Rillor

METHODS NOTE

Writing annotation guidelines: definitions, worked examples and edge cases

How to write annotation guidelines from published practice: label definitions with inclusion and exclusion rules, worked examples, edge cases, pilot rounds and a versioned change log.



Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Annotation guidelines decide how evenly a label set is applied, and published practice shows what they contain, how they are tested and how often datasets leave them out [1, 26].

- Annotation guidelines are the written rules for deciding when each label applies. A 2024 audit of 591 dataset papers separates the annotation scheme, which defines the labels, from the guidelines, which describe how to decide between them [1].
- Each label needs a definition with inclusion and exclusion criteria, and examples of both. Published guideline formats share the same core parts: a label, a definition, what fits and what does not, and examples, and one format adds negative examples with the reason for each [6, 7].
- Examples in the instructions shape data that annotators write. In 14 benchmarks where annotators wrote the data, about 73% of instruction examples shared a few patterns, and those patterns appeared in about 61% of the collected data [11].
- Detailed guidelines raise agreement when one reading of a label is wanted. In a 2022 experiment on a binary label for 200 social media posts, a group given detailed guidelines reached 93.72% agreement, against 72.50% for a group given a short definition [16].
- Guidelines are tested in pilot rounds and revised until agreement meets the project's own target. The OntoNotes corpus project checked the agreement of each 50-sentence sample at once, and scores below 90% led it to revise how it grouped each word's senses [3].
- Guidelines change during a project, so each version and each change needs a record. Universal Dependencies marks each notable change to its guidelines by type, such as an amendment or a clarification, and names the release it belongs to [22].
- The same guidelines can serve as a model's prompt, with results that vary by task. A 2023 study that replicated 27 annotation tasks with a commercial chat model found a median F1 of 0.707, ranging from 0.059 to 0.969, and recommends giving people and the model the same guidelines [24].
- Many papers that collect new labels say little about the instructions. In a 2020 review of machine learning papers on Twitter data, 47.30% of the 74 that used new human labels gave no information about the instructions annotators received [26].

Background

Labels are the answers a model learns from and is scored against. Before anyone applies them, someone writes down what each label means and how to choose between labels. In a 2024 audit of 591 papers introducing text datasets, Klie, Eckart de Castilho and Gurevych separate the two documents involved: the annotation scheme defines the labels, and the guidelines describe how to decide when to apply which label [1]. They call properly written guidelines essential for consistency and reproducibility, for example when a dataset is re-annotated or extended, and note that the way guidelines are written can itself introduce bias [1].

Social scientists call the document a codebook. Halterman and Keith write that codebooks, "documents that operationalize concepts and outline annotation procedures", are used almost universally by social scientists coding political texts [7]. Hovy and Lavid use the same word in corpus linguistics, call writing it "not a trivial task", and note that it is never complete before annotation starts [2].

Guidelines now have a second reader. Recent studies give the human codebook to language models as their prompt and compare the labels with human ones [7, 12, 24]. One study found a catch-all label so loosely defined that it was "a problem for humans and machines alike" [12].

This note is for annotation leads, engineers and dataset buyers. It covers what published guidelines contain, how definitions, examples and edge-case rules are written, how pilot rounds test them, how changes are recorded and what a dataset should publish. It ends with a template, an example entry and a checklist drawn from the sources.

Method

What was examined. Published annotation guidelines and the research on writing and testing them: structure and length, codebook entry formats, the effect of instructions and examples on collected data, prescriptive and descriptive annotation, pilot and revision cycles, change records, guidelines given to models as prompts, and what dataset papers report about their instructions.

Sources. Thirty sources, each opened on 8 November 2026 and read for the passages cited: 9 journal and magazine articles (one read in its arXiv version), 7 peer-reviewed conference papers (two read in their arXiv versions), 3 preprints, 5 published annotation guidelines, 3 pages of the Universal Dependencies guidelines, documentation from a cloud labeling service and a labeling tool, and a dataset card template.

How. Each practice is described from the source that sets it out, and each figure is as the source prints it. Page counts and the Universal Dependencies change counts are marked "derived": they come from the PDFs and from counting the project's change table on 8 November 2026 [8, 13, 19, 20, 22]. Model results are kept to the tasks, models and dates each study tested, and models are described by family. No guideline or tool is ranked.

Where sources disagree. Hovy and Lavid give the Penn Treebank codebook as three hundred pages [2], and the bracketing guidelines PDF has 318 (derived) [19]. The audit's 22% is two different counts, kept apart here [1]. Bayerl and Paul report percentage agreement, and Röttger and colleagues add Fleiss' kappa, so their figures are not compared with each other [5, 16].

Findings

1. What published guidelines contain and how long they run

The parts. In Hovy and Lavid's account, the second step of an annotation project specifies "the set of tags to use, their conditions of applicability", and begins the annotator instructions, "often called the Codebook or Manual" [2]. The guidelines examined here share these parts.

Part	What it holds	Published example	Source
Label set	The labels and how they nest	ACE entity types with subtypes, entity classes and mention levels	[8]
Definitions	When each label applies	OntoNotes sense groups, each with a definition, a syntactic frame and examples	[3]
Decision rules	Which label wins when two fit	ACE "trumping rules" for organization subtypes, with exceptions	[8]
Worked examples	A definition applied to real cases	The rater guidelines' examples	[13]
Flags	Special cases with their own rule	"Did Not Load"	[13]
Formal specification	The markup for structured annotation	TimeML's "How to annotate" section, BNF and attributes for each tag	[20]
Change record	What changed and when	The Universal Dependencies change table, the rater guidelines' change log	[13, 22]

The length.

Guideline	Version and date	Pages (derived)	Source
Penn Treebank bracketing guidelines, Treebank II style	January 1995	318	[19]
ACE English annotation guidelines for entities	Version 6.6, 13 June 2008	72	[8]
TimeML annotation guidelines	Version 1.2.1, 31 January 2006	71	[20]

Guideline	Version and date	Pages (derived)	Source
A search engine's published rater guidelines	11 September 2025	182	[13]

Guidelines for complex projects "are often quite detailed and span many pages", while in crowdsourcing they usually fit into the annotation screen [1]. One cloud labeling service's documentation splits short instructions on the task page from full instructions in a dialog box, and recommends "multiple examples showing edge cases and other difficult situations" [14]. It says the short instructions should take less time to understand than the task takes to complete, and that full instructions should supplement them without repeating them [14]. One labeling tool's documentation lets a project show its instructions in a pop-up when users enter the labeling stream [15].

The number of labels. Bayerl and Paul's meta-analysis covered 96 annotation studies in three domains and 346 agreement indices [5]. Agreement "decreased with a rising number of categories in the coding scheme" ($\beta = -0.28$, $p < 0.001$), in all three domains [5]. They conclude that schemes "should be as sparse as possible, if the reliability of annotations has to be high", and that fine-grained schemes, whose categories differ only slightly in meaning, are especially problematic [5]. Roberts and colleagues note that the percentage of agreement falls as the number of codes rises [6].

2. Definitions with inclusion and exclusion criteria

The entry format. Roberts, Dowell and Nie wrote each code with a label, a definition, a description, qualifications or exclusions, and examples from the raw data, following Boyatzis [6]. Halterman and Keith propose a machine-readable format: the exact label, a definition that is "generally a single sentence", a clarification and negative clarification covering extra inclusion criteria and "what does not fit into the class", and optional positive and negative examples with explanations [7]. They suggest writing codebooks from the outset to be both human- and machine-readable [7].

Part of the entry	Qualitative codebook [6]	Machine-readable codebook [7]
Name	Label	Label
Meaning	Definition and description	One-sentence definition
Boundaries	Qualifications or exclusions	Clarification and negative clarification
Cases	Examples from the raw data	Positive and negative examples, with explanations

Exclusion rules. The MUC-7 named entity task requires "a single, unambiguous output for any relevant string", and writes its boundaries as exclusions with examples [9]:

- Generic entity names such as "the police" and "the government" are not tagged.
- Titles such as "Mr." and role names such as "President" are not part of a person name.
- Political groups, laws named after people, diseases and prizes named after people, and saints are not PERSON, since removing a saint's title leaves a non-unique name.
- Adjectival forms of location names, as in "American exporters", are not LOCATION.

OntoNotes left non-specific references such as "officials" out of coreference, "since coreference for them is frequently unclear" [3].

Precedence and context. Where an organization fits more than one category, the ACE guidelines assign the most specific type possible, then list exceptions: a country's military organizations are Government, media conglomerates Commercial and the media organizations they own Media [8]. Where context decides, MUC-7 shows both readings: "congress" may name an event or an organization, as in "member of the 91st Congress" and "commencement of the 91st Congress" [9].

Testing an exclusion. Halterman and Keith add an exclusion criterion tied to a trigger word to both a document and the codebook. A model should reject the label only when the trigger is in the document and the criterion is in the codebook [7].

What a definition leaves in the data. Writers of natural language inference examples were given three definitions for a premise p and hypothesis h: entailment, where h is "definitely true given p", neutral, where h "might be true", and contradiction, where h is "definitely not true" [10]. A simple classifier reading only the hypothesis labeled about 67% of SNLI and 53% of MultiNLI correctly, through artifacts Gururangan and colleagues trace to the writers' strategies, such as negation for contradiction [10].

Where a label name is not enough. Sainz and colleagues found labels such as MEDIA, TRAILER and TASK "inherently polysemous", hard to apply from the name alone [12]. For one movie dataset's Trailer label, F1 rose from 0.0 without guidelines to 76.4 with them [12].

3. Worked examples and instruction bias

What examples do. The rater guidelines call their examples "very important to illustrate how the concepts defined in each section should be applied", and warn that content and websites may change after an example is added [13]. Their change log records examples added for clarity in September 2025 and outdated examples removed in January 2025 [13]. OntoNotes grouped the 22 WordNet 2.1 senses of "drive", each group with a definition, a syntactic frame and examples [3]. Geiger and colleagues write that formal definitions and concrete examples help annotators determine edge cases, and that without them there could be misunderstandings that agreement metrics do not capture, if all annotators share them [26].

Negative and near-miss examples. In Halterman and Keith's codebook excerpt, each negative example ends with the reason it belongs to the other class [7]. The ACE guidelines tag fictional characters' names, but not a character name used as a TV show title when it refers to the show [8]. They list "Morris the cat" among names that yield no mention, since animal names do not refer to person entities [8].

Instruction bias. Parmar, Mishra, Geva and Baral studied 14 benchmarks, most of them question answering datasets, in which annotators wrote the examples [11]. The authors hypothesize, and their counts support, that annotators "pick up on patterns in the crowdsourcing instructions" and write many similar examples [11].

Measure (2023)	Figure	Source
Instruction examples sharing a few clear patterns, average	About 73%	[11]
Collected instances showing those patterns, average	About 61%	[11]
QUOREF instruction examples starting "What is the name"	About 36%	[11]
QUOREF collected instances with the same start	About 59%	[11]
Gap when models trained only on pattern examples were tested on the rest	Up to 58%, in DROP	[11]
Score advantage on pattern examples, base and large models	About 7% and 3%	[11]

The authors ask whether models then learn "the dataset creator's instructions, rather than the task" [11]. They recommend diverse instructions, such as many instructive examples or examples sampled from previously collected data, with each annotator shown a different random sample, and analysis of word patterns during collection [11].

Order and checks. Every model in Halterman and Keith's behavioral tests was sensitive to the order in which the codebook described its categories [7]. The audit suggests checking understanding with task examples from the guidelines, and with unseen tasks as the harder test [1]. In a study with 40 participants, ordering instances by a simple heuristic cut total annotation time while preserving quality [29].

4. Edge cases and decision rules

Pilot data. The audit advises that pilot data "contain as many corner cases and difficult instances as possible", which reduces the chance of costly re-annotation later when changes are not backward compatible [1]. It adds that "Improving the annotation guidelines to incorporate edge cases should therefore be preferred over filtering" [1].

Rules for doubt. The rater guidelines give a tie-break for page quality ratings: raters unsure between two ratings use the lower one, and raters torn between three choose the middle [13]. They define flags for special cases, such as "Did Not Load" for pages with "absolutely no content on the page created by the website" [13].

What prespecification cannot do. Hovy and Lavid write that precise instructions for every conceivable case would give 100% agreement and remove the need for annotation, "since one can simply encode the instructions as a computer program", and that fully prespecifying any nontrivial phenomenon is "patently impossible" [2]. Röttger and colleagues add that inevitable gaps leave annotators to apply personal judgement, and that long, complex guidelines may need extra training [16].

When a distinction does not hold. Low agreement between similar categories hints that annotators cannot tell them apart, and one remedy is to collapse them into a single label [1]. Hovy and Lavid describe "neutering" the theory: when categories are hard to define, annotation can go ahead with a simpler, more neutral set [2].

When labels and guidelines disagree. Sainz and colleagues' model labeled years as TIME entities in one dataset, correctly under its guidelines, while the dataset itself does not label years [12]. The mismatch surfaced in their error analysis [12].

Feedback. The audit stresses feedback when annotators find the guidelines hard to understand or find errors in them [1]. Of 298 papers with human annotators, 18 (6%) reported annotator feedback during annotation [1].

5. Prescriptive and descriptive paradigms

Two aims. Röttger, Vidgen, Hovy and Pierrehumbert define two paradigms for subjective tasks. The descriptive paradigm encourages annotator subjectivity, so different beliefs can be surveyed and modeled. The prescriptive paradigm discourages it, so models can be trained to apply one belief consistently [16]. Well-formulated guidelines, they write, give clear instructions on how to decide between classes, with explanations and examples, which creates accountability because people can review and challenge the stated belief [16]. Under the prescriptive paradigm, disagreements are "a call to action": the guidelines were either applied incorrectly or inadequate [16].

The experiment. Sixty annotators were randomly assigned to three groups of 20, and each group gave the same 200 social media posts a binary label [16].

Group (2022)	Instructions	Percent agreement	Fleiss' kappa
G1	Descriptive prompt	73.90%	0.20
G2	Prescriptive prompt with detailed guidelines	93.72%	0.78
G3	Prescriptive prompt with a short definition only	72.50%	0.15

Source: [16]. Agreement was significantly higher in G2 ($p < 0.001$), 19.82 points above G1 (derived) [16]. The authors conclude that comprehensive guidelines are "instrumental in facilitating high agreement in the prescriptive paradigm", and that dataset creators should aim explicitly for one paradigm or the other [16].

The cost of precision. Aroyo and Welty list "Detailed Guidelines Help" among seven myths of human annotation [17]. They found that increasingly precise guidelines "do eliminate disagreement but do not increase quality", by forcing annotators into choices they may not think valid [17]. Uma and colleagues report human disagreement from part-of-speech tagging to subjective tasks, and hold that a consensus on how to evaluate models remains essential even without a gold standard [18]. Hovy and Lavid conclude that "the intended use of the annotated corpus is probably the single most important guideline" [2].

6. Pilot rounds, training and agreement as a test of the guidelines

The cycle. Hovy and Lavid set out the steps: specify the tags and begin the codebook, annotate a fragment to test both, and if agreement is not satisfactory, repeat with changes to the theory, the manual and the instructions [2]. Their generic pipeline takes 90% agreement as the acceptability threshold [2]. OntoNotes annotated a 50-sentence sample and checked its agreement at once, and scores below 90% led to "a revision and clarification of the groupings" [3]. The project reports higher accuracy and "at least a three-fold increase in annotator productivity" [3].

Thresholds are a project's own. OntoNotes based its 90% rule on the observation that when people agree at N%, systems reach (N − 10)%, and it considered 80% realistic for word senses [2]. For many complex problems, Hovy and Lavid add, 90% is "simply unreachable" [2]. Bayerl and Paul ask authors who stop training at 80% agreement to say so, and do not advocate identical thresholds for every purpose [5].

Training. In the meta-analysis, trained annotators reached 81% average agreement and untrained ones 70%, and high-intensity training gave 86% against 75% for low intensity and 70% for none [5]. Repeating a sample until agreement is stable risks overtraining, but shows that the codebook is clear for the variants met so far [2].

Consensus sessions. In Deleger and colleagues' medical corpus, every document was labeled twice and disagreements were resolved in consensus sessions [4]. During the training period, agreement was computed at the end of each day, a session was held and the guidelines gained clarified rules and examples [4]. Once agreement was high, the guidelines were mostly stable and sessions were held every three to four days [4].

How often pilots happen. Of the audit's 298 papers with human annotators, 68 (22%) reported an iterative refinement loop, mainly to refine guidelines after pilot studies (10%) or to correct instances until they reached sufficient quality (12%) [1]. Overall, 22% of the publications mentioned a pilot study [1]. The audit also suggests publishing each annotator's raw annotations alongside the adjudicated corpus [1].

7. Versions and change records

Why guidelines change. Hovy and Lavid write that the codebook "is invariably developed and perfected during the annotation process", which poses problems for annotation already completed, and that later changes should be made "always ensuring that no earlier decisions are superseded" [2]. Updating the guidelines may require discarding earlier annotations or at least reviewing them [1]. Fernald and Duclos advise teams to write down code definitions, revisit them and update them as the analysis progresses, to catch inconsistencies and changes in meaning [30].

Changes in practice. In one published study that trained a model family on human preference rankings, the labeling instructions evolved as the authors gave feedback, changed metadata fields and understood better what they wanted to measure, and confusing or inconsistent instructions were amended [25]. The priority order changed as well: helpfulness came first while training data was labeled, and truthfulness and harmlessness came first in the final evaluations [25].

A change table. Universal Dependencies launched its v2 guidelines on 1 December 2016, with later revisions [21]. Their change history classes each notable change by type [22].

Type	Definition in the change table	Count, December 2021 to June 2026 (derived)
Amendment	A change that negates a policy of an earlier version	7, decided January 2022 to March 2023
Clarification	Resolves an important ambiguity without negating a policy	10, some also validator changes
Validator	An important change to what the official validator checks	1 on its own

Source: [22]. The change table assigns the 18 changes to releases 2.10 to 2.19 [22]. Updates to v2 keep the inventory of basic top-level dependency relations and part-of-speech tags, and some clarify how they apply to particular phenomena [22]. The move from v1 to v2 came with an executive summary written as a checklist for treebank developers, recording changes such as renaming the tag CONJ to CCONJ [23].

Versions and logs. ACE, MUC-7 and TimeML publish numbered, dated versions [8, 9, 20]. The rater guidelines are "updated regularly to remain relevant over time", with two years of changes in an appendix [13].

Tools fix the version. One cloud labeling service's documentation states that worker instructions cannot be modified once a labeling job is created, and that changing them means stopping the job, cloning it and editing the copy [14].

8. Guidelines given to models as prompts

The same codebook for both. Pangakis, Wolken and Fasching replicated 27 annotation tasks across 11 datasets with a commercial chat model [24]. Their workflow gives human coders and the model the same codebook, as the model's prompt, since different codebooks could leave "a conceptual gap" [24]. If the model does poorly on a first human-labeled subset, the codebook is refined around the incorrect classifications and then tested on the remaining human-labeled samples [24]. From their analyses, they put the number of human-labeled samples between 250 and 1,250 random texts [24].

Result across 27 tasks (2023)	Figure	Source
Median F1	0.707	[24]
Lowest and highest F1	0.059 and 0.969	[24]
Median accuracy	0.85	[24]

Performance varied with "prompt quality, text data idiosyncrasies, and conceptual difficulty", and the authors call a clear codebook essential [24].

Guidelines in the prompt. Sainz and colleagues trained a model to follow annotation guidelines given in its input [12]. In their ablation, mean zero-shot F1 was 55.3 with every component, 49.9 without the example candidates drawn from the guidelines and 42.3 for the baseline with all components removed [12]. The full model scored 13.0 points above the baseline, and removing only the candidates cost 5.4 (derived) [12]. The authors find definitions and representative candidates complementary, and detailed guidelines key for good results [12].

Codebooks with open-weight models. Halterman and Keith tested open-weight models of 7 to 12 billion parameters on political science codebooks [7]. Zero-shot weighted F1 ranged from 0.21 on one dataset to 0.65 and 0.57 on two others [7]. In an ablation on one model and one codebook, including all codebook components improved performance, and supervised instruction tuning raised that 7-billion-parameter model from 0.53 to 0.82 F1 on the same dataset [7].

9. What dataset papers report about their guidelines

One field's record. Geiger and colleagues labeled 164 papers on machine learning with Twitter data, 142 of them original classification tasks and 74 with original human annotation [26].

What the paper reported about instructions (2020)	Papers	Share
No information	35	47.30%

What the paper reported about instructions (2020)	Papers	Share
Instructions with formal definitions or examples	32	43.24%
No instructions beyond the question text	7	9.46%

Source: [26].

Across text datasets. In the 2024 audit, 60% of publications with manual annotation described their annotation scheme, showed their interface or published their guidelines with the dataset in some form [1].

Documentation templates. One dataset hub's dataset card template asks for the annotation process: tools, the amount of data annotated, the guidelines provided, agreement statistics and validation [27]. Datasheets for Datasets asks creators to read its questions before labeling and answer them afterward, including a description of any labeling, and asks whether the dataset is updated to correct labeling errors [28]. It does not name annotation guidelines as a field [28].

Implications

A guideline template, derived from the sources.

1. Purpose: the intended use of the data and whether the guidelines are prescriptive or descriptive [2, 16].
2. Label set: as small as the intended use allows [5, 6].
3. One entry per label: the label, a one-sentence definition, inclusion criteria, exclusion criteria, positive examples, and negative examples that each state why they belong elsewhere [6, 7, 9].
4. Decision rules: precedence between labels, rules that depend on context and tie-breaks for doubt [8, 9, 13].
5. Edge cases and flags [1, 13, 14].
6. Short instructions for the task screen, with the full guidelines behind them [14, 15].
7. A version number, a date and a change log that classes each change as an amendment or a clarification [13, 22].

An example entry built from published rules. This entry fills section 3 with rules from two published entity guidelines. The definition is left for the project to write.

Part	Content	Source
Label	PERSON	[9]
Definition	One sentence, written for the project	[7]
Include	Names of fictional characters	[8]
Exclude from the name	Titles such as "Mr." and role names such as "President"	[9]
Exclude from the label	Political groups, laws named after people, diseases and prizes named after people, saints	[9]
Near misses	A character name used as a TV show title, when it refers to the show. "Morris the cat", since animal names are not tagged	[8]

A checklist before labeling starts.

1. The intended use and the paradigm open the guidelines [2, 16].
2. Every label has inclusion and exclusion criteria, and negative examples with their reasons [6, 7].

3. Where annotators write data, the example pool is varied and each annotator sees a different sample [11].
4. The pilot sample holds as many corner cases and difficult instances as possible [1].
5. The agreement target and its basis are written down, and the pilot repeats until the target is met or the label set is simplified [1, 2, 3, 5].
6. Annotators are checked on examples from the guidelines and on unseen items [1].
7. Each version has a number, a date and a change log, and each label can be traced to the version it followed, so labels made under an earlier version can be found and reviewed [1, 2, 22].
8. Where a model labels, its prompt is the same codebook, refined on one human-labeled subset and tested on another [24].
9. The dataset ships with its guidelines, its raw annotations and its agreement statistics [1, 27].

For engineers. Store the guideline version beside each label, and for model labels the prompt and the model as well, since instructions change during a project [24, 25].

For dataset buyers. Ask for the guidelines, their version and their change log with the data.

Limits of this analysis

- Most sources come from language processing and social science, and every example is a text task.
- The instruction bias evidence comes from tasks where annotators write examples, mostly question answering [11]. It does not measure the effect when annotators choose labels.
- The paradigm experiment covers one binary label on 200 posts [16], and the meta-analysis reports percentage agreement in three domains [5].
- Model results belong to the models, tasks and dates each study tested [7, 12, 24].
- The 2020 review covers machine learning papers on Twitter data [26].
- Page counts include front matter, and the Universal Dependencies counts were taken on 8 November 2026 [19, 22]. Documentation pages can change after the viewing date [14, 15].
- No source reviewed sets a required guideline length, number of examples per label or number of pilot rounds, and agreement thresholds here are each project's own [2, 3, 5].

Sources

1. Analyzing Dataset Annotation Quality Management in the Wild. Jan-Christoph Klie, Richard Eckart de Castilho and Iryna Gurevych, Computational Linguistics 50(3), pp. 817-866, September 2024, viewed 8 November 2026. https://aclanthology.org/2024.cl-3.1/

2. Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. Eduard Hovy and Julia Lavid, International Journal of Translation 22(1), January-June 2010, author copy, viewed 8 November 2026. https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf

3. OntoNotes: The 90% Solution. Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel, Proceedings of HLT-NAACL 2006, Companion Volume: Short Papers, pp. 57-60, Association for Computational Linguistics, June 2006, viewed 8 November 2026. https://aclanthology.org/N06-2015/

4. Building Gold Standard Corpora for Medical Natural Language Processing Tasks. Louise Deleger, Qi Li, Todd Lingren and six others, AMIA Annual Symposium Proceedings 2012, pp. 144-153, 3 November 2012, viewed 8 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC3540456/

5. What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. Petra Saskia Bayerl and Karsten Ingmar Paul, Computational Linguistics 37(4), pp. 699-725, December 2011, viewed 8 November 2026. https://aclanthology.org/J11-4004/

6. Attempting rigour and replicability in thematic analysis of qualitative research data: a case study of codebook development. Kate Roberts, Anthony Dowell and Jing-Bao Nie, BMC Medical Research Methodology 19, article 66, 28 March 2019, viewed 8 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC6437927/

7. Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts. Andrew Halterman and Katherine A. Keith, preprint, arXiv 2407.10747, version 2, 9 January 2025, viewed 8 November 2026. https://arxiv.org/abs/2407.10747

8. ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6. Linguistic Data Consortium, 13 June 2008, viewed 8 November 2026. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

9. MUC-7 Named Entity Task Definition, Version 3.5. Nancy Chinchor, National Institute of Standards and Technology, 17 September 1997, viewed 8 November 2026. https://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html

10. Annotation Artifacts in Natural Language Inference Data. Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman and Noah A. Smith, Proceedings of NAACL-HLT 2018, Volume 2, pp. 107-112, Association for Computational Linguistics, June 2018, viewed 8 November 2026. https://aclanthology.org/N18-2017/

11. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. Mihir Parmar, Swaroop Mishra, Mor Geva and Chitta Baral, Proceedings of EACL 2023, pp. 1779-1789, Association for Computational Linguistics, May 2023, viewed 8 November 2026. https://aclanthology.org/2023.eacl-main.130/

12. GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction. Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau and Eneko Agirre, ICLR 2024, arXiv 2310.03668, version 5, 6 March 2024,

viewed 8 November 2026. https://arxiv.org/abs/2310.03668

13. General Guidelines (Search Quality Evaluator Guidelines). Google, 11 September 2025, viewed 8 November 2026. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf

14. Create instruction pages. Amazon SageMaker Developer Guide, Amazon Web Services, undated page, viewed 8 November 2026. https://docs.aws.amazon.com/sagemaker/latest/dg/sms-creating-instruction-pages.html

15. Set up your labeling project. Label Studio documentation, undated page, viewed 8 November 2026. https://labelstud.io/guide/setup_project

16. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. Paul Röttger, Bertie Vidgen, Dirk Hovy and Janet B. Pierrehumbert, Proceedings of NAACL-HLT 2022, pp. 175-190, Association for Computational Linguistics, July 2022, viewed 8 November 2026. https://aclanthology.org/2022.naacl-main.13/

17. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. Lora Aroyo and Chris Welty, AI Magazine 36(1), pp. 15-24, 25 March 2015, viewed 8 November 2026. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564

18. Learning from Disagreement: A Survey. Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank and Massimo Poesio, Journal of Artificial Intelligence Research 72, pp. 1385-1470, 27 December 2021, viewed 8 November 2026. https://www.jair.org/index.php/jair/article/view/12752

19. Bracketing Guidelines for Treebank II Style, Penn Treebank Project. Ann Bies, Mark Ferguson, Karen Katz and Robert MacIntyre, Linguistic Data Consortium, January 1995, viewed 8 November 2026. https://catalog.ldc.upenn.edu/docs/LDC99T42/prsguid1.pdf

20. TimeML Annotation Guidelines, Version 1.2.1. Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer and James Pustejovsky, 31 January 2006, viewed 8 November 2026. https://timeml.github.io/site/publications/timeMLdocs/annguide_1.2.1.pdf

21. UD Guidelines. Universal Dependencies, undated page, viewed 8 November 2026. https://universaldependencies.org/guidelines.html

22. Guidelines Changes. Universal Dependencies, undated page, viewed 8 November 2026. https://universaldependencies.org/changes.html

23. Executive summary of changes from v1 to v2. Universal Dependencies, undated page, viewed 8 November 2026. https://universaldependencies.org/v2/summary.html

24. Automated Annotation with Generative AI Requires Validation. Nicholas Pangakis, Samuel Wolken and Neil Fasching, preprint, arXiv 2306.00176, version 1, 31 May 2023, viewed 8 November 2026. https://arxiv.org/abs/2306.00176

25. Training language models to follow instructions with human feedback. Long Ouyang and 19 others, preprint, arXiv 2203.02155, version 1, 4 March 2022, viewed 8 November 2026. https://arxiv.org/abs/2203.02155

26. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang and Jenny Huang, Conference on Fairness, Accountability, and Transparency (FAT* '20), January 2020, arXiv 1912.08320, viewed 8 November 2026. https://arxiv.org/abs/1912.08320

27. Dataset card template (datasetcard_template.md). Hugging Face, huggingface_hub repository, last changed 18 January 2024, viewed 8 November 2026. https://github.com/huggingface/huggingface_hub/blob/main/src/huggingface_hub/templates/datasetcard_template.md

28. Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford, Communications of the ACM, December 2021, arXiv 1803.09010, version 8, 1 December 2021, viewed 8 November 2026. https://arxiv.org/abs/1803.09010

29. Annotation Curricula to Implicitly Train Non-Expert Annotators. Ji-Ung Lee, Jan-Christoph Klie and Iryna Gurevych, Computational Linguistics 48(2), pp. 343-373, June 2022, viewed 8 November 2026. https://aclanthology.org/2022.cl-2.4/

30. Enhance Your Team-Based Qualitative Research. Douglas H. Fernald and Christine W. Duclos, Annals of Family Medicine 3(4), pp. 360-364, July/August 2005, viewed 8 November 2026. https://www.annfammed.org/content/annalsfm/3/4/360.full.pdf

About this report

Published by Rillor on 9 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Writing annotation guidelines: definitions, worked examples and edge cases. 9 November 2026. https://rillor.com/insights/writing-annotation-guidelines

Work with Rillor

Send a sample of the records, the labels your schema sets and the accuracy you need. Contact us at rillor.com/contact?type=labeling or contact@rillor.com. More on labeling and annotation: rillor.com/data#labeling