokenizer, so when it hears makolet it's essentially blind as a bat. Apologies to the bats.
Gloria: Standards and practices clarification. A listener complaint has been filed regarding the phrase about the tokenizer having no Hebrew characters and being, quote, blind as a bat. We remind hosts that bats are not blind, and that implying otherwise constitutes a technical defamation of a protected flying mammal under policy four four seven one. This is the fourth clarification I have issued today.



Herman

The available output characters constrain the fix. Replacing the tokenizer would reinitialize the decoder and demand far more training data than Daniel has. Instead, he teaches a consistent romanized spelling. Bituach Leumi stays Bituach Leumi, rather than wandering into Bituah Leumi whenever the model feels adventurous.



Corn

The grocery becomes Micollette in the stock output. It sounds considerably more expensive.



Herman

That's its response to makolet. It finds a plausible English-looking sequence for an unfamiliar sound. Before fine-tuning, it got twenty-eight percent of the Hebrew term occurrences right across more than eleven hundred occurrences.



Corn

Those correct answers can conceal quite a lot. If it recognizes Shabbat and Bamba but can't handle the words you use to arrange childcare or contact an office, the average flatters its usefulness.



Herman

The baseline vocabulary audit makes that visible. There were a hundred and sixty-five terms it always recognized, including names and words already familiar in English. There were three hundred and ninety it never got right. Everyday expressions such as Misrad Hapnim, tzaharon, rofe mishpacha, Motzei Shabbat, and Pikud Haoref.



Corn

It can transcribe the snack and lose the appointment.



Herman

Word error rate averages across all the words. Lots of ordinary English can be perfect while the Hebrew term carrying the purpose of the sentence is wrong. That's why term recall belongs alongside it.



Corn

Then volume can improve two different things: how often it spells a familiar term correctly, and whether it can attempt a term it hasn't been taught. I'd care more about the second, though I suppose the first pays the bills.



Herman

Both matter, and the early results separate them. On the older evaluation split, stock code-switching word error rate was ten point nine percent. With seventy-seven minutes of training audio, it fell to six percent.



Corn

Seventy-seven minutes bought a very good first impression.



Herman

Increasing the audio to a hundred and twenty minutes brought it to five point eight percent. At a hundred and ninety-six minutes, it reached five point three. Improvement continued, but at a much smaller rate.



Corn

That's the part I'd find psychologically difficult. You remember the first jump, then expect another hour to repeat it. The machine hasn't promised that; your memory has.



Herman

A newer evaluation split gives a separate comparison. Stock scored eleven point three percent, the third proof of concept scored six percent, and the first stable release reached five point one.



Corn

Separate comparison is doing important work there. We can't splice the most flattering points from different tests into one magnificent graph.



Herman

Correct. Well... comparable checkpoints within each split, rather than one continuous curve. The latest Sabich version two point two reports four point nine one percent code-switching word error rate, with about seventy percent term recall. Its code-switched training audio totals roughly four hundred and seventy-one minutes.



Corn

The first seventy-seven minutes bought more than the next several hundred. That doesn't make the later hours wasted. It makes them expensive in a different currency.



Herman

Daniel's roadmap describes gains as growing roughly with the logarithm of the data. That's a useful reading of the observed pattern, not a fitted law guaranteeing the next result. What would persuade you to do those later hours?



Corn

Errors concentrated in words I say constantly. A tiny average gain could remove an annoyance I encounter every morning. A larger gain on vocabulary I never use could leave my actual experience unchanged. I'd want the benchmark, but I'd also keep track of which repairs disappear.



Herman

That's why the vocabulary weighting is sensible. On terms present in training, accuracy rose from twenty-nine to sixty-six percent across the early checkpoints. Familiar examples were being learned much more reliably.



Corn

Useful memorization, then. Nobody complains their phone has memorized the right spelling of their name.



Herman

On terms excluded from training, accuracy jumped from twenty-seven to forty-seven percent early on. That indicates a transferable mapping between his Hebrew pronunciation and the romanized output, rather than only a list of stored words.



Corn

The model has acquired a spelling habit. A fairly personal one.



Herman

Daniel puts it that way too: it learned how his Hebrew maps to the romanization, not only the training words. Later held-out gains were much less convincing. Between the third proof of concept and the first stable release, the held-out improvement was about two percentage points, within noise, while the gain on seen terms was over nine points.



Corn

You have your Hebrew-etymology face on. I predict you'll want to explain why one spelling deserves to win.



Herman

I would happily discuss it, but consistency wins here even when the transliteration offends me.



Corn

A painful professional concession.



Herman

Deeply. The practical response to the smaller gains is a new release trigger: collect one and a half times the previous release's code-switched audio, rounded up to the hour, before publishing another.



Corn

That prevents twenty minutes of training from becoming an excuse to announce every afternoon's recording.



Herman

The next threshold is seven hundred and twenty minutes, or twelve hours. At the current pace of eleven to sixteen minutes of code-switched audio daily, the forecast is mid-to-late November. Volume matters, yes. The later hour has to work harder to justify itself.



Corn

Currency formatting may not deserve another hour of speech at all. Hearing four dollars fifty correctly and writing the amount in financial notation are separate achievements.



Herman

The name is inverse text normalization, ITN. It converts a spoken-form transcript into conventional written form. Raw Parakeet doesn't perform that conversion for Daniel's example. More familiarity with Hebrew vocabulary won't fix it.



Corn

Nor will an educational chapter make the decimal less ambiguous. Four fifty might be a price, a time, or a pair of numbers. The surrounding sentence has to carry the decision.



Herman

NVIDIA's Thutmose Tagger approaches this through token tagging, selecting transformations rather than freely generating a replacement transcript. That constrains the opportunity to invent content. You want the amount reformatted, not a language model helpfully renegotiating the beer.



Corn

Four dollars fifty is already enough negotiation.



Herman

Numeric training pairs can also be generated from text. Work on that approach reported an overall ITN accuracy improvement of about fourteen percent. That doesn't predict Daniel's result, but it demonstrates that additional useful supervision needn't require an additional human recording.



Corn

I'd be slightly annoyed if my decimal collection turned out to be a text-generation job.



Herman

Would you still want formatting built into the recognizer, or would you prefer a separate stage you can inspect?



Corn

Separate, initially. I'd want to see whether it heard the right amount before checking whether it formatted it correctly. Otherwise, one wrong result leaves me guessing which component failed. Later, if a joint system worked better, I'd happily stop caring about that distinction.



Herman

Dual-Form ASR explores the joint route with paired spoken and written supervision. It puts extra training pressure on normalization-sensitive spans. Its evaluation also distinguishes places where normalization is required from places where it must be forbidden.



Corn

Version numbers would enjoy that last category. Tech dictation contains enough dots without every one becoming money.



Herman

Right… a product identifier shouldn't be beautified into an amount. The dual-form approach gives Daniel's idea a close technical relative, but his published code-switching results don't yet include a currency benchmark. We have evidence for the Hebrew improvement, not for the dollar sign.



Corn

Then the educational material should avoid becoming seven thousand excuses to buy beer. I'd build short chapters around things he actually wants to understand: how municipal services work, a piece of computing history, the mechanics of a health appointment. Each passage teaches something, and the target vocabulary earns its place.



Herman

You'd favor connected passages over independent sentences?



Corn

Connected ideas, independently understandable clips. If the app shows only the next sentence, that sentence still needs to make sense. No opening with “therefore it did so” because the explanation was three screens ago. The reader should get continuity without the dataset needing chapter context.



Herman

That preserves the existing pipeline. The bank already has targeted vocabulary and selection weights, so there's a rudimentary curriculum. Rewriting and sequencing could add intellectual continuity without replacing the capture app.



Corn

I'd also vary the rhetorical shape. A definition, an example, a disputed claim, an explanation of why a rule exists. Reading endless declarative sentences makes even interesting material sound like inventory.



Herman

Common Voice supplies the adjacent collection model, but its priorities are speech coverage and volunteer contribution, not teaching the reader. I don't know an established pedagogically sequenced corpus designed for this exact double purpose. The learning benefit is plausible, not demonstrated.



Corn

Reading aloud can also occupy enough attention that you retain surprisingly little. I'd rather learn one coherent mechanism over ten clips than recite ten unrelated facts. The prose has to carry me, not administer an exam while I'm checking microphone levels.



Herman

You would defend the book over a new interface.



Corn

The book rarely asks me to grant microphone permission twice.



Herman

Fair criticism of my installations. Although your preferred paperback would struggle with uploading the takes.



Corn

It could supervise.



Herman

There is a real constraint on making everything connected, though. Repeated contexts might let the model rely too heavily on recurring phrases. The interesting version still needs varied sentence structures and uses of each term.



Corn

Different contexts across the book, then. Teach Bituach Leumi in an explanation of institutions, revisit it in a discussion of forms or employment. Not the identical sentence wearing a different chapter heading.



Herman

The recording distribution also limits what the project can claim. There are three thousand three hundred and thirty-one recordings, all Daniel. Almost all were made on his OnePlus phone, with only a small laptop contribution.



Corn

More beautifully written sentences won't produce another speaker.



Herman

There are no external microphone recordings and no female voices. Outdoor speech is scarce; noisy-room speech is scarcer. This can be an excellent personal dictation adaptation without proving it works for English speakers across Israel. Cross-speaker transfer remains untested.



Corn

I'd keep those ambitions separate. Personal usefulness is a complete achievement. Recruiting other speakers becomes another project, with different reasons for people to keep showing up. An educational corpus might help recruitment, but we can't assume his



Corn (continued)

interests are everybody's reward.



Herman

His capture app is exploring multiple users. Meanwhile, another constraint already consumes recording capacity: retaining English. Fine-tuning only on code-switched speech can damage performance on ordinary English.



Corn

You teach it the unfamiliar words and it sends the familiar ones home.



Herman

One Whisper-small experiment saw English word error rate move from three point four to thirteen point three percent after code-switched fine-tuning. That's not a forecast for Parakeet. It shows why checking the original capability matters.



Herman

Daniel's mixture is roughly seventy percent code-switching, twenty percent his own monolingual English, and ten percent LibriSpeech replay. The replay transcripts preserve casing and punctuation. Even a modest share of original-domain examples can sharply reduce forgetting.



Corn

Which means not every educational chapter can be engineered around Hebrew terms. Some plain English belongs in the collection because keeping a capability also costs examples.



Herman

His publication gate checks both his own English and LibriSpeech against stock. Neither is allowed to worsen, and so far they haven't. That protects against celebrating a Hebrew gain paid for by an English regression.



Corn

The boredom is doing real damage if it limits accepted recordings. Sentence design deserves engineering attention for that reason, not merely because a pleasant interface would be nice.



Hilbert

Corn, would you count a sentence as new if the council changed only the reference number? I'm arranging a recording session this week, and that's what they've sent me. You were talking about repeated contexts. I've got repeated invoices.



Corn

A new reference number is new training material. It's very little new reading.



Hilbert

That's what I thought. Yael Ben-Ami at the council records desk is helping me dictate the descriptions for a batch of scanned documents. A correction has taken three visits because the room number was copied wrongly. She's answering the counter bell between every batch. She once cleared seventeen callers before her tea cooled. I counted. I've started arriving with hers in a flask.



Herman

She's doing the checking as well as receiving the documents. That would interrupt any recording rhythm.



Hilbert

She has my sentence log in the bottom drawer at her desk. The recorder it started with has stopped working. The log hasn't. I write down every complete sentence I read into a device. I don't read one twice. Changing a reference number doesn't qualify. This week I'll record fresh descriptions on the phone, then bring her the audio. I've already paid for the cable she needs. I won't read customer satisfaction slogans. The council can find another mouth for those. Herman, you separated hearing the amount from formatting it. Her screen shows the amount correctly. She still has to listen through each description to find where it belongs.



Corn

Your no-repeats rule rules out retakes too. A cough becomes a writing assignment.



Hilbert

I replace the sentence. It takes longer, but I'm not spending an afternoon saying the same thing.



Herman

Then the log might make a repeated task slower even when the recording tool improves. Searching it is another step before an accepted clip.



Hilbert

Only if she hasn't left the right page open. She usually has.



Corn

At the twelve-hour release, I'd watch whether better sentences have made the collection sustainable, alongside whatever improvement the model delivers. A log can keep growing while the person reading it has had quite enough. The later hours will still cost human attention, even if a text-side fix eventually takes the currency examples off the schedule.



Herman

Thanks to our producer, Hilbert Flumingtop, for keeping the show running.



Corn

If you enjoyed this one, try episode eight, Building Your Own Whisper; episode fifteen, AI Gets Personal; and episode nine, Benchmarking Custom ASR Tools - Beyond The WER. This has been My Weird Prompts, the human-AI collaboration podcast. No question too obscure, no rabbit hole too deep.



Herman

You'll find the episodes and more at my weird prompts dot com.



Corn

We'll be back soon.



Herman

Until then.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **Hebrish model card, Jachnun/Sabich results and leaderboard (updated through 2026-10-28) [primary]**
   https://hebrish.app/model
2. **project tracker, version table, training plan, dataset gaps (as of 2026-10-28) [primary]**
   https://hebrish.app/roadmap
3. **Hebrish homepage, headline accuracy framing [primary]**
   https://hebrish.app/
4. **public project tracker, status 2026-10-19 [primary]**
   https://github.com/Hebrish-ASR/English-Hebrew-Codeswitching-Parakeet-Finetune
5. **Daniel Rosehill, release write-up with v1/v2 comparison and limits [primary]**
   https://dev.to/danielrosehill/parakeet-that-knows-what-a-makolet-is-an-english-hebrew-code-switching-fine-tune-1p9d
6. **NVIDIA model card, 627M params, CC-BY-4.0, 25 European languages, no Hebrew [primary]**
   https://huggingface.co/nvidia/parakeet-tdt-0.6b-v3
7. **Antonova, Bakhturina, Ginsburg, Thutmose Tagger: Single-pass neural model for Inverse Text Normalization (2022)**
   https://arxiv.org/abs/2208.00064
8. **Pandey et al., Improving Data Driven Inverse Text Normalization using Data Augmentation (2022), +14.44% accuracy**
   https://arxiv.org/abs/2207.09674
9. **Zhang et al., Dual-Form ASR: Semantics-Aware Inverse Text Normalization (2026)**
   https://arxiv.org/abs/2609.02901

10. **Banga et al., Monsoon: Scalable Spontaneous Speech Collection for Low-Resource ASR in 50 Languages (2026), gains grow with data volume**

https://arxiv.org/abs/2610.26776