Episode Book

EPISODE #6158

Naming the ASR Sub-Tasks Nobody Mapped

PUBLISHED	20 October 2026 at 13:46 UTC
DURATION	23:52
LISTEN	myweirdprompts.com/episode/asr-subtask-taxonomy-acronym-expansion-30945/
HOSTS	Corn, Herman, Larry, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.10
BOOK GENERATED	21 October 2026 at 07:50 UTC

SYNOPSIS

Daniel fine-tunes Whisper-style models and keeps finding hundreds of tiny jobs hiding inside the big "ASR" label — like teaching a model that "F Y I" should be written out as "for your information." It turns out Sproat and five co-authors named this exact decision back in 2001, in a taxonomy built for text-to-speech front ends. The catch: their arrow points the other way. This episode digs into what the old tags EXPN, LSEQ and ASWD actually cover, why the direction of the mapping matters, and what happens when two purpose-built datasets disagree on purpose. Plus: why ASR scoring deliberately normalises away most formatting, and why that means nobody ever built the map Daniel is looking for.

TRANSCRIPT



Corn

Somebody split this exact hair in two thousand and one, which is either comforting or humbling and I haven't decided which. Daniel fine-tunes Whisper-style models by blending small purpose-built datasets, and today he got an acronym set, where the speaker says "F Y I" and the transcript reads FYI. Then it hit him that he also wants the sibling, where the same audio comes out as "for your information", fitted to the context, for the fine-tunes where he prefers it written out. He called it acronym expansion and wondered whether a proper word for it exists. The bigger itch is that Hugging Face has the big task labels, ASR is one of them, but inside ASR he keeps finding hundreds of little jobs. Has anyone mapped them, and can he just coin names so a blend recipe can say "pull from the acronym-expansion set"?



Herman

The word exists, and it has been sitting in a table for twenty-five years. Sproat and five co-authors, a paper called "Normalization of non-standard words", Computer Speech and Language, two thousand and one. They were building the front end of text-to-speech systems, so they had to label every odd token in a text by how it should be read aloud. The alphabetic tags are EXPN for abbreviations you read as full words, LSEQ for letter sequences, and ASWD for letter strings you read as a word. CIA is a letter sequence. NATO is read as a word. "Gov't" is an expansion.



Corn

Three tags for three ways of saying the same four letters. Tidy.



Herman

It gets better, because they hit Daniel's exact problem. The expansion tag, they say, is not used for expanding acronyms to their full name, "unless it would be more natural to say the full expansion in that genre." Their examples are IBM, which gets labelled a letter sequence, against NY, which gets labelled an expansion. New York.



Corn

So in two thousand and one a committee had already written "fitting the context" into the rules. Daniel's two datasets are in there, with a footnote saying it depends.



Herman

The footnote is the good part. One caveat though, Corn. Their direction is text to speech. They see IBM on the page and decide whether the voice says "I B M" or "International Business Machines". Daniel's case is the mirror image. He hears "eff why eye" and decides whether the page says FYI or the full phrase. Same decision, other side of the glass.



Corn

So it's not a new species. It's an old one with the arrow flipped.



Herman

That's my read, and part of it is inference, because I haven't come across anyone publishing an ASR set under that description. The taxonomy carries over. The direction doesn't.



Corn

You're enjoying this table a little too much.



Herman

I'm enjoying it exactly the right amount. It's a table with CDs in it.



Corn

Go on, then. What's the CDs problem?



Herman

The authors admit it's borderline. They say CDs could be labelled a letter sequence or a split token, because of the lowercase s on the end. And they say the whole taxonomy was designed partly around what a system could detect automatically. So a coined category is always partly a tooling decision. That was true in two thousand and one and it'll be true for Daniel.



Corn

Which is reassuring, since half my filing system is whatever the drawer would hold.



Herman

Another one I like. They added the punctuation tag late, after labelling experience suggested it would be reliable and useful. The taxonomy grew while they labelled.



Corn

Punctuation, the guest who gets invited after the seating plan.



Herman

Here's what I want your take on, Corn. Daniel's two sets disagree on purpose. Same audio, F Y I, one target says FYI and the other says "for your information". If he were building a transcript for you, which would you want?



Corn

Mine would want FYI, and I'd want it quick, because I read slowly. But I'd also bet that if the transcript went to a legal team, they'd want the long one. It depends where the words are going, which is the genre line again.



Herman

It is, and it's why neither set is wrong. They're output-style choices. But they're dangerous in one blend, because a conflict in training data is just the same input mapped to different targets. Put both sets into one fine-tune with nothing telling the model which style you want, and it learns a coin flip. Some utterances come out short, some long, and you can't predict which.



Corn

Could you tell it? Tag the style somewhere in the input?



Herman

I'd expect a style cue, a tag or a prompt or something, would let one model hold both. Or you keep two fine-tunes. I'm inferring that. I haven't seen it done for this particular distinction. What I'm sure of is that the recipe has to know the two sets are mutually exclusive unless there's a switch.

L

Larry

Verbatim. Breathe in. The fragrance of a thing said aloud, before anyone has tidied it. Letters, falling softly through a velvet afternoon. F. Y. I. A whisper that has not yet decided what it wants to become. Notes of unfinished sentence, of a room just left, of a question that was almost a sound. Linger. Let it settle on the collar, on the wrist, on the lower shelf of the supermarket, because Verbatim is a mop head. Fibres of pure intention. Sixty centimetres of quiet. Verbatim, for the one who is never quite written down. Wring gently. Available in the family pack.



Corn

That had your fingerprints on it, Herman. Breathy, vague, and it ends in a mop.



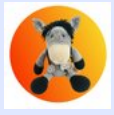
Herman

I was at the archery range from nine until twenty past one, the range keeps a sign-in sheet, and my entry is in green pen because the black one had run dry, which the man at the desk will confirm, his name is Yossi, he wears a blue cap.



Corn

Noted. Back to the arrow flipping the other way.



Herman

Inverse text normalisation, then. Text normalisation is written to spoken, so "one two three" gets read as "one hundred and twenty-three". Inverse text normalisation, ITN, is the ASR post-processing step going back, spoken to written. NVIDIA's NeMo toolkit defines it exactly that way, and its ITN is a set of hand-written grammars, one family of rules per class of thing. The class names follow Sproat and Jaitly, whose dataset gives you labels like LETTERS for a letter sequence and VERBATIM for reading a character string as it stands.



Corn

So Daniel's FYI set is letters turning into the written abbreviation. ITN with a small hat.



Herman

In classic terms, yes. The expansion set is the odd one out. It goes spoken to written, but it deliberately writes out the verbal form. My own reasoning here, nobody has published it as far as I know: it's an anti-ITN choice for one class. You switch off the written-form default for acronyms only.



Corn

Is there already a mechanism for that, or does he invent one?



Herman

NeMo has a whitelist, a file of custom mappings, and when the grammar finds an ambiguity the whitelist has the highest priority. A line saying FYI maps to "for your information" is the nearest thing. It's a post-processing tool, not a training set. But the idea is the same.



Corn

A priority list for the pushiest entry in the file. Sounds like your side of the family.



Herman

My side of the family is a donkey from Connecticut. Yours is the one that was supposedly born in Mongolia.



Corn

Supposedly. Fine. You win that one.



Herman

And the "fitting the context" part is its own field. Whether "IT" is the department or the pronoun, whether "AM" is a time or a radio band. That's acronym disambiguation. There were shared tasks on identifying acronyms and on disambiguating them at an AAAI workshop, around fifty teams on the first and low forties on the second. And a clinical one, a sense inventory of abbreviations in medical notes, from the University of Minnesota: four hundred and forty abbreviations carrying nine hundred and forty-nine senses between them.



Corn

In a doctor's note, a two-letter abbreviation can mean more than one condition, and nobody speaking it aloud says which. Let me guess. All text, no audio.



Herman

All text. Nobody has picked up the audio side under those names that I know of.



Corn

So the prior art covers writing, speaking and choosing the meaning, and nobody has glued the three into a transcript. Daniel's niche has a population of one.



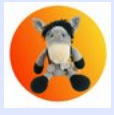
Herman

Probably a few, but not many.



Corn

So what does he call it? Not the config, the idea. A colleague asks what's in that set.



Herman

Three candidates and none is perfect. "Acronym expansion" is what the text-only literature calls resolving an abbreviation to its long form, so the words are taken but people would understand them. "Verbalised output" or "spoken-form output" comes from the text normalisation side, accurate about the target but it describes the mirror image, so it



Herman (continued)

might confuse. And the Sproat tag, EXPN, which is compact and unfriendly.



Corn

I'd keep his own phrase for conversation. Everyone knows what it means in four seconds. The formal name can live in the config.



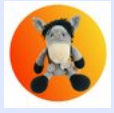
Herman

Same here, with EXPN somewhere in the formal name so the literature finds it. Which leaves his bigger question. Why is there no map of ASR sub-tasks?



Corn

Because the field decided formatting was lint.



Herman

Close. Look at how ASR gets scored. The Whisper paper has an appendix with an English normaliser that runs before any errors are counted. It strips anything in brackets or parentheses, deletes fillers like um and uh and hmm, expands contractions, removes commas inside numbers, strips symbols and diacritics, turns spelled-out numbers into digits, swaps British spellings for American, and collapses whitespace. Lowercasing happens in the code. The stated aim is to penalise real mistranscription and not formatting or punctuation.



Corn

So the exam deliberately ignores most of what Daniel is trying to teach the model.



Herman

Most of it. FYI versus the long form is not a spelling difference the normaliser touches, so I'm not certain how that exact case is scored, but the principle is plain. And the effect is large. In one evaluation paper I've only seen the headline numbers of, stacking the normaliser steps took the error rate on the same output from about twenty-two percent to



Herman (continued)

about ten.



Corn

Half the errors were the scorer's taste.



Herman

By that measurement, yes. And the Whisper training side is blunt too. They filtered out transcripts that were all uppercase or all lowercase, to avoid what they call transcript-ese. They also note that many ASR systems include some inverse text normalisation, but simple and rule-based. A field that has chosen to grade around formatting is not going to grow a vocabulary for formatting jobs.



Corn

Is that wrong, though? If I'm comparing two models I'd want the scorer to stop arguing about commas.



Herman

For comparing models, fair. The cost is that formatting stops being a measured thing, so nobody labels it, so no vocabulary forms. When did you last see a leaderboard column for handles acronyms in written form?



Corn

Never. There's no column for the one thing I'd actually check, which is whether FYI shows up as FYI.



Herman

Not everyone ignores it, though. There is a defender.



Corn

Hm. Name one.



Herman

The ESB benchmark, from Gandhi, von Platen and Rush, runs one system across many datasets and says outright that punctuation and casing artefacts are hard and are included in the evaluation. So the idea that style is part of the task has a defender. It never turned into a list of named sub-tasks.



Corn

And Hugging Face itself? Daniel says the big labels are there.



Herman

The ASR task page lists one variant. Multilingual ASR. That's the whole sub-task tree. Dataset cards have a field for task categories from a fixed list, and a field for task IDs, whose examples are question answering and image classification, nothing for speech. I can't swear a speech ID is impossible. Then there are tags, which are free-form, and configs, where you give each subset a name and point it at its own files.



Corn

One variant. That's a short map.



Herman

So the answer to "can he just coin names" is yes, and it's the sanctioned route. A config name is exactly where this goes. Nobody is going to stop him.



Corn

And the tags? If configs carry the names, what are tags for?



Herman

Search. Tags are where the Sproat words go, the letter-sequence and expansion names and inverse text normalisation, so someone typing the literature's vocabulary lands on his set. Configs are for building, the named subsets with their own files. The card text is for the human who reads it in a year: the mapping, the genre rule, the conflict line. Different readers for each.



Corn

Then how does he avoid a config called final-final-three?



Herman

Borrow what exists, coin the rest. Let me sketch the axes, and push on them, Corn, because I haven't tested this on anybody. First axis, what is spoken. Sproat's tags already cover it: letters said one at a time, a letter string said as a word, an abbreviation said as its full words. Second axis, what's written, the output policy. The abbreviation, which is the FYI set. The expansion, which is the new one. And a third value, fully verbatim, if he ever wants the spoken form word for word. Third axis, the domain.



Corn

Does the domain need to be a separate axis? If the policy is "expand where natural for the genre", the genre is doing the deciding.



Herman

You've found the weak point. If the expansion set means "expand when natural", it needs a lexicon per domain. This acronym expands here, stays letters there. IBM stays, NY expands. Without a domain label the set contradicts itself internally, not just against its twin.



Corn

So the name has to say which house rules it follows. Okay, what would you call them?



Herman

Spoken class first, then output policy, then domain if it matters. Letter-sequence, written-form, general. Letter-sequence, expanded-form, general. If he specialises, a medical suffix. That's my coinage and I'd flag it as his in the dataset card, with a line mapping it to the Sproat tags so anyone searching with the literature in hand can find it.



Corn

The card as a Rosetta stone. And then the blend recipe reads like a sentence: take this much of the letter-sequence expanded set, this much of the written one for the other model.



Herman

Never both in the same fine-tune. The card should say so in one line, a conflicts field.



Corn

Here's my doubt. Two sets, each with its own audio, means Daniel records or generates everything twice. Is that sane?



Herman

Not necessarily, and this part is design advice, not literature. Keep one master table where each clip has the spoken form, the written abbreviation and the expansion as separate columns, plus the class tag and the genre. The two named sets are then views of it: a filter, plus a choice of which column becomes the target. "Pull from the acronym-expansion set" turns into a query, and the audio exists once.



Corn

Which also means when he finds a clip where the expansion is wrong for the genre, he fixes one row.



Herman

Once, and both sets stay honest. I'd keep a column for the genre the expansion was chosen for, because that's the thing a future Daniel won't remember.



Corn

Future me remembers nothing. I've made my peace.



Herman

Do you buy the fine-tune at all, though? Or is a lookup enough?



Corn

For fixed pairs, a lookup. FYI goes to the long form in a millisecond, done. The fine-tune earns its place when the choice depends on context, the department versus the pronoun, because a lookup can't tell. So the expansion set is worth building only where the genre decides. Everywhere else it's a spreadsheet pretending to be a model.



Herman

That's a sharper test than the one I had. Where the answer depends on context, build the dataset. Where it doesn't, write the mapping.



Corn

One more borrowed vocabulary, if you want it. Transcription houses already say clean verbatim, full verbatim, edited. Output style with a shared name.



Herman

The one place where ASR output style has a shared tier vocabulary, and it comes from the transcription industry, not machine learning. Sources disagree on exactly where clean ends and intelligent begins, so I'd borrow the words loosely. And there's a speech-side extension of the Sproat taxonomy, Van Esch and Sproat, which built many more classes from recognition and synthesis systems across languages. I've only read the abstract, so I won't claim its class list. Daniel should check it before coining the next name.



Corn

You'd borrow from everybody and credit nobody.



Herman

I'd credit all of them in the card. That's the point of the card.



Hilbert

Brass plate, nine centimetres, twenty-six shekels, from Avrum on Agrippas Street. The plate, not the letters. He charges the letters separately, forty agorot each, and that is what we are not agreeing on this month.



Herman

A per-character tariff. The expansion set has the same problem in miniature, since a long target costs more of everything.



Hilbert

I ordered forty labels for the patch bay. Three letters each, XLR. He delivered forty plates with a long phrase on them, nothing like what I wrote, and says the short form was a misunderstanding on my part and that he corrected it for my protection. For the kettle I wrote FYI. The plate says FOR YOUR INFORMATION THIS KETTLE IS HOT. Thirteen shekels twenty. FYI would have been one twenty.



Corn

Avrum runs the expansion policy with no domain switch. On a kettle it reads fine, though. He picked the right genre by accident.



Hilbert

Herman, you read the whole normaliser list out in one go earlier. Avrum does that with the invoice. Nineteen lines, and the last line is the cat.



Herman

The cat has a line item?



Hilbert

The cat has the account. Avrum put the shop in the cat's name in the spring. The cat signs with a paw, the bank takes it, and the cat has been in credit since the third week. I've been going for eleven weeks. Nine. Weeks when his machine is waiting on the shipping forecast don't count.



Corn

Waiting on the forecast is the part that needs explaining.



Hilbert

The engraver runs off a radio. When the station plays the shipping forecast it stops, so he cuts those plates by hand. Twelve agorot a letter extra, goes on the cat's side. The forecast is seven minutes long and he starts every long plate at the beginning of it.



Herman

So the long form costs more letters, and the letters come out slower. For a model that's the plain version of it. More tokens to produce for the same few syllables of audio, and it changes how hard that set pulls in training.



Hilbert

On Tuesday a man from the council bought the kettle plate off the counter. He has no kettle. He wanted the plate.



Corn

Then the plate is the kettle now. Nothing hot, one warning, everyone fully informed.



Hilbert

I did hospital signage for a while. Nobody there wanted the long form either. I go in Thursday. He says the kettle plate is final and the cat has been paid.



Herman

In a blend, that's the case for weighting the long set on purpose and leaving it out of the model that doesn't need it. The pull depends on how many characters it adds, not only on how many rows it has.



Hilbert

Letters. Not characters. The spaces are free, he says, it's a gift. Thirty-three letters, so thirteen twenty is right.



Corn

Daniel's pair is three letters against eighteen. Six times the length, and that's the number the recipe has to know about.



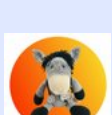
Herman

Corn, if you were setting the weights, by clips or by characters?



Corn

Clips. It's what I can count. Then I'd listen to the outputs, and if the model starts expanding things in contexts that should stay short, I turn the long set down. A dial you adjust by ear.



Herman

I'd go by characters for the first pass. That's from first principles, I haven't seen anyone report it either way. The ear check afterwards, though, I'd keep. That is the part that tells you the names mean something.



Corn

Then Daniel's sentence in the recipe works if the name carries three things: what was spoken, what was written, and which genre made the call. Sproat's tags give him the first, one coined word gives him the second, and the card gives him the third and the conflicts.



Herman

What I don't know is whether the length of the long set will ever make a measurable difference once the blend is tuned. Daniel will find out before anyone publishes it.



Corn

Thanks to our producer, Hilbert Flumingtop, for the production. This has been My Weird Prompts. If you want more, my weird prompts dot com has the back catalogue.



Herman

See you tomorrow.