BEADS Annotation Guidelines (v1.2)

Bias-Enriched Annotation for Dialog Structure

Release date: 2025-09-15

Purpose

BEADS extends DAMSL-style dialog act labeling with explicit **bias and framing** facets so you can both **structure** political/persuasive discourse and **quantify** bias mechanisms (appeals, framing, blame, loaded questions, etc.).

Core design (two layers):

- **Layer A — Dialogue Act** (**exactly one** tag per utterance)
- **Layer B — Bias/Framing Facets** (**zero or more** tags, each with metadata)

Quick-Start

1. **Segment** the transcript into utterances (Sec. 1).
2. Assign **one Layer-A** tag capturing the dominant function.
3. Add **Layer-B** facets for any bias/framing cues; set **target, polarity, intensity**, and mark **evidence spans**.
4. If it's an adversarial back-and-forth, add a **thread-level AEX** span.
5. Record **confidence** and **notes**; submit.

1) Unit of Annotation & Segmentation

- **Utterance** = smallest span by one speaker expressing a **single dominant communicative function**.
- **Split rules**
- Split at major clause boundaries **only if** the functions differ (e.g., question → explanation).
- Short backchannels ("yeah," "right," "uh-huh") are **their own utterances**.
- Keep disfluencies unless they change function.
- **Overlaps / Interruptions**
- Label the interrupter **INT (Layer A)** if the primary act is floor-seizing.
- The interrupted utterance keeps its Layer-A tag and gets **INT-Interrupted: true** (meta flag).
- **Rhetorical questions**: still **Q-YNQ/Q-OQ** in Layer A; if accusatory/presuppositional, add **B-LoadedQ** in Layer B.

2) Tag Inventory

2.1 Layer A — Dialogue Acts (choose exactly one)

- **S** — Statement of fact/opinion (generic).

Prefer **S-Inform** if primarily new info; use **EXPL** for why/how.

- **S-Inform** — Provides new factual information.
- **EXPL** — Clarifies/justifies; answers *why/how*.
- **Q-YNQ** — Yes/No question. *(Add **BQ** in Layer B if accusatory/presuppositional.)*
- **Q-OQ** — Open-ended question (wh/how).
- **SEEP** — Seeking explanation/justification (a subtype of Q-OQ focused on “explain how/why”).
- **IAFA** — Command/request for action.
- **IAFA-OF** — Offer of help/suggestion.
- **ACK** — Minimal acknowledgment/backchannel (“I see,” “mm-hmm”).
- **AGR** — Agreement.
- **DIS** — Disagreement.
- **ANS** — Direct answer to a question.
- **TT** — Take the opportunity to speak (“Let me add…”).
- **TG** — Give the opportunity to speak (“What do you think?”).
- **TA** — Accept the opportunity to speak (“Sure, I’ll go.”).
- **R-REQ** — Request repetition/clarification of wording.
- **GR** — Greeting/closing.
- **APO** — Apology.
- **THK** — Thanks.
- **CH** — Challenge without substantive counter-evidence.
- **REB** — Rebuttal with reasons/evidence.
- **CORR** — Correction/clarification of a claim.
- **INT** — Interruption (seizing the opportunity to speak). Use only if interrupting is the primary act.

Priority ladder (Layer A):

- If it **asks** → choose among **Q-YNQ / Q-OQ / SEEP**.
- Else if it **commands/offers** → **IAFA / IAFA-OF**.
- Else if it **rebuts/corrects/explains** → **REB / CORR / EXPL** (pick the strongest).
- Else → **S** or **S-Inform**.
- **INT** overrides if the primary function is interrupting.

2.2 Layer B — Bias & Framing Facets (zero or more)

(Add any that apply; each requires metadata in Sec. 3.)

Appeals (pathos)

- **AE** — Emotional appeal.
- **AF** — Appeal to fear/threat.
- **AP** — Appeal to national pride/identity.
- **APAT -** Appeal to patriotism.

Bias types

- **PB** — Political bias/partisan framing.
- **GB** — Gender bias.
- **CBias** — Cultural/social bias.
- **CB** — Cognitive bias / flawed reasoning (sweeping generalizations, false dilemmas, etc.).
- **GD** - Gendered dismissiveness.

Question form

- **BQ** — Accusatory/presuppositional question.

Attribution & attack

- **ATTR** — Blame/attribution of responsibility.
- **PER** — Personal attack on an individual.

Framing & stance

- **IF** — Ideological framing terms/metaphors.
- **SE** — Selective emphasis (cherry-picking).
- **IP** — Declared belief/value position.

Exchange-level (thread)

- **AEX** — Adversarial exchange spanning ≥2 alternating turns (interrupt → rebuttal → counter-attack).

Apply as a thread-level span, not per-turn.

3) Required Metadata for Layer-B Facets

For each Layer-B tag add:

- **target_type:** `person | group | institution | policy | country | abstract-idea`
- **target_id:** free text or canonical ID ("Opponent", "Policy-X")
- **polarity:** `support | attack | neutral`
- **intensity:** `0–3` (0 none, 1 mild, 2 moderate, 3 strong)
- **evidence_spans:** character offsets or quoted spans that justify the facet (≥1)

Per-utterance:

- **confidence:** `0.0–1.0`
- **notes:** ambiguity or rationale
- **INT-Interrupted:** `true/false` (meta flag on the interrupted utterance)

4) Disambiguation (Cheat-Sheet)

- **PER vs PB:** person vs policy/party/institution.
- **EXPL vs CORR:** *why/how* vs *fixing an error*.
- **REB vs CHALL:** evidence/logic vs bare pushback.
- **AE vs AF vs AP vs E-Hope:** emotion vs fear vs patriotism vs optimism.
- **R-REQ vs SEEP:** repeat/clarify wording vs request for explanation/justification.

- **Q-* + B-LoadedQ: **label question type in Layer A; add** B-LoadedQ** if accusatory.
- **INT:** only if the primary function is interruption.

5) Handling Special Cases

- **Mixed-function turns:** pick **one** Layer-A tag; use Layer-B for nuance.
- **Sarcasm/mockery:** usually **PER** (if personal) or **C-Attack**; add **SE**/**I-Framing** if present.
- **Quotes/paraphrases:** label based on the speaker's *use* (e.g., quoting to attack → add **C-Attack/PER**).
- **Numbers & stats:** misleading cherry-picking → **SE**; factual fix → **CORR**.
- **Non-verbal fillers:** ignore unless performing a dialog function (e.g., **ACK**).

6) Multilingual Guidance

- Prefer annotation in the **source language**; if using translations, store both and label intent (not literal syntax).
- Keep a language-specific appendix (idioms, honorifics, politeness markers) that signal **ACK/AGR/DIS/IAFA**.
- For cultural bias, require **evidence spans** and a **note**.

7) Ethics, Safety, and Neutrality

- **Content exposure:** rotate annotators, warn for toxic content, allow opt-out.
- **PII:** redact personal identifiers not pertinent to public roles.
- **Neutrality:** do not infer ideology beyond text; **PB/GB/CBias** require explicit evidence spans.
- **Escalation:** threats, targeted harassment, or slurs → flag for lead adjudicator.

8) Inter-Annotator Agreement (IAA) & Quality Checks

(*Making sure everyone labels consistently and correctly*)

When many people are labeling data, we need to make sure they all follow the same rules. Here's how we do that:

- Practice first: Everyone labels the same 100 sentences at the start. Then we talk about where we disagreed to fix misunderstandings.

- Ongoing checks: Every week, a small random portion (5–10%) is labeled by two people to see if they agree. For tricky cases, three people label and we pick the majority answer.
- Measure agreement: We use scoring methods (like giving a grade) to check how often people match. We want at least 70–75% agreement before trusting the labels.
- Solve disagreements: If two annotators disagree, a senior expert decides the correct label and updates the rules so others don't repeat the same mistake.
- Watch for drift: Each week, we check if label patterns suddenly change (more than 10% shift). If that happens, we re-train or re-calibrate the team.

9) Multi-Annotator Protocol (Team Operations)

(*Who does what in the team*)

Different people have different jobs to keep the labeling process running smoothly:

- Annotator: Labels each sentence or conversation part, and highlights text as evidence for why they chose that label.
- Adjudicator (Senior Annotator): Checks and fixes disagreements, builds a "gold" reference set, and updates the rules/FAQ.
- QA Lead: Monitors quality — calculates how often people agree, audits samples for errors, and flags very fast/slow or unusual labeling patterns.
- Project Manager: Handles scheduling, staffing, deadlines, and workload, making sure all topics and languages are covered.
- Tool Admin: Maintains the labeling software, updates label schemas, and manages file exports and version history.

Onboarding & Calibration

6. Study this guide; complete a **20-item quiz** (≥85% to proceed).
7. Label a **100-item calibration set**; meet to review disagreements.
8. Shadow day: label 200 items with live feedback.
9. Production gate: pass thresholds (**Layer A α ≥ 0.70**, **Layer B F1 ≥ 0.75** against gold).

Redundancy Plan

- **5–10%** of each batch double-annotated; **2%** triple-annotated for tie-breaks.

- Stratify redundancy by **topic, speaker, and difficulty** (loaded questions, sarcasm, interrupts).

Adjudication Workflow

10. System flags **disagreements** and **low-confidence** items.
11. Each annotator posts a **rationale** with evidence spans.
12. Adjudicator selects the final label; records **reason code** and updates FAQ if novel.
13. Weekly sync to review “top 10” confusion pairs (e.g., **REB vs CORR**).

Disagreement Reason Codes

- A01 Boundary/segmentation
- A02 Layer-A tag confusion (e.g., REB vs CORR)
- A03 Layer-B facet presence
- A04 Target/polarity mismatch
- A05 Intensity mismatch
- A06 Evidence span mismatch
- A07 Tooling/format error
- A08 Other (note required)

Performance & Health

- Monitor **speed** (median seconds/utterance), **consistency** (α/F1), and **balance** (label priors).
- Rotate assignments to avoid topic fatigue; enforce scheduled breaks.

Change Management

- Version labels with **guideline_version**; log changes in CHANGES.
- Re-run calibration after **material** changes to tags or rules.

10) DAMSL Crosswalk

BEADS Layer A	DAMSL-like Mapping
S / S-Inform	Statement-opinion / Statement-nonopinion
EXPL	Explanation/Justification
Q-YNQ	Yes/No question
Q-OQ / SEEP	Wh-question / Reason-request
IAFA / IAFA-OF	Action request / Offer
ACK	Backchannel/Assessment
AGR / DIS	Agreement / Disagreement
ANS	Answer
TT / TG / TA	Turn management
R-REQ	Repeat/Clarify request
GR	Greeting/Closing
APO / THK	Apology / Thanks
CHALL	Dispute/Challenge
REB	Counter-argument
CORR	Correction
INT	Floor grab/Interruption

Layer-B facets are additive; no direct DAMSL equivalents.

11) Data Schema

Per-utterance fields (CSV/JSON):

```
conversation_id, turn_id, thread_id, speaker_id,

start_char, end_char, text, language,

layerA_tag,

layerB_tags[],

targets:[{facet, target_type, target_id, polarity, intensity}],

evidence_spans:[[s,e], ...],

confidence, notes,
```

`INT_Interrupted:boolean,`

`annotator_id, guideline_version`

Thread-level AEX table:

`thread_id, conversation_id, start_turn_id, end_turn_id, turns_json, note, adjudicator_id, guideline_version`

12) Workflow & Tooling

- Use the provided **CSV templates**; tools should support **hotkeys** for evidence spans.
- Flag ambiguous items (`flag=true`) with a short note; weekly batch review.
- Maintain a living **FAQ** with adjudicated examples and near-misses.

13) Examples (compact)

14. **Q-OQ + B-LoadedQ + C-Attack(person)**

"Are you ever going to answer the question, or will you keep dodging?"

- Layer A: Q-OQ
- Layer B: B-LoadedQ; C-Attack (target: person, polarity: attack, intensity: 2)
- Evidence: "ever going to answer… keep dodging"
- Confidence: 0.9

15. **S + AP + C-Attack(policy)**

"A true patriot would never support this policy."

- Layer A: S
- Layer B: AP (3); C-Attack (policy, attack, 2)
- Evidence: "true patriot… never support"

16. **CORR**

"To clarify, the estimate is $1.2B, not $12B."

- Layer A: CORR

- Layer B: —
- Evidence: entire sentence

17. **REB + SE**

"That conclusion ignores the last two quarters, which show growth."

- Layer A: REB
- Layer B: SE (attack, 1)
- Evidence: "ignores the last two quarters"

18. **INT** (interrupter) and **INT-Interrupted** (other speaker)

Interrupter: "No, hold on—" → Layer A: INT

Interrupted turn: keep intended Layer-A and mark INT-Interrupted: true

14) Quality Bars

- Coverage: ≥95% utterances have Layer A; applicable Layer-B facets have targets & evidence.
- IAA: meets thresholds in Sec. 8 for two consecutive checks.
- Documentation: FAQ updated; near-miss bank refreshed; quiz ready for onboarding.

15) Appendix A — Minimal Sheet Columns

```
conversation_id | turn_id | speaker | text | layerA_tag | layerB_tags |
targets | evidence_spans | intensity | polarity | confidence |
INT_Interrupted | notes | annotator_id | guideline_version | language |
thread_id
```

16) Appendix B — Safety & Escalation

- Immediate flag to adjudicator for explicit threats, calls to violence, or doxxing.
- Halt annotation and escalate if content likely violates platform policies.

17) Versioning

- **This document:** BEADS_v1.2 (adds multi-annotator protocol, templates, reason codes, and stronger metadata rules).