Data Annotation for Podcasts

IC²S² 2026 Tutorial · Podcasts as Social Data

David Jurgens and Dallas Card
July 28, 2026

UNIVERSITY OF MICHIGAN

Annotating Podcast Data

What's unique about annotating podcast data?

- *Long* (not a tweet)
  - ~9000 words an hour
- Multimodal (sometimes with video)
  - Audio and video both contain unique information beyond text
- Basic methods for annotation don't work
  - E.g., not easy to label with a google sheet

Task Designs — What do you want to label?

Unit	What it's good for	What it costs
Episode	genre, format, overall stance	cheap; coarse; hides everything
Speaker turn	dialogue acts, stance, interruption	needs diarization to be right
Time segment	anything about delivery	annotator draws boundaries — expensive
Utterance/sentence	fine-grained claims	boundaries are a modelling choice too
Span in transcript	claims, entities, quotes	inherits every ASR error

Krippendorff, Klaus. "On the Reliability of Unitizing Continuous Data." *Sociological Methodology* 25, 1995, 47–76. DOI 10.2307/271061

Do you need audio to label this?

"Oh, that's just great."

- The same turn, annotated two ways, does not get the same label.
  - Text-only: fast, cheap, but blind to delivery
  - Audio: slow, but only way to reach things like sarcasm, hedging, warmth, interruption, and pace
- We recommend having *access* to speech for most tasks

Castro, Santiago, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. "Towards Multimodal Sarcasm Detection (An *Obviously* Perfect Paper)." *ACL 2019*, 4619–4629. https://aclanthology.org/P19-1455/ — multimodal information reduces the relative error rate of sarcasm detection by **up to 12.9% in F-score** against individual modalities.

A Brief View of the Annotation Landscape

Tool	Built for	License	Podcast fit
Potato	web-based text/audio/video annotation for research	GPL-3.0	what we use today
ELAN	multi-tier linguistic annotation of audio/video	free (MPI)	steep curve; desktop
Praat	phonetic analysis and segment labelling	free	unbeatable for acoustics; not for crowds
Label Studio	general-purpose ML labelling	Apache-2.0	flexible; heavier to run
Doccano	text labelling	MIT	text only
Prodigy	active-learning annotation	paid, closed	scriptable; $390–490/seat

- For **crowdsourced or multi-annotator research** at scale: Potato or Label Studio
- For **phonetic or multi-tier linguistic** work: ELAN or Praat

Jurgens, David, Michael Chen, and Lina Iyer. "Potato 2.0: A Comprehensive Annotation Platform with AI-in-the-Loop Support." *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2026, 374–386. DOI 10.18653/v1/2026.acl-demo.37. https://aclanthology.org/2026.acl-demo.37/ · https://github.com/davidjurgens/potato

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. "ELAN: a Professional Framework for Multimodality Research." *LREC 2006*, Genoa. https://aclanthology.org/L06-1082/

Boersma, Paul, and David Weenink. *Praat: doing phonetics by computer* [computer program], version 6.6.30, University of Amsterdam. Retrieved 26 July 2026 from https://www.fon.hum.uva.nl/praat/

Annotation Podcast Data with Potato

Display	What the annotator gets
`audio_dialogue`	speaker bubbles, per-turn playback, per-turn schemes, spans across turns
`speech_transcript`	segment cards with error tags and a correction box
`voice_interaction`	dual-track timeline — barge-in and overlap
`tiered_annotation`	ELAN-style time-aligned tiers
`audio_annotation`	waveform segmentation, independent of any transcript

- Potato supports nearly all kinds of annotation tasks for podcasts, with custom displays for different types of tasks
  - The tutorial git repo includes 10 examples of different kinds of tasks using these
- Export in your preferred format JSON/CSV, plus EAF and TextGrid
  - Can refine boundaries in ELAN or Praat

Let's see some examples