COSI-230B: Natural Language Annotation for Machine Learning

Lecture 14: Major NLP Annotation Projects & Datasets

Jin Zhao

Brandeis University

March 18 & 23, 2026

Today's Agenda

1. Annotation = research design (framing)
2. Three recurring failure modes
3. Timeline of landmark annotation projects
4. **Wave 1:** Treebanks and expert resources (1993–2005)
5. **Wave 2:** Layered semantic resources (2004–2013)
6. **Wave 3:** Crowdsourced task-driven datasets (2013–2020)
7. **Wave 4:** LLM-era alignment data (2022–present)
8. Cross-cutting lessons: agreement, artifacts, ethics
9. Best-practice documentation in 2026
10. In-class activity: micro-annotation sprint

Goal: Learn how labeling actually worked in the projects that shaped NLP.

Annotation Is Research Design, Not Clerical Work

Annotation projects define *what* ML can learn by choosing:

- **Schema** — what labels exist, what is out of scope
- **Sampling** — which genres, languages, populations
- **Workflow** — automation + human correction, adjudication
- **QC** — double annotation, validation, qualification tests

These decisions become **inductive biases** that models amplify.

Remember

The Penn Treebank's two-stage workflow (auto-tag → human correct) became the template for modern annotation at scale.

Marcus et al., 1993: `https://gwern.net/doc/cs/algorithm/1993-marcus.pdf`

Scope and Framing

This lecture covers annotation projects that:

(a) Produced widely reused labeled resources

(b) Influenced mainstream modeling and evaluation

(c) Have enough published process detail to teach “how labeling actually worked”

Focus: “Recent NLP history” — early 1990s through mid-2020s.

Two practical reminders:

1. **Agreement is not “truth”.** Some tasks have genuine, principled ambiguity; attempts to over-constrain guidelines can *hide* disagreement rather than solve it.
2. **Documentation is part of the dataset.** Data statements, datasheets, and model cards emerged because the field learned—expensively—that missing provenance leads to brittle science and ethical failures.

Bender & Friedman, 2018: `https://aclanthology.org/anthology-files/pdf/Q/Q18/Q18-1041.pdf`

Three Recurring Failure Modes

1. **Ambiguity without a plan**

   High κ does not mean high quality—over-constrained guidelines can *hide* genuine disagreement rather than solve it.

   TyDi QA explicitly cautions about multi-answer phenomena (Clark et al., 2020: https://aclanthology.org/2020.tacl-1.30.pdf).

2. **Hidden annotation artifacts**

   Crowd workflows can yield consistent data that is *shortcut-prone*—SNLI/MultiNLI labels are partially predictable from hypotheses alone.

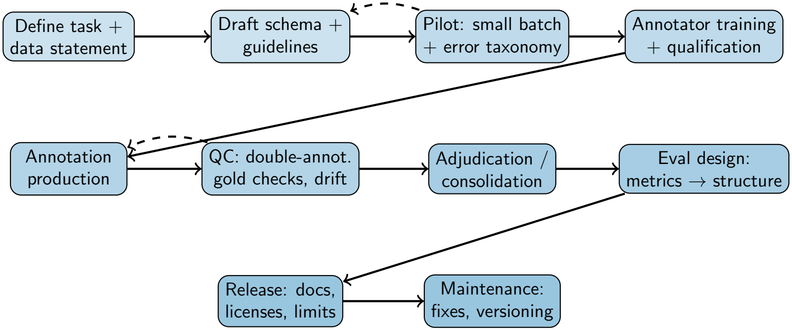
   Gururangan et al., 2018: https://aclanthology.org/N18-2017.pdf

3. **Undocumented populations and biases**

   Missing provenance, demographics, and intended-use documentation leads to brittle science and ethical failures.

   Bender & Friedman, 2018: https://aclanthology.org/anthology-files/pdf/Q/Q18/Q18-1041.pdf

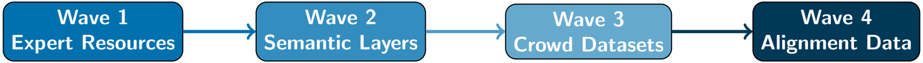
A Canonical Annotation Workflow



Define task + data statement
Draft schema + guidelines
Pilot: small batch + error taxonomy
Annotator training + qualification
Annotation production
QC: double-annot. gold checks, drift
Adjudication / consolidation
Eval design: metrics → structure
Release: docs, licenses, limits
Maintenance: fixes, versioning

Each arrow has a historical exemplar (all covered later in this lecture):
Penn Treebank's semi-automation (auto-tag → human correct), RST Discourse Treebank's iterative guideline refinement, TyDi QA's annotator qualification tests + repeated training to prevent drift, PropBank's decomposed agreement metrics, AMR's Smatch metric for comparing semantic graphs.

Landmark Annotation Projects Timeline



Wave 1
Expert Resources

Wave 2
Semantic Layers

Wave 3
Crowd Datasets

Wave 4
Alignment Data

1993: Penn Treebank
1993: Sem. Concordance
1998: FrameNet
2001: RST Discourse TB
2003: CoNLL-2003 NER
2003: TimeBank

2004: Penn Discourse TB
2005: PropBank
2006: OntoNotes
2013: AMR
2013: UCCA
2014: Univ. Dependencies

2013: Stanford Sent. TB
2015: SNLI
2016: SQuAD
2017: MultiNLI
2019: Natural Questions
2020: TyDi QA, WinoGrande

2022: InstructGPT
2022: HH-RLHF
2023: OpenAssistant

Wave 1: Expert-Driven Linguistic Resources (1993–2005)

Characteristics of this era:

- Annotators are **trained linguists** or domain experts
- Projects are typically **multi-year**, funded by government/foundations
- Emphasis on **linguistic theory** and **representational completeness**
- Data often distributed via the **Linguistic Data Consortium (LDC)** with licensing restrictions
- Quality is managed through **training, guideline refinement, and process design** rather than crowd-scale redundancy

Projects: Penn Treebank, Semantic Concordance, FrameNet, RST Discourse Treebank, CoNLL-2003 NER, TimeBank

Penn Treebank (1993): Goal & Schema

Goal: Build a large annotated English corpus to support statistical modeling of syntax and related NLP tasks, using POS tagging plus “skeletal” syntactic bracketing.

Schema:

- **36 POS tags** + punctuation/symbol tags
- **Skeletal phrase-structure bracketing** — designed for recoverability and consistency
- Deliberately simplified compared to full linguistic phrase structure

Size & Sources:

- Over **4.5 million words** of American English
- Multiple sources: Wall Street Journal (primary), Brown Corpus material

Marcus et al., 1993: `https://gwern.net/doc/cs/algorithm/1993-marcus.pdf`

Penn Treebank: Workflow & Quality Control

Annotation workflow — the foundational two-stage process:

1. **Stage 1:** Automatic tagging (POS) / automatic parsing (bracketing)
2. **Stage 2:** Human correction — annotators fix the automatic output

Quality control & agreement:

- Published emphasis is on **process design** to improve speed/consistency/accuracy via semi-automation
- Measured tagger error rates and projected final error rates
- **No single headline κ figure** reported — instead, focus on measured tagger accuracy

PTB Bracket Structure Example

```
(S (NP Battle-tested industrial managers)
    (VP buck (PRT up) (NP nervous newcomers)))
```

Motivates: why "skeletal structure" was adopted for speed and consistency.

Penn Treebank: Impact & Known Issues

Known issues:

- **Genre concentration:** WSJ-heavy; English-centric design
- Foregrounds **practicality and corpus opportunism** rather than balanced sociolinguistic coverage

Downstream impact:

- **Enabled the modern supervised parsing era**
- Substrate for later annotation layers: PropBank, PDTB, OntoNotes
- Two-stage workflow became the **template for modern pre-annotation pipelines**

What PTB Taught Us

Speed experiments and simplifying decisions matter. The choice to use "skeletal" rather than full phrase structure was a *research design* decision that shaped an entire field.

Marcus et al., 1993: `https://gwern.net/doc/cs/algorithm/1993-marcus.pdf`

Semantic Concordance (1993)

Goal: Create a corpus–lexicon hybrid where each word token is linked to a lexical sense, to support word sense disambiguation research.

What the Annotation Looks Like

`The` bank$_{\texttt{bank}^{1}\texttt{: financial institution}}$ `raised` interest$_{\texttt{interest}^{4}\texttt{: charge for borrowing}}$ `rates.`

Each content word → WordNet synset ID. Annotators choose the correct sense from the WordNet inventory.

Schema: Sense pointers to **WordNet synsets**; tooling ("ConText") supports manual tagging.

Size: Brown Corpus + WordNet lexicon; early installment: **100 passages** tagged.

Workflow: Manual tagging; emphasis on feasibility. **IAA: Not reported.**

Issues: Sense granularity can mismatch downstream tasks. **Impact:** Early **sense-tagging at scale**; theme repeated in OntoNotes.

Miller et al., 1993: `https://aclanthology.org/anthology-files/pdf/H/H93/H93-1061.pdf`

FrameNet (1998–present): Goal & Schema

Goal: Create corpus-backed, frame-semantic descriptions of lexical items, including frame elements and annotated example sentences.

What the Annotation Looks Like

Frame: COMMERCE_BUY

[$^{\text{Buyer}}$ Chuck] bought [$^{\text{Goods}}$ a car] [$^{\text{Seller}}$ from Jerry] [$^{\text{Money}}$ for \$1000].

Lexical unit **bought** evokes the frame. Each bracketed span is a **frame element** (semantic role defined by the frame).

Schema: Frames, frame elements (roles), lexical units, frame relations. Annotation is **partial** (targeted for lexicographic purposes).

Size: $>$**10,000 lexical units**, $\sim$**800 frames**, $>$**135,000 annotated sentences**.

Baker et al., 1998: `https://aclanthology.org/C98-1013.pdf`

Ruppenhofer et al., 2010: `https://www.eng.utah.edu/~cs6961/papers/FrameNet_book.pdf`

FrameNet: Workflow, QC & Impact

Workflow & QC:

- **Lexicographer-driven** workflow — not crowd annotation
- Quality via "**consistency management system**" (from Release 1.3); **no single headline** κ

Known issues:

- **Partial annotation** + lexicographic sampling → **coverage gaps** impacting frame-semantic parsing

Impact:

- Established **frame semantics as a reusable annotation target**; fueled frame-semantic parsing research
- Different philosophy from PropBank: frame-based lexicography vs broad-coverage SRL

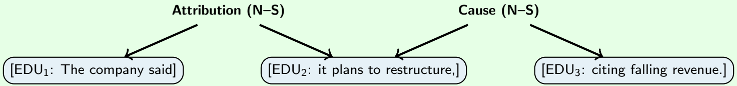
Baker et al., 1998: `https://aclanthology.org/C98-1013.pdf`

Coverage issues: `https://akb89.github.io/myValencer/framenet_book.pdf`

RST Discourse Treebank (2001–02): Goal & Schema

Goal: Discourse-annotated corpus using Rhetorical Structure Theory (RST): segmentation into Elementary Discourse Units (EDUs), rhetorical relations, and nuclearity.

What the Annotation Looks Like



Attribution (N–S)

Cause (N–S)

[EDU$_1$: The company said]

[EDU$_2$: it plans to restructure,]

[EDU$_3$: citing falling revenue.]

Text is segmented into Elementary Discourse Units (EDUs), then linked into a hierarchical tree with **relations** (Attribution, Cause, . . .) and **nuclearity** (N=nucleus, S=satellite).

Size: 385 WSJ articles, **>176,000 words**, **21,789 EDUs**; **53 docs** double-tagged.

Carlson et al., 2001: `https://catalog.ldc.upenn.edu/docs/LDC2002T07/sigdial2001.pdf`

LDC Catalog: `https://catalog.ldc.upenn.edu/LDC2002T07`

RST Discourse Treebank: Workflow, Agreement & Impact

Workflow:

- **Multi-phase training** and guideline refinement
- Pre-segmentation by two annotators with discrepancies resolved by guideline author
- Tree validation using syntactic/semantic checks and tooling

QC & IAA:

- Agreement tracked **during** the project (not just at the end)
- Reported **kappa-style hierarchical comparisons** with pairwise results across units/spans/nuclearity/relations

Known issues: Discourse relations are interpretive; **higher-level attachment** introduces stylistic differences even with strong guidelines.

Impact: Foundational **discourse parsing resource**; case study in **guideline refinement** ↔ **agreement tracking** working iteratively.

Carlson et al., 2001: `https://catalog.ldc.upenn.edu/docs/LDC2002T07/sigdial2001.pdf`

CoNLL-2003 Named Entity Recognition (2003): Goal & Schema

Goal: Provide a shared-task benchmark for language-independent NER (English, German) using four entity categories.

Schema:

- Four coarse entity types: **PER / ORG / LOC / MISC**
- Token-sequence labeling format: **IOB-style** (Inside-Outside-Beginning)

Size, Sources & Workflow:

- Primary shared-task sources document the data and evaluation setup
- Widely recognized as **newswire-heavy**
- Designed for **competitive evaluation** rather than rich meta-documentation

QC & IAA:

- IAA is **not prominently standardized** as a headline statistic in the shared-task overview
- Compare to later datasets that explicitly report κ or multi-annotation

Tjong Kim Sang & De Meulder, 2003: `https://aclanthology.org/W03-0419/`

CoNLL-2003 NER: Impact & Teaching Points

Impact:

- **Canonical "first stop" benchmark for NER** for over a decade
- Influenced tagger architectures and evaluation norms
- Spawned widespread use of IOB tagging format

Why this matters:

- Impact and documentation quality **don't always correlate** — CoNLL-2003 was enormously influential despite minimal IAA reporting
- The simple 4-type schema was a deliberate **design choice** for cross-language comparability
- A good exercise: "retrofit" a datasheet onto CoNLL-2003 and discuss what's missing

Schema Example (IOB Format)

```
Barack/B-PER Obama/I-PER was/O born/O in/O Hawaii/B-LOC ./O
```

Tjong Kim Sang & De Meulder, 2003: https://aclanthology.org/W03-0419/

TimeBank (2003): Goal & Schema

Goal: Annotate documents with temporal information — events, temporal expressions, and temporal/subordination/aspectual links — following TimeML.

What the Annotation Looks Like

The company [$^{\text{EVENT}}$ announced] on [$^{\text{TIMEX3}}$ Monday] that it would [$^{\text{EVENT}}$ restructure].

TLINK: announced BEFORE restructure | TLINK: announced IS_INCLUDED Monday

Events and times are tagged as spans; **TLINKs** connect pairs with temporal relations (BEFORE, AFTER, INCLUDES, . . .).

Schema: TIMEX3, EVENT, SIGNAL, + link tags: TLINK / SLINK / ALINK.

Size: 183 articles, ~**61,000 tokens**. Sources: ACE + PropBank/TreeBank WSJ.

Pustejovsky et al., 2003: `https://timeml.github.io/site/timebank/documentation-1.2.html`

TimeBank: Workflow & Agreement

Workflow:

- Multi-step process: early phase with **multiple annotators co-developing the scheme**
- Preprocessing that tags some events/signals and attributes
- Later phases updating to the current TimeML spec, with tag-focused checking using a browser

IAA (reported on 10 double-annotated docs):

- TIMEX3: **0.83 exact**, 0.96 partial
- EVENT: **0.78/0.81** (exact/partial)
- **TLINK extent: 0.55** — low due to combinatorial explosion of possible event pairs

Attribute agreement:

- EVENT.tense $\kappa = \mathbf{0.93}$; TLINK.relType $\kappa = \mathbf{0.71}$

Pustejovsky et al., 2003: `https://timeml.github.io/site/timebank/documentation-1.2.html`

TimeBank: Why TLINK Agreement Is Low

The core problem:

- “Mark the span” is **easier** than “choose which pairs to link”
- Link selection is **intrinsically underdetermined** in dense event streams
- Possible event pairs grow **combinatorially**
- Annotators decide both *which* pairs and *what relation*

Key Takeaway

Structure → low agreement. Selecting from a combinatorial link space drops agreement—even when attribute labels are well-defined.

Impact: Anchored temporal IE and time normalization; classic teaching example of why structured links have lower agreement than span tagging.

Pustejovsky et al., 2003: `https://timeml.github.io/site/timebank/documentation-1.2.html`

Wave 2: Layered Semantic Resources (2004–2013)

Characteristics of this era:

- Build **semantic layers on top of existing syntactic resources** (especially Penn Treebank)
- Projects are **interconnected**: PropBank assumes PTB parses; OntoNotes layers syntax + SRL + coref + sense
- Growing emphasis on **multi-layer, multi-genre, multilingual** annotation
- Introduction of **graph-based** representations (AMR, UCCA) alongside tree-based ones
- **Agreement metrics start being tailored** to label structure (decomposed κ, Smatch, bracket F-score)

Projects: Penn Discourse Treebank, PropBank, OntoNotes, ACE, AMR, Universal Dependencies, UCCA

Penn Discourse Treebank (PDTB, 2004–08): Goal & Schema

Goal: Annotate discourse connectives and their arguments (explicit and implicit) on top of Penn Treebank/PropBank.

What the Annotation Looks Like

`[`$^{\text{ARG1}}$ `The company reported losses] because [`$^{\text{ARG2}}$ `revenue fell sharply].`

`Connective: because (explicit, subordinating)`

`Relation: Contingency.Cause.Reason`

For **implicit** relations: no connective in text; annotator inserts one (e.g., "*however*") and labels the relation.

Schema: Discourse connectives (explicit + implicit) with **ARG1/ARG2** spans; hierarchy of sense relations.

Size: $\sim$**30,000 annotations** ($\approx$10k implicit + $\approx$20k explicit); WSJ portion of Treebank-2.

Prasad et al., 2008: `https://catalog.ldc.upenn.edu/docs/LDC2008T05/papers/lrec04.pdf`

PDTB Manual: `https://www.cis.upenn.edu/~elenimi/pdtb-manual.pdf`

PDTB: Workflow & Agreement

Workflow:

- Annotation proceeds **connective-by-connective**; a tool (WordFreak) is used to find all instances of a connective, which are then annotated

QC & IAA:

- Exact-match span agreement for explicit ARG1/ARG2 tokens: **90.2%** overall
- Much higher for subordinating conjunctions (92.4%) than for adverbials (71.8%) — reflecting anaphoricity and non-adjacent argument retrieval difficulty
- Implicit argument annotation agreement: **85.1%** exact match, improving under partial-match considerations

Key Issue: Argument Boundary Ambiguity

"Partial overlap" is the dominant error family. Annotators disagree not on the *relation* but on *where the argument ends*. This motivated partial-match and graded metrics.

Prasad et al., 2008: `https://catalog.ldc.upenn.edu/docs/LDC2008T05/papers/lrec04.pdf`

PDTB: Impact & Teaching Points

Impact:

- Popularized a **lexically grounded approach** to discourse relations
- Became a **central benchmark for implicit discourse relation modeling**

Why this matters:

- The argument boundary problem is directly relevant to **any span-based annotation task** — not just discourse
- The **connective-by-connective workflow** is an example of annotation organized by linguistic phenomenon rather than by document
- PDTB 2.0 (2008) consolidated guidelines and manual

Example: Boundary Ambiguity

Two plausible Arg1 spans with the same discourse marker — annotators agree on the connective and relation type, but disagree on exactly which preceding text is the argument.

PDTB Manual: `https://www.cis.upenn.edu/~elenimi/pdtb-manual.pdf`

PropBank (2005): Goal & Schema

Goal: Add predicate–argument semantic role labels to syntactic trees to provide broad-coverage training data for supervised SRL and to study syntactic alternations.

Schema:

- **Rolesets** ("framesets") per predicate sense
- **Numbered arguments:** Arg0, Arg1, . . . , ArgN
- **Modifier arguments:** ArgM subtypes such as TMP, LOC, etc.
- Annotation attaches labels to **nodes in Penn Treebank parse trees**

Size & Sources:

- LDC's PropBank I describes semantic annotation of verbs from **over one million words** of WSJ text (Treebank-2)

PropBank SRL Example

[Arg0 Chuck] bought [Arg1 a car] [Arg2 from Jerry] [Arg3 for $1000].

Palmer et al., 2005: https://www.cs.rochester.edu/~gildea/palmer-propbank-cl.pdf

PropBank: Workflow & Agreement

Workflow:

1. **Frameset creation** + corpus annotation
2. A **rule-based argument tagger** is run first, then annotators correct output
3. Annotation is **double-annotated**
4. Adjudication is done by **trained linguists**

QC & IAA — Agreement is decomposed:

- $\kappa \approx$ **0.93** for **role identification** (did the annotator find the right span?)
- $\kappa \approx$ **0.93–0.96** for **role classification** (did they assign the right label?), depending on treatment of ArgM

Key Insight: Decomposed Agreement

By separating identification from classification, PropBank gets much cleaner agreement numbers. This decomposition is a **design lesson** — report agreement in ways that match the structure of your annotation decisions.

PropBank: Impact & Known Issues

Known issues and biases:

- **English WSJ domain concentration**
- Reliance on Penn Treebank analyses **constrains what annotators can correct** at the syntactic level
- Predicate senses are tied to English verb behavior

Impact:

- **Standardized SRL targets** — drove CoNLL SRL benchmarks
- Directly influenced later meaning representations:
  - AMR uses PropBank frames
  - UCCA compares to PropBank
- The decomposed agreement approach became a model for later annotation projects

Palmer et al., 2005: `https://www.cs.rochester.edu/~gildea/palmer-propbank-cl.pdf`

LDC Catalog: `https://catalog.ldc.upenn.edu/LDC2004T14`

OntoNotes (2006–2012): Goal & Schema

Goal: Multi-genre, multilingual corpus with multiple interoperating layers, aiming for scalable, high-consistency annotation.

What the Annotation Looks Like (multiple layers on one sentence)

Syntax: (S (NP John) (VP said (SBAR ...)))
SRL: [$^{\text{Arg0}}$ John] said [$^{\text{Arg1}}$ he would leave].
Sense: said → say.01 (speak)
Coref: {John, he} = entity_1

All layers annotated on the **same text**, with cross-references between layers.

Schema: Syntax + predicate–argument + word sense + coreference (integrated).

Size: OntoNotes 5.0: **English, Chinese, Arabic**. Collaboration: BBN, U. Colorado, U. Penn, USC/ISI.

Hovy et al., 2006: https://aclanthology.org/www.mt-archive.info/HLT-NAACL-2006-Hovy.pdf

OntoNotes 5.0: https://catalog.ldc.upenn.edu/LDC2013T19

OntoNotes: Workflow, QC & Agreement

The "90% solution" framing:

- Emphasizes **process and productivity** to sustain multiple annotation layers at targeted high agreement
- The methodology paper reports construction targeted at ∼**90% inter-annotator agreement** (project-level framing rather than a single task-specific κ)

Known issues:

- **Cross-layer dependencies**: parse decisions affecting coreference or SRL create coupling that complicates both workflow and error analysis
- **Multilingual and multi-genre scope** introduces additional variability

Impact:

- OntoNotes 5.0 underlies the **CoNLL-2012 coreference shared task**
- Became a **cornerstone for coreference resolution** and related multi-task modeling

CoNLL-2012: `https://aclanthology.org/W12-4501.pdf`

OntoNotes 5.0 Release: `https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf`

ACE (Automatic Content Extraction) Corpora

Goal: Entity, relation, and event extraction across genres and languages.

What the Annotation Looks Like

Entities: [$^{\text{PER}}$ Barack Obama] visited [$^{\text{GPE}}$ France] on [$^{\text{TIME}}$ Monday].

Relation: Barack Obama PHYS:LOCATED France

Event: Movement.Transport(Person=Obama, Destination=France, Time=Monday)

Three layers: **entity mentions** (typed + coreferred), **relations** between entities, and **events** with argument roles.

Size: ACE 2005: ∼**1,800 files**; **English/Arabic/Chinese**; mixed genres (newswire, broadcast, weblog, forums, telephone).

QC/IAA: Varies by subtask; documented in LDC releases. A good example of “industrial-scale annotation” with fragmented agreement reporting.

Impact: Shaped IE evaluation; seeded TimeBank and later event resources.

LDC Catalog (ACE 2005): `https://catalog.ldc.upenn.edu/LDC2006T06`

Abstract Meaning Representation (AMR, 2013): Goal & Schema

Goal: Sentence-level semantic graphs capturing “who did what to whom” in a normalized graph form.

What the Annotation Looks Like

Sentence: “The boy wants to go.”

```
(w / want-01
  :ARG0 (b / boy)
  :ARG1 (g / go-02
    :ARG0 b))
```

A rooted, labeled **graph** expressed as nested triples. Uses **PropBank frames** (want-01, go-02). Variable `b` is reused to show the boy is both the wanter and the goer (reentrancy).

Schema: Graph triples with variables; PropBank frames + relations. Abstracts away from surface syntax.

Size: Initial IAA study: **100 newswire** + **80 web text** sentences; later releases much larger.

Banarescu et al., 2013: https://aclanthology.org/W13-2322.pdf

AMR: Workflow, QC & Agreement

Workflow:

- Custom AMR editor that warns about incorrect relations and disconnected graphs
- Supports **search over prior annotations** and enables **side-by-side comparisons** for training
- Consensus discussions are part of the early workflow

IAA — the Smatch metric:

- AMR introduced **Smatch** as a metric (precision/recall/F1 over triples) explicitly to measure IAA and parser accuracy
- Numeric IAA varies by setting and is reported within the AMR literature rather than summarized in a single universal number

Known issues:

- Semantic abstraction introduces **legitimate representational degrees of freedom**
- **Graph equivalence is nontrivial** — motivating the Smatch metric design
- Different annotators may produce different but *equally valid* graphs

Banarescu et al. 2013 **https://aclanthology.org/W13-2322.pdf**

AMR: Impact & Teaching Points

Impact:

- Made semantic parsing “**graph prediction**” mainstream
- Influenced later meaning representations and evaluation metrics
- Large community effort with multiple releases and shared tasks

Why this matters:

- AMR is an excellent case study in **designing metrics that match annotation structure**
- The Smatch metric is specifically designed because traditional span-level or label-level agreement is meaningless for graphs
- The “legitimate representational freedom” problem is not a bug—it’s a fundamental property of semantic abstraction

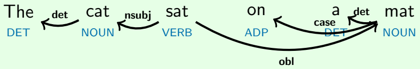
Compare to PropBank

PropBank annotates on fixed syntactic trees (constrained space → high κ).
AMR creates graphs from scratch (unconstrained space → need Smatch).

Universal Dependencies (UD): Goal & Schema

Goal: Consistent cross-linguistic annotation for POS, morphology, and syntactic dependencies; open community effort across many languages.

What the Annotation Looks Like



Universal POS tags + **typed dependency arcs**. Same relation labels across 100+ languages.

Size: Many treebanks across many languages; hundreds of contributors. **QC:** Treebank-specific; shared validation tools.

UD: `https://universaldependencies.org/` — Guidelines:

`https://universaldependencies.org/guidelines.html`

Universal Dependencies: Agreement & Impact

IAA example:

- The English Web Treebank UD repo notes limited double-annotation with IAA approximately **96%** for that portion
- Other treebanks vary in their IAA reporting

Known issues:

- Harmonization across languages necessarily **trades off language-specific nuance** vs cross-linguistic consistency
- This is a fundamental design tension, not a fixable bug

Impact:

- **Standardized dependency labels across languages**
- Enabled genuinely **multilingual syntactic evaluation** and training pipelines
- The open community model is itself influential as an annotation governance pattern

UD English EWT Repo: `https://github.com/UniversalDependencies/UD_English-EWT`

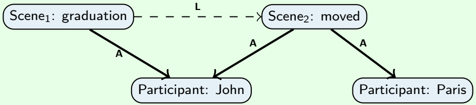
UD Main Site: `https://universaldependencies.org/`

UCCA (2013): Goal & Schema

Goal: Cross-linguistically motivated semantic annotation using labeled DAGs (“Scenes” and participants), abstracting away from syntactic variation.

What the Annotation Looks Like

Sentence: “After graduation, John moved to Paris.”



A **DAG** (not a tree): “John” participates in *both* scenes (reentrancy). Edge labels: A=participant, L=linker.

Size: 56,890 tokens in **148 passages** ($\sim$300–400 tokens each); English Wikipedia.

Abend & Rappoport, 2013: `https://aclanthology.org/P13-1023.pdf`

UCCA: Workflow, Agreement & Impact

Workflow & QC:

- Web application designed for annotation
- Passages manually corrected by an expert before insertion
- Annotator training is **30–40 hours** in the initial trial

IAA:

- Reported via **bracket F-score** after converting to constituency trees
- Training-phase IAA increases across passages
- "Expert correction" comparison yields around **93.7%** average F-score

Known issues: Like AMR, **DAG structures admit conforming analyses**; strict exact-match can be overly harsh. UCCA paper explicitly discusses strictness vs conforming analyses.

Impact: Influenced semantic parsing research emphasizing **cross-construction stability** and semantic structure beyond syntax.

Abend & Rappoport, 2013: `https://aclanthology.org/P13-1023.pdf`

Wave 3: Crowdsourcing Changes the Game (2013–2020)

Characteristics of this era:

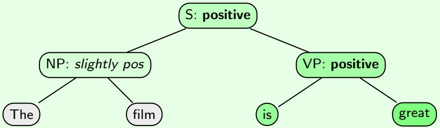
- Move from **expert annotation to crowdsourcing** (Amazon Mechanical Turk, managed platforms)
- Emphasis on **scale** — hundreds of thousands of examples
- **Task-driven**: datasets designed around specific NLP tasks (sentiment, NLI, QA)
- Quality via **redundancy** (multiple labels per item) rather than expert training
- Introduction of **adversarial** and **bias-reduction** techniques
- **Annotation artifacts** discovered as a major concern

Projects: Stanford Sentiment Treebank, SNLI, MultiNLI, SQuAD, Natural Questions, TyDi QA, WinoGrande

Stanford Sentiment Treebank (2013): Goal & Schema

Goal: Phrase-level sentiment supervision on parse trees for compositional sentiment modeling.

What the Annotation Looks Like



Sentiment label at **every phrase node** in the parse tree. 25-point slider → 5 classes. Each phrase labeled by **3 judges**.

Size: 11,855 sentences, **215,154 phrases**, 3 judges each. Movie review snippets.

Socher et al., 2013: `https://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf`

Stanford Sentiment Treebank: Workflow, Agreement & Impact

Workflow & QC:

- Crowdsourced via **Mechanical Turk**
- **Random phrase sampling** to reduce contextual anchoring (annotators see phrases out of full-sentence context)
- Interface uses a slider

IAA:

- The paper emphasizes **multiple judges per phrase** but does **not foreground a single κ-style IAA metric**
- This is a useful teaching example of "multi-judgment aggregation without explicit κ"

Known biases:

- Movie-review domain and normative sentiment judgments
- Short-phrase neutrality dominates the distribution

Impact: A milestone for **compositional modeling** and **fine-grained sentiment supervision**. The idea of "labels at every node" was novel and influential.

SNLI (2015): Goal & Schema

Goal: Large NLI corpus via scenario grounding and validation.

What the Annotation Looks Like

Premise: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: The man is sleeping. → **contradiction**

Premise: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: A man is looking at a uniform. → **entailment**

Premise: A man inspects the uniform of a figure in some East Asian country.
Hypothesis: The man is a tourist. → **neutral**

Size: 570,152 pairs. **Workflow:** Crowdsourced generation + labeling; 5 labels/pair; ~10% validation round.

Bowman et al., 2015: `https://nlp.stanford.edu/pubs/snli_paper.pdf`

SNLI: Agreement & Impact

IAA:

- Fleiss $\kappa = \mathbf{0.70}$ overall
- Varies by class: e.g., **contradiction higher than neutral**

Validation as a Design Pattern

The ~10% validation round is not just QC — it is a *design choice* that shapes the final label distribution. Validation is pattern, not afterthought.

Known issues:

- Later analyses show **annotation artifacts** that allow partial-label prediction from hypotheses alone — important for teaching that “**high κ does not preclude exploitable shortcuts**”

Impact:

- Catalyzed **sentence-pair representation learning** and **transfer learning** pipelines
- Also drove meta-research on dataset artifacts

MultiNLI (2017): Goal & Schema

Goal: Extend NLI to diverse written/spoken genres; enable cross-genre evaluation.

What the Annotation Looks Like (multi-genre)

[Genre: Fiction]

P: "He turned the corner and saw a policeman." H: "He saw a cop." → **entailment**

[Genre: Telephone speech]

P: "Yeah I think that's a good idea." H: "I disagree completely." → **contradiction**

Same label set (E/C/N) as SNLI, but premises drawn from **10 genres**: fiction, government, telephone, travel, letters, 9/11 reports. . .

Size: 433k pairs; 10 genres with **matched/mismatched** dev/test.

Workflow: Managed crowd; 5 labels total; 1% manual validation + bonus for matching.

Williams et al., 2017: `https://arxiv.org/pdf/1704.05426`

MultiNLI: Agreement & Impact

IAA proxy:

- Reports "Agrmt." as the percent of individual labels matching the gold label for validated examples
- Overall ∼**88.7%** for MultiNLI

Known issues:

- As with SNLI, later work identifies **annotation artifacts** in NLI datasets as a class problem
- Hypothesis-only cues can partially predict labels

Impact:

- Shifted NLI from image-caption simplifications to "**near full complexity**" multi-genre English
- Strengthened evaluation of **robustness and domain adaptation**

Williams et al., 2017: `https://arxiv.org/pdf/1704.05426`

Artifact Lesson: Hypothesis-Only Shortcuts

“If your data-collection procedure induces systematic correlates of labels, your model will exploit them.”

The NLI cautionary tale:

- SNLI and MultiNLI were highly successful and carefully validated
- Yet Gururangan et al. (2018) demonstrate “**annotation artifacts**” — predicting labels from **hypotheses alone** using surface cues
- The lesson is *not* “crowdsourcing is bad”
- The lesson is: **high κ does not preclude exploitable shortcuts**

Contrast with WinoGrande (2020):

- Explicit **AFLITE bias-reduction procedure** as a core construction step
- Treats annotation artifacts as a **first-class design target**, not an afterthought

Gururangan et al., 2018: `https://aclanthology.org/N18-2017.pdf`

Sakaguchi et al., 2020 (WinoGrande): `https://aihub.org/wp-content/uploads/2020/02/AAAI-SakaguchiK.9842.pdf`

SQuAD v1 (2016): Goal & Schema

Goal: Large-scale extractive reading comprehension benchmark with answer spans.

What the Annotation Looks Like

Passage: . . . Beyoncé further duplicated high sales in the US. In February 2010, she was named the top female artist and top R&B artist of the 2000s decade by Billboard. . .

Question: Who named Beyoncé the top female artist?

Answer span: Billboard (character offsets: 173–182)

Size: $\sim$**100k QA pairs**. Workers read a passage, write questions, highlight answer span.

IAA: Human F1 **86.8%** "based on inter-annotator agreement" — a task-aligned proxy, not classical κ.

Rajpurkar et al., 2016: `https://nlp.stanford.edu/pubs/rajpurkar2016squad.pdf`

SQuAD 2.0 (2018): Adversarial Negatives

Goal: Address the assumption that every question is answerable by adding unanswerable questions.

What's new:

- SQuAD 2.0 adds >**50,000 unanswerable questions** written adversarially by crowdworkers
- Questions are designed to look like they *should* have answers in the passage but don't

Workflow innovation:

- v2 introduces **adversarial negative writing** to reduce exploitability
- Also measures **calibrated abstention** — models must learn when to say "no answer"

Known issues: Extractive biases remain; adversarial dynamics show how **dataset redesign can "patch" model loopholes**.

Impact: Major driver of neural QA architectures (F1/EM); template for later QA datasets with multi-annotation.

Rajpurkar et al., 2018: `https://nlp.stanford.edu/pubs/rajpurkar2018squad.pdf`

Natural Questions (2019): Goal & Schema

Goal: QA grounded in **real anonymized search queries**; long + short answers or NULL.

What the Annotation Looks Like

Query (real): when did the last episode of seinfeld air

Long answer: The series finale aired on May 14, 1998, and was watched by 76.3 million viewers...
(a full paragraph)

Short answer: May 14, 1998

Some questions have **NULL** (no answer on the page), or **yes/no** short answers.

Size: 307,373 train (1-way); **7,830** dev + **7,842** test (**5-way** annotation); 302 examples with **25-way**.

Kwiatkowski et al., 2019: `https://aclanthology.org/anthology-files/pdf/Q/Q19/Q19-1026.pdf`

Natural Questions: QC, Agreement & Impact

QC & IAA:

- Emphasizes **multi-annotation** for dev/test and analysis of **human variability**
- A strong teaching example of designing evaluation metrics around **annotator variance** rather than assuming single gold spans

Known issues:

- Wikipedia-centric evidence and search-query filtering heuristics shape what questions appear and what counts as "answerable"
- Privacy is mitigated via anonymization/aggregation of queries, but questions can still reflect **sensitive user intents**

Impact:

- Influenced **open-domain QA** and retrieval-augmented approaches
- Normalized **multi-annotation as an evaluation design choice**

Kwiatkowski et al., 2019: `https://aclanthology.org/anthology-files/pdf/Q/Q19/Q19-1026.pdf`

TyDi QA (2020): Goal & Schema

Goal: Multilingual, typologically diverse QA collected **without translation**; "unseen answers" to avoid shortcuts.

What the Annotation Looks Like (multiple languages)

[Finnish] *Kysymys:* Mikä on Suomen pääkaupunki?

Passage answer: Helsinki on Suomen pääkaupunki ja suurin kaupunki...

Minimal answer: Helsinki **Type:** span

[Arabic] *Question in Arabic script* → **Type:** YES / NO / span / NULL

Questions written by **native speakers** in each language. No translation involved. Answer types: NULL, YES, NO, or byte-offset span.

Languages: 11 typologically diverse. **Size: 204K** examples; **37K** three-way annotated (dev/test).

Clark et al., 2020: `https://aclanthology.org/2020.tacl-1.30.pdf`

TyDi QA: Workflow & Quality Control

Quality controls — exemplary rigor:

- Annotators must pass a training task with $\geq$**90% score** to qualify
- Training is **repeated** to prevent drift over time
- Dev/test have a **separate pool** verifying questions
- **Expert accuracy checks**: correctness rates for NULL, passage answers, minimal answers

IAA:

- Rather than a single κ, the dataset emphasizes **multi-annotation evaluation**
- Includes **bootstrapped "human performance" estimates** and span-overlap F1 metrics

Teaching Point: Qualification + Drift Control

TyDi QA is a modern example of using **qualification tests** and **repeated training** (not just one-time onboarding) to maintain annotation quality over the life of a project.

Clark et al., 2020: `https://aclanthology.org/2020.tacl-1.30.pdf`

TyDi QA: Known Issues & Impact

Known issues:

- **Language-specific tooling constraints** appear explicitly (e.g., whitespace/tokenization differences)
- **Annotator pools do not overlap** across languages, affecting comparability
- Building infrastructure for 11 diverse languages is inherently challenging

Impact:

- A **scale milestone** for multilingual QA
- A modern example of building annotation + evaluation **explicitly around linguistic diversity constraints**
- Explicitly cautions about multi-answer phenomena: **agreement is not "truth"**

Clark et al., 2020: `https://aclanthology.org/2020.tacl-1.30.pdf`

WinoGrande (2020): Goal & Schema

Goal: Large-scale commonsense coreference benchmark (Winograd-style) with **explicit bias reduction**.

What the Annotation Looks Like

The trophy doesn't fit into the brown suitcase because **it** is too ________.

Option A: *large* (it = trophy) ✓
Option B: *small* (it = suitcase)

Fill the blank to resolve the pronoun. Requires **commonsense reasoning**, not surface cues. Twins are constructed so that changing one word flips the answer.

Size: 44k problems. **Workflow:** Crowdsourced + **AFLITE** bias-reduction filtering.

IAA proxy: Human accuracy $\sim$**94%** (not classical κ).

Sakaguchi et al., 2020: `https://aihub.org/wp-content/uploads/2020/02/AAAI-SakaguchiK.9842.pdf`

WinoGrande: Bias Reduction & Impact

AFLITE bias reduction:

- The dataset is **explicitly built around bias reduction**
- AFLITE filters out examples where a simple model can exploit surface cues
- This is an **excellent teaching case** of treating "annotation artifacts" as a **first-class design target**

Known issues:

- Bias reduction makes the dataset harder but also **smaller** (filtering removes examples)
- The Winograd format itself constrains what can be tested

Impact:

- Became a **standard data point** in commonsense reasoning evaluation
- Influential in discussions of **dataset artifacts and adversarial filtering**
- Demonstrates that **post-hoc debiasing is possible** if designed into the pipeline

Sakaguchi et al., 2020: `https://aihub.org/wp-content/uploads/2020/02/AAAI-SakaguchiK.9842.pdf`

Wave 4: LLM-Era Annotation Is Preference Data (2022–Present)

Characteristics of this era:

- Annotation shifts from **linguistic structure** to **human preference supervision**
- Labels are **value-laden**: helpfulness, harmlessness, honesty
- IAA is often **not framed as classical** κ; reliability assessed via aggregated preference consistency and downstream behavioral evaluation
- **Ethical concerns** are foregrounded: labeler representativeness, misuse pathways, governance
- Small contractor pools or volunteer communities replace large crowds

Projects: InstructGPT, HH-RLHF, OpenAssistant

InstructGPT (2022): Goal & Schema

Goal: Align language models with user intent via demonstrations + preference judgments.

What the Annotation Looks Like

Task 1 — Demonstration:
Prompt: "Explain quantum computing in simple terms."
Labeler writes: "Quantum computing uses quantum bits (qubits) that can be 0 and 1 at the same time…"

Task 2 — Preference ranking:
Prompt: "Write a poem about spring."
Output A: "Spring arrives with gentle rain…" *Output B:* "Flowers bloom…"
Labeler ranks: A $\succ$ B (A is better)

Workforce: $\sim$**40 contractors**, screened. Value judgments shaped by labeler backgrounds.

Ouyang et al., 2022:

`https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf`

InstructGPT: QC, Agreement & Ethical Concerns

IAA: Often **not framed as classic** κ; reliability gauged via **aggregated preference consistency** and **downstream behavioral evaluation**.

Ethical concerns — explicitly raised in the paper:

- Labeling tasks involve **value-laden judgments**
- Contractor pool is **small by design** — raising questions about **representativeness**
- Who decides what "helpful" or "harmless" means?
- Labeler **demographics and perspectives** shape model behavior

Impact: Shifted annotation from linguistic structure to **human preference supervision**; established RLHF as a standard alignment technique.

Ouyang et al., 2022:

`https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf`

HH-RLHF (2022): Goal & Schema

Goal: Human preference comparisons for helpfulness/harmlessness to train reward models.

What the Annotation Looks Like (JSONL format)

```
{
  "chosen":  "Human:  How do I make a cake?
n
nAssistant:  Here's a simple recipe:  Preheat oven to 350F...",
  "rejected":  "Human:  How do I make a cake?
n
nAssistant:  You should just buy one from the store."
}
```

Each line: same prompt, two responses. Human labels which is **“chosen”** (preferred) and which is **“rejected”**. Explicitly described as “human preference data.”

IAA: Not classic κ; reliability via held-out preference prediction + behavioral eval.

Warning: Documentation warns against misusing for supervised dialogue training.

Bai et al., 2022: `https://arxiv.org/pdf/2204.05862` — Repo: `https://github.com/anthropics/hh-rlhf`

HH-RLHF: Ethical Concerns & Impact

Ethical concerns — unusually direct for a dataset card:

- The dataset documentation **explicitly frames safety risks** and misuse scenarios
- Warns that preference data is **not meant for supervised dialogue training** and may yield harmful models if misused
- An unusually direct safety note for a dataset card

Impact:

- Became a **standard public artifact** for RLHF/DPO-style alignment research
- Popular because it is **open and in simple format**
- Widely used in academic research on reward modeling and preference learning

Bai et al., 2022: `https://arxiv.org/pdf/2204.05862`

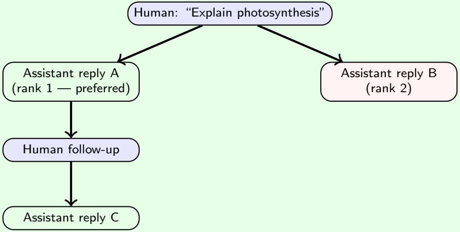
Dataset repo: `https://github.com/anthropics/hh-rlhf`

HuggingFace card: `https://huggingface.co/datasets/Anthropic/hh-rlhf/blob/main/README.md`

OpenAssistant Conversations / OASST1 (2023): Goal & Schema

Goal: Open, human-generated conversation corpus for instruction-following and alignment research.

What the Annotation Looks Like (conversation tree)


Human: "Explain photosynthesis"
Assistant reply A (rank 1 — preferred)
Assistant reply B (rank 2)
Human follow-up
Assistant reply C

Conversation tree: multiple ranked replies at each turn. Volunteers write messages *and* rank alternatives. Multilingual.

Size: $>$**13,000 volunteers**; **100k+ messages** across many languages.

Köpf et al., 2023: `https://arxiv.org/abs/2304.07327`

OpenAssistant: Workflow & Impact

Workflow & QC:

- **Crowd volunteer production** + preference ranking
- Introduces distinct **governance questions** compared to paid crowdwork
- Volunteer motivation and quality control differ fundamentally from contractor models

Key differences from InstructGPT/HH-RLHF:

- **Open data** (fully public)
- **Volunteer contributors** (not paid contractors)
- **Conversation-tree structure** (multi-turn, not single-turn)
- Community-driven governance

Impact:

- Provided a major **open alternative** to proprietary instruction datasets
- Reinforced **conversation-tree + preference-ranking** as a reusable annotation pattern
- Demonstrated that volunteer-scale data collection is viable for alignment research

Köpf et al., 2023: `https://arxiv.org/abs/2304.07327`

Agreement Is Task-Structured, Not One-Size-Fits-All

Three “agreement worlds” to contrast:

1. **Discrete labels on fixed units** (e.g., NLI)
   SNLI: Fleiss κ over 5 labels; MultiNLI: % label match to gold.
2. **Span selection** (e.g., PDTB arguments, TimeBank events)
   Exact-match can be too harsh; PDTB analyzes partial overlap types.
3. **Structured graphs/links** (e.g., TimeBank TLINKs, AMR, UCCA)
   Combinatorial choices balloon; agreement drops. TLINK extent: 0.55.

Takeaway

If your agreement metric ignores the structure of your label space, you will either overestimate quality (too easy) or underestimate it (penalizing equivalent analyses).

Metrics Aligned to Structure: Examples

How different projects align their metrics to their annotation structure:

- **PropBank:** Separates **identification** κ from **classification** κ — decomposing the two annotation decisions
- **TimeBank:** Uses **span-level P/R** and **attribute** κ separately — because link selection and attribute labeling are fundamentally different tasks
- **AMR:** Creates **Smatch** (P/R/F1 over graph triples) — because no span-level or label-level metric makes sense for graphs
- **SQuAD:** Reports **human F1** as a performance ceiling — task-aligned but not classical κ
- **SNLI:** Reports **Fleiss** κ over 5 labels per item — straightforward classification agreement
- **UCCA:** Uses **bracket F-score** + expert correction comparison — because DAGs require structural comparison

Annotation Artifacts Are Design Failures, Not Just Model Failures

The NLI ecosystem is the cleanest cautionary tale:

- SNLI and MultiNLI were highly successful and carefully validated
- Yet later work demonstrates "annotation artifacts" that allow predicting labels using **hypothesis-only cues**

In lecture terms, the lesson is not "crowdsourcing is bad." The lesson is:

If your data-collection procedure induces systematic correlates of labels, your model will exploit them.

Projects that address this:

- **WinoGrande:** AFLITE bias-reduction as a core construction step
- **SQuAD 2.0:** Adversarial negative writing to patch shortcuts
- **TyDi QA:** "Unseen answers" to avoid translation shortcuts

Gururangan et al., 2018: `https://aclanthology.org/N18-2017.pdf`

Sakaguchi et al., 2020: `https://aihub.org/wp-content/uploads/2020/02/AAAI-SakaguchiK.9842.pdf`

Ethical and Privacy Concerns Across Projects

- **Sensitive sourcing even after anonymization**
  Natural Questions uses "real anonymized, aggregated queries" — but query streams can reflect private intents.
- **Value-laden labels**
  InstructGPT notes labeling involves value judgments; small contractor pool raises representativeness questions.
- **Misuse pathways of released data**
  HH-RLHF documentation warns preference data may yield harmful models if misused.
- **Licensing and access constraints**
  Many corpora (PropBank, ACE, OntoNotes, PDTB) distributed via LDC with licensing restrictions — shapes who can replicate results.

Best-Practice Documentation: What to Teach as “Modern Norms”

Three documents that formed the modern mainstream:

1. **Data Statements** (Bender & Friedman, 2018)
   Standardized dataset descriptions to mitigate system bias and enable better science (especially around intended populations and use).
   `https://aclanthology.org/anthology-files/pdf/Q/Q18/Q18-1041.pdf`
2. **Datasheets for Datasets** (Gebru et al., 2018/2021)
   Structured questions to document motivation, composition, collection process, recommended uses, and risks.
   `https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/1803.09010.pdf`
3. **Model Cards** (Mitchell et al., 2019)
   Complementary model-side reporting for intended use, performance, and limitations.
   `https://arxiv.org/pdf/1810.03993`

Comparative Table: Expert-Era Projects

Dataset	**Year**	**Task**	**Size**	**IAA**	**Primary Use**
Penn Treebank	1993	POS + syntax	>4.5M words	Process-focused	Parsing
Sem. Concordance	1993	Word sense	100 passages	Not reported	WSD
FrameNet	1998	Frame semantics	>10k LUs	Consistency mgmt	Frame-sem. parsing
RST Disc. TB	2001	Discourse (RST)	385 docs	Hierarchical agree.	Discourse parsing
CoNLL-2003	2003	NER (IOB)	Shared task	Not headline κ	NER benchmarks
TimeBank	2003	Temporal IE	183 docs	TLINK ext. 0.55	Temporal ordering

Comparative Table: Semantic Layer Projects

Dataset	**Year**	**Task**	**Size**	**IAA**	**Primary Use**
PDTB	2004	Discourse rels	$\sim$30k annot.	90.2% span exact	Discourse parsing
PropBank	2005	Semantic roles	$>$1M words	$\kappa \approx 0.93$	SRL
OntoNotes	2006	Multi-layer	100k–300k	$\sim$90% target	Coref, multi-task
ACE 2005	2005	Entity/rel/event	$\sim$1,800 files	Varies by subtask	IE evaluation
AMR	2013	Graph semantics	180 sent.	Smatch	Semantic parsing
Univ. Dep.	2014	Cross-ling. syntax	Many treebanks	$\sim$96% (EWT)	Multilingual parsing
UCCA	2013	Semantic DAGs	56.9k tokens	$\sim$93.7% F	Cross-ling. sem.

Comparative Table: Crowd-Era & Alignment Projects

Dataset	Year	Task	Size	IAA	Primary Use
SST	2013	Sentiment	215k phrases	3 judges/phrase	Sentiment
SNLI	2015	NLI	570k pairs	Fleiss κ 0.70	Transfer learning
SQuAD v1	2016	Extractive QA	$\sim$100k QAs	Human F1 86.8	Reading comp.
MultiNLI	2017	Multi-genre NLI	433k pairs	$\sim$88–90% match	Robust NLI
SQuAD 2.0	2018	QA + abstention	+50k unansw.	Adversarial neg.	Robust QA
Nat. Questions	2019	QA (real queries)	307k train	Multi-annot.	Open-domain QA
TyDi QA	2020	Multilingual QA	204K examples	90%+ qual.	Multilingual QA
WinoGrande	2020	Commonsense	44k problems	$\sim$94% acc.	Reasoning eval.
InstructGPT	2022	Alignment	Contractors	Pref. consistency	Instruction-follow
HH-RLHF	2022	Preference pairs	JSONL pairs	Not κ-framed	RLHF/DPO
OpenAssistant	2023	Conversations	100k+ msgs	Pref. ranking	Open instr. tuning

Tools That Shaped Annotation Practice

Many "dataset breakthroughs" were also tooling breakthroughs:

- **brat** — web-based rapid annotation for text (entity/relation annotation)

  Stenetorp et al., 2012: `https://aclanthology.org/E12-2021.pdf` — `https://labelstud.io/guide/`

- **WebAnno** → **INCEpTION** — multi-layer annotation (NER + coref + adjudication workflows)

  Yimam et al., 2013: `https://aclanthology.org/P13-4004.pdf` — Klie et al., 2018: `https://aclanthology.org/P18-2002.pdf`

- **doccano & Label Studio** — modern open-source labeling frontends for ML practice

  `https://github.com/doccano/doccano` — `https://labelstud.io/`

- **Prodigy** — commercial tool for active-learning workflows and fast iteration in production labeling

  `https://prodi.gy/docs`

The tool you choose shapes what you can annotate and how fast.

Discussion Questions

1. **When is disagreement a bug vs a feature?**
   Use TimeBank TLINKs and UCCA/AMR as examples where "multiple valid analyses" are plausible.
2. **What counts as "gold" when multiple answers are correct?**
   Natural Questions and TyDi QA are built around this tension.
3. **How do we prevent artifacts without making data collection impossible?**
   Contrast SNLI/MultiNLI artifacts with WinoGrande's bias-reduction step.
4. **Who decides what "helpful" or "harmless" means?**
   Use InstructGPT and HH-RLHF to discuss value judgments and governance.
5. **Should datasets come with "nutrition labels"?**
   Connect older corpora (CoNLL-2003, PTB) to modern datasheets.

Key Takeaways

1. **Annotation projects define what ML can learn** — schema, sampling, workflow, and QC choices become model inductive biases
2. **Four historical waves:** expert treebanks → layered semantics → crowd-driven tasks → LLM alignment data
3. **Agreement must match label structure** — κ for classification, span P/R for spans, Smatch for graphs
4. **High agreement $\neq$ no artifacts** — SNLI/MultiNLI show that consistent data can still be shortcut-prone
5. **Documentation is part of the dataset** — data statements, datasheets, and model cards exist because the field learned the hard way
6. **Ethics are data-design questions** — who labels, who is represented, who can access, and who might be harmed

Questions?

Questions & Discussion

✉ jinzhao@brandeis.edu