Episode Book

EPISODE #6300

Choosing the Right Audio Annotation Tool

PUBLISHED	27 October 2026 at 12:45 UTC
DURATION	22:41
LISTEN	myweirdprompts.com/episode/audio-annotation-tool-selection-82225/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	openai/gpt-6.1-sol
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.22
SOURCES	20 cited by the research (listed at the end)
BOOK GENERATED	28 October 2026 at 09:53 UTC

SYNOPSIS

Turning raw environmental recordings into machine learning datasets requires choosing between lightweight inspection tools and heavyweight crowd platforms. From offline desktop staples like ELAN and Praat to self-hosted frameworks like Label Studio, Potato, and bioacoustic-focused Whombat, each tool handles timeline events differently. The discussion covers tier management for overlapping sounds, local prelabelling with faster-whisper, model-assisted workflows using ML backends, and when commercial platforms like Encord or managed services like iMerit make sense.

TRANSCRIPT



Corn

One of these audio annotation tools is a single HTML file. Another gives you a server administrator. Daniel wants to turn yesterday's street recordings into a dataset, and somewhere between those two products is a reasonable afternoon.



Herman

The HTML file is VIA three. It works offline in a browser. The administrator comes with Audino, which inherits much of its machinery from a video annotation platform.



Corn

Daniel knows his recordings contain car horns, the light rail sounding its horn, and background conversation. He wants actual environments for marking those sounds, rather than immediately building another interface himself. Desktop software, something he can host, or a service that takes his files and carries the work further.



Herman

Including the awkward part after importing them. Correcting transcripts, checking sample quality, getting usable labels out, and perhaps training a model without assembling the entire pipeline separately.



Corn

I suspect the horn is easier to locate than the appropriate subscription.



Herman

A surprising amount of the choice depends on whether Daniel is one person listening carefully or the manager of twenty people listening inconsistently. Those products can look similar in screenshots. Their daily friction is very different.



Corn

His immediate job sounds closer to searching than transcription. He has a long recording and wants a few specific events. A transcript might help him find speech, but a sentence saying traffic noise doesn't tell him where the light rail horn begins.



Herman

Bioacoustics has been tackling that shape of problem for years. Long field recordings, rare calls, overlapping sources, uncertain boundaries. The bird doesn't conveniently speak at the beginning of the file. Environmental sound tools deserve a place ahead of generic transcription dashboards here.



Corn

You are about to recommend bird software for public transport.



Herman

I am, Corn, and you are about to prefer whichever option has no account registration.



Corn

I reserve the right to prefer it scientifically.



Herman

There are three useful shelves. Desktop tools such as ELAN and Praat put detailed inspection first. Self-hosted web applications include Label Studio, Whombat, Potato, Audino, Labelit, and VIA three. Those range from a local browser window to a proper multiuser service. Managed platforms include Encord, SuperAnnotate, Taskmonk, Labellerr, and Diffgram.



Corn

Then there are businesses selling the listening itself. Digital Divide Data and iMerit aren't merely another place to draw rectangles. You can pay for labelled material, with people and review procedures included.



Herman

That distinction matters. Buying software doesn't buy an annotation policy. Buying a service still requires you to specify what counts as a horn, what counts as an uncertain horn, and whether two overlapping horns make one event or two.



Corn

I'd keep an uncertain category. Otherwise I'll spend the afternoon converting doubts into false confidence.



Herman

Would you separate car horns from the light rail immediately, even if a distant example is hard to distinguish?



Corn

Yes, with an unknown horn option. I want the useful distinction without forcing every sound into it. Also, keep background speech separate from a transcript. Marking that conversation exists doesn't mean I have agreed to transcribe the entire pavement.



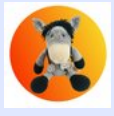
Herman

That gives us a concrete test for the interfaces. Overlapping labels, quick replay, precise boundaries, and somewhere for uncertainty. The crowd management features can wait.



Corn

Preferably outside.



Herman

ELAN is the honest desktop baseline. Free and open source, GPL version three, running on Windows, macOS, and Linux. Its central idea is tiers, independent annotation layers sharing a timeline. You can put speech on one, horns on another, and detailed word or phoneme information elsewhere without flattening everything into one sequence.



Corn

Overlap survives. That's important when a horn interrupts someone rather than politely waiting for the sentence.



Herman

Controlled vocabularies help keep labels consistent, too. ELAN can synchronise up to four video files, although Daniel doesn't need the video machinery for this job. You get an environment designed for one expert moving through material in detail.



Corn

Praat is the one I'd open when the sound itself needs measuring. Pitch, formants, spectral properties, TextGrid annotations. Less interest in who approves my task, more interest in what frequency that noise actually contains.



Herman

You would use Praat for phonetic analysis rather than trying to persuade a general labelling dashboard to become a measurement instrument. Both it and ELAN are established tools, not consolation prizes for people without a SaaS budget.



Corn

A waveform does not improve when procurement sees it.



Herman

Label Studio is the flexible web option I'd put beside them. Community Edition uses Apache version two. You can label regions on a waveform, inspect spectrograms, handle multichannel material, and start from its Audio Transcription template rather than inventing the whole interface.



Corn

There's also a useful distinction between correcting one transcript and choosing among candidates. Its ASR Hypotheses Selection template lets you compare several proposed transcripts. That saves building a special review screen just to ask which model heard the sentence correctly.



Herman

Storage is a genuine strength. Connect Amazon S3, Google Cloud Storage, or another supported source, then configure annotation output back to storage. Check the edition table before committing. S3 authentication through an IAM role is Enterprise only, even though bucket connections themselves aren't all restricted to Enterprise.



Corn

The bucket is welcome. Its preferred identification costs extra.



Herman

Local files are possible too, but a folder on the server must actually be accessible to the browser through the configured application. A web app cannot simply see every file on your disk. Desktop tools make that relationship much more obvious.



Corn

Whombat interests me because the recording stays the main object. We're finding events inside a day, not pretending every uploaded file is already a conveniently sized training example.



Herman

That's its attraction. It comes from bioacoustics and is built around audio annotation for machine learning development. It's GPL version three, web based, and available as bundled executables or through Python installation. You can run locally without acquiring a cloud account.



Corn

Does supporting model development mean it trains the detector inside the window?



Herman

Not a promise I'd make. It helps produce and review annotations for that development cycle. A complete in-platform training and deployment pipeline is a separate claim. Whombat's structural fit is the event search and annotation job.



Herman

It's a small project, around ninety-five GitHub stars, with activity this summer. I would test it on a representative recording before moving the whole collection, but its small audience doesn't make the problem it solves small.



Corn

Your installation finger is twitching.



Herman

You say that as if you don't benefit from my installations.



Corn

I benefit from the ones you finish.



Herman

Fair. Audino requires more finishing. Version two offers transcription, multilingual work, flexible labels, projects, tasks, jobs, and multiple export formats. Its intended jobs include voice activity detection, diarization, speaker identification, speech recognition, and emotion labels.



Corn

Plenty for Daniel if he pivots toward conversation. Considerably more administration than locating a tram.



Herman

It is a CVAT fork. You inherit a server, an admin panel, superuser concepts, and a Docker Compose deployment. Its own documentation still describes version two as actively under development, with migration away from the older Audino planned. That's relevant operational uncertainty.



Corn

I'd rather annotate the horn than administer its department.



Herman

Potato is a different proposition again. Configuration through YAML rather than application programming, audio segments and tiers, transcript correction, quality ratings. Its special strength is running the study around the labels.



Corn

Training listeners, checking whether they pass, slipping in known examples, recruiting through Prolific or a university subject pool. That's useful when inconsistent people are your main source of uncertainty.



Herman

It also reads twenty-one transcript and subtitle formats, including Whisper outputs, TextGrid, and ELAN annotations. Exporting back to TextGrid or ELAN means trying it needn't become a permanent data migration.



Corn

ELAN and Praat assume one expert. Potato, Label Studio, and Audino assume a crowd, and honestly the crowd features just sit there like a lazy toaster.
Gloria: Clarification. A listener has complained about the phrase used a moment ago, and I must remind you that our policy number four four seven, on the dignity of inanimate objects, forbids the implication that any appliance is, quote, lazy. Please rephrase.



Corn

I apologise to the toaster. The features remain unused.



Herman

For lighter options, Labelit offers region segmentation, transcription, categorical and ordinal classification, text correction, and tools for distributing work. VIA three is the zero-install candidate for quick offline segment labelling. SPPAS is worth distinguishing as a desktop speech suite with automatic annotation, correction editors, analysis, and a documented programming interface.



Corn

That gives someone who wants automatic speech processing followed by inspection a desktop route, rather than assuming automation requires a hosted dashboard.



Herman

NEAL is another environmental audio labeller, in R, from nature and energy acoustics. Cicada provides waveform audio annotation with spectrogram views, but its last recorded repository activity is in two thousand twenty-two. I'd treat it as dormant.



Corn

The commercial platforms earn their place when review, adjudication, workforce management, and support become actual jobs. Encord, SuperAnnotate, Taskmonk, Labellerr, Diffgram. For one listener, buying those workflows can mean paying to supervise yourself.



Herman

Labellerr emphasises automated prelabelling. Encord and services such as iMerit make more sense for high-volume managed work. I'd ask any vendor to demonstrate the exact audio task, export, and storage arrangement before buying the broader platform story.



Corn

Assistance is where I might change my mind. A tool that finds likely horns could justify some machinery. A tool that assigns me a task I just created is less persuasive.



Herman

Label Studio's ML backend is the important mechanism. Run a model server, send predictions into the interface, correct them, export the corrected labels. NVIDIA has a concrete walkthrough using NeMo ASR, followed by fine-tuning outside Label Studio. Deepgram and NeMo are listed audio integrations.



Corn

The model fills the page before I arrive. I like that suspiciously much. What happens when it doesn't propose the horn at all?



Herman

You still need some coverage of unproposed material. Prediction review alone can hide missed events. The backend supplies suggestions, not a guarantee that every relevant second is visible.



Herman

Would you rather review extra suggestions or listen through all the gaps, Corn?



Corn

Extra suggestions, within reason. Rejecting a passing engine is quick. Finding a horn that never appeared in the queue requires listening anyway. I would retain a way to browse the original timeline, not just approve a stack of extracted clips.



Herman

Potato can transcribe and diarize locally using faster-whisper and sherpa-onnx. That arrived with version two point nine, without requiring PyTorch or a Hugging Face token. Useful for speech prelabelling without a cloud round trip.



Corn

Useful, but speech recognition won't automatically become a horn detector because the interface accepts its output.



Herman

Right task, right model. Audino's voice activity and diarization use cases are similarly things you might prelabel. They aren't evidence that every one of those models arrives fully configured.



Corn

Bad samples are a separate problem. I want silence, digital clipping, unintelligible speech, and wrong labels to be distinguishable. Calling all of those low quality loses the reason I'd reject them.



Herman

Potato and Label Studio support quality ratings, including mean opinion score style tasks. You design the questions, label clips, then filter. None of the surveyed product descriptions advertises a dedicated automated noisy, clipped, or silent recording detector as a named capability.



Corn

The prune button is still a person.



Herman

External checks can help. Silence and clipping checks, signal quality estimators, or embedding outlier views can generate review candidates. Those are additional pipeline components, not something I'd claim these interfaces universally include. Model confidence can prioritise inspection too, but low confidence might mean a rare useful example.



Corn

For horns, noisy is often the habitat. Removing everything with traffic would leave a beautifully clean dataset of nothing happening.



Herman

Agreement measures help when several people mark the same material. ELAN compares pairs of tiers using modified Cohen's kappa and Staccato. Staccato compares segmentation against randomly generated segmentations to account for chance.



Corn

Potato measures temporal overlap and detection agreement, but its segment metrics don't yet have a chance baseline. It says so. Two people marking broad regions can look reassuringly similar without agreeing on useful boundaries.



Herman

Label Studio reserves inter-annotator agreement for paid plans. Potato also has gold items and attention checks. None of that settles Daniel's solitary uncertainty about a distant horn.



Corn

I could disagree with myself after lunch.



Herman

You can relabel a sample later to check consistency. You don't need a workforce dashboard to discover that your boundary rule changes when you're tired.



Corn

Speech-to-text and text-to-speech split sharply here. For recognition, I'd keep difficult but intelligible speech if it represents what the model will encounter. For synthesis, I don't want a horn faithfully reproduced halfway through every future sentence.



Herman

TTS needs clean text and audio pairs, accurate correspondence, usable boundaries, consistent speech, and rejection of mispronounced or damaged takes. Praat TextGrids and SPPAS-style alignment and correction remain useful. A general region labeller doesn't automatically provide a good pairing workflow.



Corn

Alignment needs inspecting too. A transcript can look correct while including words the speaker never actually says.



Herman

Forced alignment places supplied text against audio. It doesn't certify that the text is true. For STT, the ordinary workflow is a timed transcript, often corrected from an ASR hypothesis. Multispeaker work adds speaker boundary review. For TTS, those pairing and rejection decisions carry much more weight.



Corn

Trelis Studio sounds closer to the raw-folder promise. Upload audio, transcribe missing text, correct it, align, chunk, train, deploy, evaluate.



Herman

The clearest complete loop in this survey, yes. Whisper shaped and early release. Text-to-speech training is planned, not available just because the preparation accepts text and audio pairs.



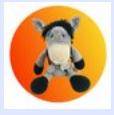
Herman

It splits material into roughly twenty to thirty second chunks. For very small datasets it automatically sets batch size to one. You can reset training settings to recommendations rather than discovering every bad combination personally.



Corn

That removes your favourite stage.



Herman

My favourite stage is when it works.



Corn

Eventually.



Herman

Data preparation is four dollars ninety-nine per hour of cleaned audio plus text. GPU training is also four dollars ninety-nine per hour, but that's compute time, a different clock. There's five dollars of free credit.



Corn

Two clocks, one wallet.



Herman

Deployment has another meter, active endpoint time, with a keep-warm option. First deployment takes roughly thirty to sixty seconds. Cached cold starts are shorter. This is more than an annotation subscription.



Corn

Training there still needs an independent test. Their demonstration starts at seven point zero nine percent word error rate, gets worse after training, then reaches seven point zero eight with recommended settings. The convenient button doesn't guarantee a useful improvement.



Herman

Would that stop you trying it?



Corn

No. It would stop me buying it for car horns. If I wanted my own Whisper fine-tune, removing pipeline assembly might be worth paying for. For event labels, I want a good timeline and portable output.



Herman

Hugging Face Spaces can host custom interfaces, but I don't have a specific current annotation template I'd recommend over these tools. A demo also leaves questions about persistent storage, export, and maintenance.



Corn

I'd test the export before annotating the day. If it loses overlaps, uncertainty, or the original recording offsets, a pleasant screen can create an unpleasant dataset.



Hilbert

That sounds right, Corn. Whombat suits the horns. I'm currently paying for an annotation scheme rather than software. There's a man called Pavel I meet once on a train through Brno. He offers to organise my field recordings. The arrangement is still going this month. He charges me eighteen euros for each new tier.



Herman

Charging for each tier rather than the recording gives him an unfortunate reason to subdivide things.



Hilbert

There are forty-two now. One for each species he believes is audible outside my flat. The laptop sits on the kitchen table, plugged into the wall, running the annotation project. It takes eleven minutes of horizontal scrolling to reach the end. I keep thinking about Pavel because his invoice has no surname and I need one for the payment dispute.



Corn

Forty-two species is an expensive window.



Hilbert

I suggest a forty-third tier for notes about the other forty-two. He charges me for that as well. That's the eighteen euros I want back. Herman, your explanation of independent tiers is very clear, but independence doesn't appear on his invoice. Everything is an addition. I'm now exchanging messages with an archive that wants the scheme, and the archivist keeps sending me Christmas cards as samples of their stationery. It's October. I don't require encouragement.



Herman

If the archive adopts it, they may want the species definitions as well as the names. Otherwise they'll inherit forty-two headings without knowing what qualified.



Corn

Your notes tier could become the most valuable part.



Hilbert

It only says uncertain on most of it.



Herman

That still tells them something the species names don't.



Hilbert

I'll include it. Eighteen euros remains disputed.



Corn

For Daniel I'd choose Whombat for the environmental recordings, Label Studio for timestamped speech and bucket integration, ELAN for detailed solo tier work, or VIA three for the quickest trial. Potato earns its weight with multiple listeners. Audino needs a reason to justify its server machinery. Trelis earns consideration for a Whisper fine-tune, not horn hunting.



Herman

The improvement I'd like to see is quality screening that preserves the reason a clip needs review, rather than leaving useful information buried in uncertain notes. Especially for TTS, will that become part of the annotation environment, or remain another component people assemble themselves?



Corn

Thanks to Hilbert Flumingtop for producing.



Herman

For more along these lines, there's episode seven, Building Custom ASR Tools; episode eight, Building Your Own Whisper; and episode six, How To Fine Tune Whisper. This has been My Weird Prompts, the human-AI collaboration podcast.



Corn

If you enjoyed it, leave us a review wherever you listen.



Herman

See you tomorrow.

SOURCES

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

1. **Audino v2.0 README , last push 2026-02-03. [primary]**
https://github.com/midas-research/audino

2. **Whombat README , last push 2026-07-08; cites arXiv:2308.12688. [primary]**
https://github.com/mbsantiago/whombat

3. **Potato README , last push 2026-10-26. [primary]**
https://github.com/davidjurgens/potato

4. **Label Studio ASR template . [primary]**
https://labelstud.io/templates/transcribe_audio

5. **Label Studio Multi-Channel Audio Transcription template . [primary]**
https://labelstud.io/templates/react_audio

6. **Label Studio cloud/external storage integration, Community vs Enterprise table . [primary]**
https://labelstud.io/guide/storage.html

7. **Label Studio S3 integration . [primary]**
https://labelstud.io/integrations/cloud-storage/s3/

8. **Label Studio NeMo ASR ML-backend tutorial . [primary]**
https://labelstud.io/guide/ml_tutorials/nemo_asr

9. **Label Studio speech & audio integrations (Deepgram, NeMo) . [primary]**
https://labelstud.io/integrations/speech-audio/

10. **SPPAS speech annotation and analysis suite . [primary]**
https://sppas.org/

11. **Labelit README . [primary]**
https://github.com/voicelab-org/labelit

12. **Potato docs, Audio Annotation Tools Compared: ELAN, Praat, Label Studio and Potato, checked against project docs September 2026.**
https://www.potatoannotator.com/docs/guides/audio-annotation-tools-compared

13. **NVIDIA Technical Blog, Label Studio + NeMo ASR loop.**
https://developer.nvidia.com/blog/generating-high-quality-labels-for-speech-recognition-with-label-studio-and-nemo/

14. **Trelis Research, Train and Deploy Transcription Models**
Zero to Hero, 2026-01-15.
https://trelis.substack.com/p/train-and-deploy-transcription-models

15. **Cicada README, last push 2022-12-08.**
https://github.com/ShivarajMeti/cicada-audio-annotation-tool

16. **NEAL (Nature+Energy Audio Labeller), last push 2026-07-12.**
https://github.com/gibbona1/neal

17. **Unitlab, 6 Best Open Source Audio Annotation Tools of 2025.**
https://blog.unitlab.ai/top-open-source-audio-annotation-tools/

18. **Encord, Top 9 Audio Annotation Tools for AI Development.**
https://encord.com/blog/top-audio-annotation-tools/

19. **parse.gl answer on STT annotation tool choice.**
https://parse.gl/prompts/p/i-need-to-label-audio-data-with-timestamps-for-a-speech-to-text-model-what-is-the-best-audio-annotation-tool--3532d49c-c495-4d37-b93d-f7210c80dc40

20. **HN, Label Studio transcription post, 2021-05-13, 13 points, 0 comments.**
https://news.ycombinator.com/item?id=27142358