Twenty-Five Years of Inter-Annotator Agreement in NLP: A Systematic Survey of Trends, Metrics, and Reporting Practices

Hend Al-Khalifa
College of Computer and Information Sciences
King Saud University, Riyadh, Saudi Arabia
hendk@ksu.edu.sa

Abstract

Large language models are now deployed as annotators across NLP, generating training labels and evaluation scores, yet the field has no established standard for verifying whether LLM-generated labels are reliable, internally consistent, or comparable across studies. This methodological gap does not begin with LLMs: a systematic corpus study of **665 ACL Anthology papers** from 1999 to 2026 reveals that even classical human annotation in NLP suffers from systematic measurement and reporting failures that have left a growing body of work scientifically incomparable. Papers were identified through keyword search and organised via topic modelling into a thematic taxonomy spanning eight annotation domains from named entity recognition to disagreement modelling. Full-text analysis of 549 papers yields three principal findings. First, Cohen's κ dominates across all task types (259 papers, 38.9%), regardless of whether its assumptions are satisfied; ICC, appropriate for continuous-scale tasks such as MT evaluation, appears in only 16 full-text papers (2.4%). Second, **226 of the 665 papers (34.0%)** name no IAA coefficient anywhere in the full text, which is a methodology-reporting concern; a further 50.1% of papers name a coefficient in the body but not in the title or abstract, so **84.1%** of papers omit the metric from the abstract layer. We frame the body-only gap as a *discoverability* concern (the methodology is reported but is not visible at the abstract layer readers and reviewers encounter first), distinct from the full-text omission. Third, annotator disagreement modelling and LLM-as-annotator constitute the fastest-growing area of the corpus (124 papers, 18.6%): 26 of these appeared between 1999 and 2022 and 98 between 2023 and 2026, the only dimension whose volume rises in that window. Fewer than 5% of papers report confidence intervals. We provide a large-scale empirical analysis of metric reporting practice within the ACL Anthology and a seven-point agenda for standardised reporting, complementing prior normative guidance (James, 2026) and the cross-disciplinary meta-analysis of agreement determinants by Bayerl and Paul (2011) with field-scale descriptive evidence.

1 Introduction

High-quality annotated data underpin supervised approaches in NLP, and their reliability is assessed by *inter-annotator agreement* (IAA): the extent to which independent annotators assign the same label to the same item. High IAA signals unambiguous guidelines and a well-defined task; low IAA may indicate under-specification, linguistic ambiguity, or subjective variation. Yet IAA is applied inconsistently: studies report raw percentage agreement, Cohen's κ (Cohen, 1960), Krippendorff's α (Krippendorff, 2013), Fleiss' κ (Fleiss, 1971), span-level F1, or no metric at all, and metric choice is rarely justified against the assumptions of each coefficient (raters, scale type, missing data, class balance). Crowdsourcing and LLM-generated labels further strain the classical two-annotator paradigm. Recent work gives prescriptive guidance on *which* metric to choose (James, 2026; Artstein and Poesio, 2008); our contribution is complementary and empirical, documenting *how* the community measures and reports IAA. Bayerl and Paul (2011) meta-analysed 587 studies to model the *determinants* of agreement levels (task difficulty, annotator training, category count) across disciplines. We ask a different question of a different population: which coefficients ACL Anthology authors name, whether those choices suit their data, and what they report alongside them. Ours is, to our knowledge, the first large-scale analysis of IAA *reporting practice* within the ACL Anthology, and the first to quantify metric–task mismatch at corpus scale.

The cost of the gap. Among the 439 post-2019 annotation papers in our corpus, only **3.6%** provide IAA evidence that is simultaneously present

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 579–602
November 6-10, 2026 ©2026 Association for Computational Linguistics

(a metric is named), task-appropriate (correct for the scale and rater configuration), and statistically interpretable (a confidence interval is given); the other 96.4% miss at least one property. The ACL Anthology grows by roughly 8,000–10,000 papers per year;[1] at a conservative 5% annotation rate, roughly 400–500 annotation papers appear annually, of which only about **16** carry IAA evidence comparable across studies. This accumulation is accelerating: disagreement modelling and LLM-as-annotator form the fastest-growing dimension of our corpus (124 papers, 18.6%), rising from 26 papers over 1999–2022 to 98 over 2023–2026, and inherit these failures at machine scale.

Contributions: (*i*) a systematic corpus of 665 papers from the ACL Anthology covering IAA across the NLP sub-fields the Anthology indexes (§3); (*ii*) an eight-dimension taxonomy mapping the literature onto dedicated survey sections (§4–§11); (*iii*) empirical analysis of metric reporting frequency, revealing Cohen's κ dominance and widespread reporting omissions (§4); (*iv*) a seven-point research agenda (§12.4).

Relation to concurrent normative guidance. James (2026) prescribes which coefficient to choose for a given task; the present paper is descriptive, documenting what the community does and quantifying the gap, so the two together supply the normative framework and the empirical baseline needed to reform IAA reporting.

2 Background

We summarise IAA metrics across three data regimes; Appendix E tabulates them by data type, rater requirements, missing-data support, and chance correction, and James (2026) and Artstein and Poesio (2008) give full practitioner guidance.

Categorical metrics. *Cohen's κ* (Cohen, 1960) chance-corrects observed agreement but is limited to two annotators and sensitive to marginal imbalance, the "κ paradox" (Gwet, 2001). *Weighted κ* (Cohen, 1968) extends it to ordinal scales; *Fleiss' κ* (Fleiss, 1971) generalises to $N \geq 3$ raters; *Krippendorff's α* (Krippendorff, 2013) handles any number of raters, missing data, and arbitrary scale types, subsuming Cohen's κ. *Gwet's AC1/AC2* (Gwet, 2001) addresses the paradox directly, and Scott's π (Scott, 1955) uses pooled marginals. Di Eugenio and Glass (2004) showed early that κ values are not comparable across studies without reporting the marginal distributions and the bias and prevalence they induce, a caveat the reporting checklist of §12.3 operationalises.

Structured and continuous metrics. Span F1 scores one annotator's spans against the other but, unlike κ-family coefficients, applies no chance correction; P_k and WindowDiff (Pevzner and Hearst, 2002) measure segmentation agreement; the γ *measure* (Mathet et al., 2015) jointly optimises boundary alignment and label agreement for unitising tasks. For continuous data, the Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss, 1979) partitions variance across rater-effect models, while Spearman ρ, Kendall τ, and Pearson r measure associative consistency rather than absolute agreement. Interpretation thresholds such as Landis and Koch (1977) are heuristic, and their application to NLP has been criticised (Artstein and Poesio, 2008; James, 2026).

3 Methodology

Search and selection (PRISMA-style flow). We searched the ACL Anthology (retrieved January 2026) by keyword matching over title and abstract metadata covering IAA terminology (*inter-annotator agreement*, *inter-rater reliability*, *annotator disagreement*, *Krippendorff alpha*, *LLM annotator agreement*, and related forms); the resulting paper list is released with the corpus, and the 61 metric-extraction regexes applied downstream are disclosed verbatim in Appendix B. Of 121,832 indexed papers, 736 matched at least one pattern; 1 was excluded as off-domain and 70 because IAA appeared in a single sentence without further analysis, yielding **665 papers**. Included papers either propose or compare IAA metrics, report and *analyse* IAA for a novel scheme, or study annotator disagreement as a substantive phenomenon.

Inclusion vs. metric naming. Inclusion asks *does the paper discuss annotation agreement?* (from IAA vocabulary in title and abstract); the metric-frequency analysis of §4 asks the independent question *does it name the specific coefficient used?* A paper on IAA that never names Cohen's κ or Krippendorff's α still belongs in the corpus; that 84.1% of included papers fall into this category is a central finding of this survey.

[1] Estimated from ACL Anthology volume metadata (https://aclanthology.org): the Anthology indexed 121,832 papers as of our January 2026 retrieval, and annual additions have exceeded 8,000 in recent years as the field's publication volume has accelerated.

Survey section / theme	*N*	%
§4 IAA metrics & reproducibility	47	7.1
§5 Corpus construction & guidelines	104	15.6
§6 Sequence labelling (NER, treebanks)	57	8.6
§7 Semantic annotation (disc., SRL, lex.)	130	19.5
§8 Temporal & event annotation	59	8.9
§9 Sentiment, opinion & hate speech	94	14.1
§10 Disagreement & LLM annotation	124	18.6
§11 MT evaluation	50	7.5
Total	665	100.0

Table 1: Corpus breakdown by thematic dimension.

Corpus overview. The corpus comprises **665 papers** (1999–2026) over 21 venues, with publication volume rising post-2019 (439 papers, 66%). Figure 1 shows the eight-section taxonomy and Table 1 the breakdown. By paper type, 81.2% are empirical studies reporting IAA for a novel scheme, 9.6% construct corpora, 5.7% study disagreement, 2.1% examine LLM-as-annotator, and only 0.6% propose new IAA metrics.

Clustering. We applied Non-negative Matrix Factorisation (NMF, k=16, seed 42) to TF-IDF representations of titles and abstracts. We chose NMF over density-based alternatives because it assigns every document to exactly one topic (89% confident at threshold 0.25, vs. BERTopic leaving 176 papers requiring keyword rescue) and recovers thematically distinct sub-fields that density-based clustering merges (e.g. sentiment vs. hate speech, NER vs. treebanks). The 16 topics were manually mapped to the eight survey dimensions of Figure 1, and the author verified all assignments by iterative review of topic keywords and representative titles.

Reliability of coding decisions. Most dimensions in Table 9 are purely algorithmic (year, venue, regex metric extraction) or read directly from paper metadata. Two require human judgment, *paper type* and *survey section*: an independent second coder, a PhD-level NLP researcher compensated at standard academic rates, coded a stratified 10% sample (50 papers) on both, with Cohen's $\kappa = 0.63$ (95% CI: 0.47–0.79) for paper type and 0.78 (0.66–0.88) for survey section, both indicating substantial agreement (Landis and Koch, 1977). The regex metric extractor, validated on a separate 50-paper sample, achieves precision 0.96, recall 0.91, and F1 0.93. Full per-category breakdowns, the disagreement analysis, and Table 7 appear in Appendix A; all bootstrap confidence intervals use 5,000 iterations.

Compliance audit. Beyond the corpus analysis, we run a secondary study on a 50-paper subset: four annotators (two human coders and two LLMs) score the most recent papers in the corpus against the reporting checklist this survey proposes. The audit measures whether the checklist can be applied consistently, and we report it in §12.3 once the checklist itself has been introduced.

Data and code availability. The 665-paper corpus with all extracted annotation flags, the derived summary tables, the four-annotator compliance-audit data, and the audit notebook are released at `https://github.com/HendSuliman/IAA`. We use ACL IDs rather than verbatim full text; all metadata derives from the ACL Anthology under its CC BY 4.0 licence.

4 IAA Metrics, Reproducibility, and Tooling

The 47 papers in this dimension address metric theory, software infrastructure, and reproducibility of IAA reporting (Mathet and Widlöcher, 2011b; Bregeon et al., 2019; Volodina et al., 2020; Popović and Belz, 2021; Stodden and Kallmeyer, 2022; Laarmann-Quante et al., 2022; Loáiciga et al., 2022; Savkov et al., 2022; Popović et al., 2022; Liu et al., 2023; Krishna et al., 2023; Groschwitz et al., 2023; Lee et al., 2023; Gao et al., 2023; Li et al., 2023).

Metric development and reproducibility. Mathet et al. (2015) introduced γ for unitising tasks, addressing a shortcoming of applying κ to span annotation. Reproducibility studies find published IAA figures hard to replicate due to under-specified preprocessing and disagreement-resolution procedures (Popović and Belz, 2021; Stodden and Kallmeyer, 2022; Gao et al., 2023), and Klie et al. (2024) notes that papers rarely document which library version computed κ, despite implementations differing in tie-breaking and yielding divergent results.

Tooling. WebAnno (Yimam et al., 2013) and INCEpTION (Klie et al., 2018) dominate recent corpus work with built-in κ/α dashboards, whereas BRAT (Stenetorp et al., 2012) computes no IAA natively. Python libraries (`nltk.metrics.agreement`, `krippendorff` (Castro, 2017), `pygamma-agreement` (Titeux and Riad, 2021)) provide reference implementations, but their versions are rarely cited. Crowdsourcing platforms

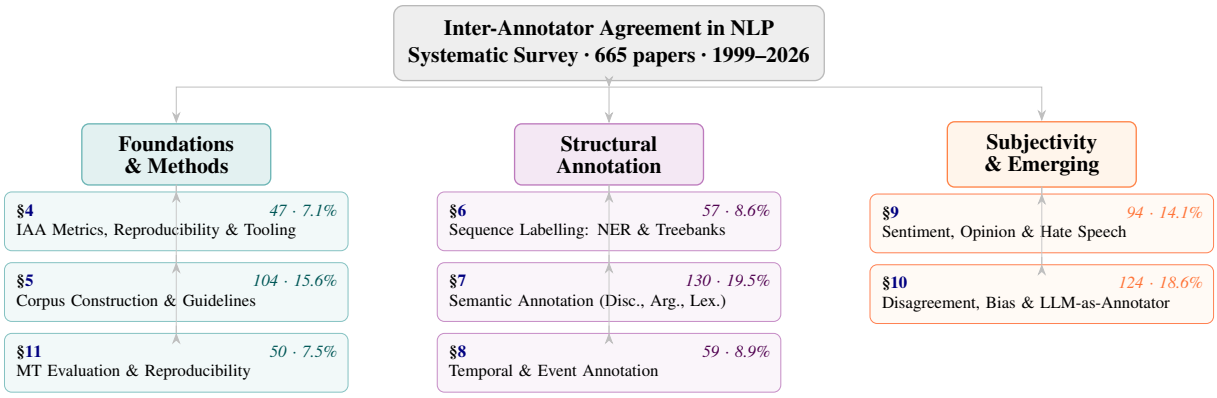


Figure 1: Taxonomy of the 665-paper corpus across eight survey dimensions, derived via NMF (k=16; TF-IDF, unigrams+bigrams). NMF was preferred over density-based methods because it assigns all documents without outliers (89% confident coverage) and naturally recovers finer distinctions that motivate separate survey sections: sentiment vs. hate speech, disagreement modelling vs. LLM-as-annotator, and NER vs. dependency treebanks. Three colour-coded branches group the eight sections thematically: **Foundations & Methods** (teal) covers metric theory, corpus construction, and MT evaluation; **Structural Annotation** (violet) covers sequence labelling, semantic annotation, and temporal relations; **Subjectivity & Emerging** (orange) covers sentiment/hate speech and the field's fastest-growing frontier, annotator disagreement and LLM-as-annotator. Each box shows the section number, paper count, and corpus share.

lack native IAA computation, and majority-vote thresholding is commonly substituted for formal agreement, a practice criticised by Artstein and Poesio (2008) and Klie et al. (2024).

Metric frequency and reporting norms. Table 2 shows named metric frequencies across the corpus. Cohen's κ is the most-cited metric at both scanning levels: 259 papers name it in full text (38.9% of the 665-paper corpus), though only 43 name it in the abstract (6.5%, 95% CI: 4.8–8.6%). Krippendorff's α appears in 157 full texts (23.6%) but only 36 abstracts (5.4%); Fleiss' κ in 114 full texts, 28 abstracts.

Temporal trend in reporting quality. Reporting has improved: the share of papers naming no coefficient in full text falls from 76.5% (1999–2009) to 58.3% (2010–2018) to **20.0%** (2019–2026; Cramér's $V = 0.42$). It varies by venue ($\chi^2(6) = 123.2$, $p < 0.001$): Findings papers are 3.6% unnamed against 62.6% for workshop papers. Abstract *length* does not explain the gap: a logistic regression predicting metric-naming from abstract word count and year leaves word count non-significant once year is controlled ($\beta = -0.20$ vs. year $\beta = 1.14$), so the venue effect reflects review-process compliance, not formatting constraints.

Across the corpus (full text parsed directly for 549 papers, and assessed from the abstract for the remaining 116 where no machine-readable PDF was available), **226 of the 665 papers (34.0%)** name no specific coefficient anywhere in the text examined. The abstract-level rate is **559 papers (84.1%)**; that is, 50.1% of authors name the metric in the body but not in the title or abstract. We frame the body-only gap as a *discoverability* concern: the methodology is reported, but it is not visible to automated indexing, meta-analytic screening, or citation-tracking tools that operate over titles and abstracts. We do not suggest that reviewers should assess reliability from abstracts; the full text is and remains the appropriate basis for that judgement. The gap is distinct from, and less severe than, the full-text omission: it concerns *where* a metric is reported rather than *whether* it is reported. Confidence intervals appear in only 29 of the 665 papers (4.4%) (Reichenheim, 2004; James, 2026). See Appendix F for the full metric × task matrix.

Naming versus reliability evidence. "No coefficient named" is a measurement-vocabulary signal, not a direct reliability-evidence signal: a paper that reports "annotators agreed on 87% of items" or "two coders cross-checked all annotations" provides reliability evidence without naming a coefficient. The 34.0% full-text figure should therefore be read as an *upper bound* on the share of papers that report no reliability evidence of any kind; an unknown subset of those papers describe agreement numerically or procedurally in language that a coefficient-name scan does not capture. A named, chance-corrected coefficient is the prerequisite for cross-study comparison, and its absence is what the 34.0% figure measures and what the reporting

	Abstract scan		Full-text scan†	
Metric	*N*	%	*N*	%
Cohen's κ	43	6.5	259	38.9
Krippendorff's α	36	5.4	157	23.6
Fleiss' κ	28	4.2	114	17.1
Percentage agreement	6	0.9	77	11.6
Span F1 (as IAA proxy)	4	0.6	36	5.4
CrowdTruth	1	0.2	37	5.6
MACE	0	0.0	27	4.1
Dawid–Skene	0	0.0	17	2.6
ICC	0	0.0	16	2.4
Gwet's AC1/AC2	0	0.0	16	2.4
Scott's π	0	0.0	16	2.4
Gamma (γ)	1	0.2	12	1.8
Correlation ($\rho/r/\tau$)	0	0.0	5	0.8
WindowDiff	0	0.0	4	0.6
IRT / Rasch	0	0.0	1	0.2
No coefficient named*	**559**	**84.1**	**226**	**34.0**

* **Full-text** (primary): papers naming no coefficient anywhere in the full paper (549 parsed, 116 fallback to abstract). **Abstract** (secondary): papers omitting the coefficient from title and abstract (what readers see without the full paper). The 2.5× gap (34.0% vs 84.1%) measures the abstract-level under-reporting burden.

† Full-text scan: 549/665 PDFs successfully downloaded and parsed (82.6%). The 116 fallbacks divide as follows: ∼34 pre-2010 papers are scanned image PDFs with no extractable text layer; ∼31 have no Anthology abstract and were assessed from title only; ∼6 are hosted outside the Anthology (JEPTAL-RECITAL, MT Summit); ∼51 failed with HTTP errors during the Colab session (timeouts, transient 503s). Fallback papers were scanned on title+abstract (same as abstract column). Bias assessment: ∼34 pre-2010 scanned PDFs constitute all papers from that period, so the 1999–2009 full-text rates are *identical* to abstract rates, so there is no additional bias. HTTP failures (∼51) are distributed across all years (7.7% of post-2019 papers) and do not confound period comparisons. Specialised methods such as MACE, ICC, Dawid–Skene, Scott's π, and Gwet appear only in full text, validating the abstract-level omission documented here.

Table 2: IAA metric frequency across the 665-paper corpus. **Full-text column is the primary analysis** (549/665 papers parsed at full-text level; 116 fall back to abstract). Abstract-level counts are provided to quantify the reporting gap: what readers see without the full paper. Methods such as ICC, MACE, Dawid–Skene, Gwet, and Scott's π are entirely absent from abstracts but present in full text. The 2.5× gap between 34.0% (full text) and 84.1% (abstract) measures the abstract-level under-reporting burden.

checklist in §12.3 targets.

5 Corpus Construction and Annotation Guidelines

The 104 papers in this dimension address annotation-scheme design and reference-corpus production across languages and genres (Gagliardi, 2018; Stymne and Östman, 2020; Papay and Padó, 2020; Farzana et al., 2020; Marcińczuk et al., 2020; Proisl et al., 2020; Baledent et al., 2022; Gecchele et al., 2022; Schaefer and Stede, 2022; Gold et al., 2023; Aracena et al., 2024; Anisimova and Zikánová, 2024; Elmadani et al., 2025; Habash et al., 2025; Antunes et al., 2025; Brants, 2000). Recommended practice is to pilot guidelines, compute IAA, and iterate before full-scale annotation, targeting $\kappa \geq 0.67$ (Artstein and Poesio, 2008). Multilingual and low-resource settings add complexity: translated guidelines may introduce ambiguity and smaller pools weaken IAA estimates, with Arabic, Nordic, Slavic, and South Asian corpora recurring throughout. Crowdsourced corpora increasingly replace expert annotation, but majority voting discards the disagreement structure needed to model annotator variation (Aroyo and Welty, 2015); agreement-with-justification designs (Pavlick and Kwiatkowski, 2019) improve quality but remain rare.

6 Sequence Labelling: NER, Treebanks, and Parsing

The 57 papers span NER across fifteen-plus languages and domain genres (newswire, biomedical, archaeology, biography) and syntactic treebank construction (Friedrich et al., 2020; Thompson et al., 2020; Jørgensen et al., 2020; Brandsen et al., 2020; Ruppenhofer et al., 2020; Luoma et al., 2020; Hahn et al., 2012; Bhat et al., 2014; Solberg et al., 2014; Agić et al., 2014; Bawden et al., 2014; Nguyen et al., 2016). Agreement for broad NER types (PER, ORG, LOC) is typically high ($\kappa > 0.85$) but drops sharply for fine-grained and nested entities, largely independent of annotation platform, suggesting the difficulty is intrinsic to entity definition. For dependency treebanks, Labelled Attachment Score is the de facto measure despite not correcting for chance. Cohen's κ predominates here even with three or more annotators, where Fleiss' κ or Krippendorff's α would be more appropriate.

7 Semantic Annotation: Discourse, SRL, and Lexical Semantics

This is the largest thematic dimension (130 papers), reflecting the difficulty of semantic tasks where interpretation of abstract categories is most variable (Mathet and Widlöcher, 2011a; Mathet et al., 2015; Bourgonje and Stede, 2020; Dhanwal et al., 2020; Peng et al., 2020; Trott and Bergen, 2021;

Goldin et al., 2025; Prévot et al., 2025; Fu et al., 2025; Mírovský et al., 2010; Iñurrieta et al., 2016; Moon et al., 2018). Discourse annotation in the tradition of Rhetorical Structure Theory (Mann and Thompson, 1988) and PDTB-style schemes frequently falls below $\kappa = 0.70$ for fine-grained relation inventories, and the γ measure has gained traction because uncertain span boundaries make κ-on-fixed-spans inappropriate (Mathet et al., 2015). PropBank SRL schemes achieve higher agreement than FrameNet, whose larger frame inventory is more ambiguous (Lo and Wu, 2013). In word-sense annotation, annotators may agree a word is ambiguous yet disagree on the sense, and distributional sense-difficulty models correlate with disagreement (Zaśko-Zielińska and Piasecki, 2018; Pedersen et al., 2018; Danieli et al., 2016; Schlechtweg et al., 2018; Camacho-Collados et al., 2017).

8 Temporal and Event Relation Annotation

The 59 papers cover TimeML-style annotation of events, time expressions, and temporal relations (Pan et al., 2006; Verhagen et al., 2006; Puşcaşu and Barbu Mititelu, 2008; Ljubešić et al., 2010; Matsuyoshi et al., 2010; Morante and Daelemans, 2012; Bitterman et al., 2020; Goel et al., 2020; Vallurupalli et al., 2022; Lee et al., 2022; Tammewar et al., 2022). Event identification achieves relatively high agreement ($\kappa \approx 0.75$–0.90), while temporal relation classification is harder ($\kappa \approx 0.50$–0.70), reflecting the combinatorial explosion of relation types and pervasive vagueness in temporal expressions, a pattern replicated across English, Hindi, and Portuguese corpora. These patterns motivate schemes that distinguish *clear* from *uncertain* relations or let annotators flag underspecified cases rather than forcing a choice (Verhagen et al., 2006); corpus papers also report annotator fatigue, with agreement dropping after the first 100 tokens (Tammewar et al., 2022).

9 Sentiment, Opinion, Stance, and Hate Speech

The 94 papers annotate inherently subjective phenomena (Mohanty et al., 2020; Øvrelid et al., 2020; Assimakopoulos et al., 2020; Titeux et al., 2020; Chakravarthi, 2020; Chakravarthi et al., 2020; Barberia et al., 2025; Parent et al., 2025; Zaghouani and Biswas, 2025; Domínguez et al., 2016; Plank et al., 2014; Stoyanov and Cardie, 2008). Coarse polarity typically achieves $\kappa > 0.75$, but fine-grained emotion schemes (six or more classes) often fall below $\kappa = 0.50$, raising the question of whether the taxonomy reflects real distinctions or annotator interpretation. Hate-speech annotation is among the most contested tasks in NLP: annotators of different demographic, political, or cultural backgrounds diverge on the same item, and inter-group agreement is often lower than intra-group agreement, yet most studies pool annotators without demographic stratification (Saeidi et al., 2021; Spinde et al., 2021; Valizadeh et al., 2021). These findings motivate *distributional labelling* that preserves disagreement and releases all annotations.

10 Annotator Disagreement, Bias, and LLM-as-Annotator

This is the fastest-growing dimension (124 papers, 18.6%), reflecting a field-wide re-evaluation of what annotation disagreement *means* and how it should be handled. The growth is recent and steep: 26 of the 124 papers appeared between 1999 and 2022 against 98 between 2023 and 2026, with 47 in 2025 alone, more than half the corpus's peak-year volume (91 papers; Figure 2a). Every other survey dimension contracts over the same window (Meaney, 2020; Amidei et al., 2020; Aralikatte and Søgaard, 2020; Pamulapati et al., 2020; Aroyo and Welty, 2015; Goot et al., 2020; Huang et al., 2020; Asakura et al., 2020; Han et al., 2021; Klie et al., 2024; James, 2026). The classical view treats disagreement as noise to adjudicate away; an emerging consensus treats it as a signal of linguistic complexity or social variation to preserve and model (Plank, 2022). Probabilistic models such as MACE (Hovy et al., 2013) and CrowdTruth (Aroyo and Welty, 2015) decompose item difficulty from annotator reliability without a gold standard, and several papers link systematic annotation differences to annotator age, gender, and cultural background (Saeidi et al., 2021; Spinde et al., 2021; Valizadeh et al., 2021). Since 2022, a growing body of work examines **LLMs as annotators** (Klie et al., 2024; Occhipinti, 2024; Periti et al., 2024; Sánchez-Montero et al., 2024; Arzt and Hanbury, 2024; Terryn and Lhoneux, 2024): LLMs match crowdworkers on well-defined classification but diverge on subjective phenomena, and their agreement with humans varies with prompt phrasing (James, 2026). Human–LLM agreement is itself an IAA problem and should be quantified with chance-

corrected coefficients, not raw accuracy.

11 Machine Translation Evaluation

The 50 MT evaluation papers occupy a distinct niche: human evaluation of translation quality is itself an IAA problem, since evaluators may rank the same output differently (Castilho, 2020a; Fonteyne et al., 2020; Castilho, 2020b; Popović, 2021; Castilho, 2021; Grundkiewicz et al., 2021). For continuous Direct Assessment scores, ICC and Spearman ρ are more appropriate than Cohen's κ, yet this is rarely acknowledged, a concrete instance of the one-size-fits-all κ pattern of Table 2. Document-level evaluation adds holistic coherence judgements that resist decomposition into item-level labels, requiring segment-level aggregation or a unitising metric such as γ.

12 Discussion

12.1 Metric Appropriateness Analysis

A named coefficient is necessary for reporting but not sufficient for *appropriate* measurement: a paper that names Cohen's κ but uses it on continuous Direct-Assessment scores has reported a metric, but the wrong one. This subsection quantifies the gap between the metric a task *should* use (per the assumptions in §2 and the practitioner guides of James (2026) and Artstein and Poesio (2008)) and the metric corpus papers use, on the 439 papers in our corpus that name at least one coefficient. The appropriateness framework (the metric each annotation setting should use, and the less-appropriate choices observed in the corpus) is given in Appendix C, Table 10; Table 3 reports the corpus frequencies.

Three patterns dominate (Table 3). **Unweighted κ on ordinal labels** is the most prevalent misuse, in 67 papers (15.3% of metric-naming papers) and present in every survey section, peaking in §9 (24%) and §11 (23%) where ordinal scales are common. **Cohen's κ on continuous DA scores** occurs in 16 of 31 metric-naming MT papers (52%), the sharpest single-domain finding. **Raw accuracy for human–LLM agreement** appears in 22 LLM-as-annotator papers (5.0%), and pairwise κ with >2 annotators in a further 18 (4.1%). Across all five patterns, an estimated **25.5% of metric-naming papers** exhibit at least one detectable misuse.

Misuse pattern	Denominator	N	%†
Cohen κ used pairwise with >2 annotators	all metric-naming papers	18	4.1
Cohen κ on continuous DA scores	metric-naming MT papers (§11)	16	52
Unweighted κ on ordinal labels	all metric-naming papers	67	15.3
Fixed-span κ on unitising tasks	all metric-naming papers	11	2.5
Raw accuracy for human vs. LLM agreement	all metric-naming papers	22	5.0
Pooled lower bound	all 439 metric-naming papers	**112**	**25.5**

Table 3: Misuse frequencies from automated detection across the corpus. Counts are lower bounds: a paper is flagged only when the task type and metric choice are both unambiguously inferable from text. A single paper may appear in multiple rows. The pooled bound deduplicates and expresses the fraction of all 439 metric-naming papers that exhibit at least one detectable misuse pattern. † percentages are within each row's denominator (column 2); the MT figure (52%) is the fraction of §11 metric-naming papers rather than of the full corpus.

12.2 Cross-Cutting Findings

1. **Cohen's κ dominates all task types.** In full text κ (259 papers, 38.9%) is the most-named coefficient in every survey section, ahead of Krippendorff's α (157) and Fleiss' κ (114), and is used regardless of whether its assumptions hold; ICC appears in only 16 papers despite suiting continuous MT and emotion tasks (§12.1).

2. **34.0% name no coefficient in full text; 84.1% omit it from the abstract layer.** The 34.0% body-level rate is a methodology-reporting concern and an *upper bound* on papers with no reliability evidence of any kind, since agreement is sometimes reported numerically or procedurally without naming a coefficient. The 50.1-point body-to-abstract gap is a *discoverability* concern: the methodology is reported but is invisible to title/abstract triage.

3. **Confidence intervals are almost never reported** (under 5% of papers), making inter-study comparison unsound (Reichenheim, 2004; James, 2026).

4. **Annotator metadata is routinely omitted**: compensation is reported in 18% of classical annotation papers, selection criteria in 11%, and session duration in under 5%, though crowdsourced and LLM papers disclose compensation

Item	Why it matters
Annotation task type	Determines suitable metric
Label scale (nom./ord./cont./span)	Determines metric family
Number of annotators	Cohen vs. Fleiss vs. α
Double-annotated item count	Needed to interpret reliability
Metric name and variant	Avoids ambiguity
Library / implementation version	Supports reproducibility
Raw agreement	Sanity check
Chance-corrected agreement + CI	Controls expected agreement
Per-label agreement breakdown	Reveals minority-label issues
Annotator selection criteria	Supports bias analysis
Compensation and session duration	Enables annotator-effect study
Adjudication procedure	Explains gold-label construction
Individual labels released	Enables disagreement modelling

Table 4: Per-paper minimum reporting checklist for ACL submissions. Pair with Table 6 for the setting-indexed metric recommendations.

far more often (58% and 64%).

5. **Disagreement is underutilised as a signal**: most pipelines adjudicate away minority labels, discarding genuine ambiguity (Plank, 2022).

6. **Cross-lingual IAA is under-theorised**: no principled framework compares κ across typologically distinct languages.

12.3 A Reporting Checklist for ACL Submissions

Combining the metric-selection guidance of §12.1 with the reporting gaps documented in §12.2 yields a concrete checklist for authors and reviewers. Table 6 maps annotation settings to recommended metrics and minimum reporting items, and Table 4 gives a complementary per-paper checklist, indexed by manuscript content and intended as a final pass an author or reviewer can run against a completed draft. Each item is phrased as a surface-verifiable fact, for the reasons the compliance audit below makes clear.

Compliance audit. To test whether the checklist can be applied reliably, four annotators scored the 50 most recent papers in the corpus (37 from 2025, 13 from 2026) on the 13 items of Table 5: two human coders working independently from the item definitions, and two LLMs (Claude Opus 4.8 and GPT-4o, accessed via aiXplain) given the same rubric. The audit is a measurement study of the checklist itself, not a ranking of annotators.

No pair of annotators agrees well. The two humans reach Cohen's $\kappa = \mathbf{0.26}$ (95% bootstrap CI: 0.20–0.33), *below* the $\kappa = 0.36$ of the two LLMs; human–LLM pairs span 0.06–0.42. Mean compliance estimates vary by a factor of three across annotators (48.9% and 36.3% for the humans, 45.1% and 16.8% for Claude Opus 4.8 and GPT-4o), with the human estimates bracketing most of the LLM range. Since two humans applying the same rubric to the same 50 papers agree only fairly, the low LLM–LLM agreement cannot be read as evidence that LLMs in particular are poor compliance annotators: the unreliability is a property of the task as operationalised, not of the annotator type.

Splitting the items sharpens this. On the nine objectively checkable items (a coefficient is named or it is not; a CI is present or absent) the two humans reach $\kappa = 0.23$, against $\kappa = 0.18$ on the four judgment-laden items (C05, C09, C10, C12), where a coder must decide whether a prose description of adjudication, annotator selection, compensation, or expertise counts as reported. Agreement is therefore weak across the board and weakest where the items require interpretation, and per-paper annotator notes attribute most disagreements to partial reporting, such as an adjudication procedure mentioned in passing without detail.

We draw two conclusions. First, the compliance figures in Table 5 support only a range, and we report the average recent ACL annotation paper as satisfying somewhere between one in six and one in two of the minimum reporting items. All four annotators nonetheless agree on the qualitative picture: the least-reported items are confidence intervals (C03), raw agreement (C04), and session duration (C11), none of which require additional data collection, while the metric name and label scale are reported most often. Second, and for the checklist itself more consequential, reporting compliance cannot currently be audited reliably by *any* annotator type at this granularity. A checklist intended for reviewers must therefore be operationalised so that each item is a surface-level, verifiable fact rather than a judgement, and we phrase the items of Table 6 to that standard.

12.4 Open Challenges and Research Agenda

The findings above motivate a seven-point agenda. **(1) Confidence intervals:** report a 95% CI with

Item	Checklist element	Human		LLM	
		H1	*H2*	*Cl.*	*GPT*
Objectively checkable items					
C01	Metric named	78	72	78	36
C02	Annotator count stated	62	34	84	26
C03	CI reported	2	6	10	2
C04	Raw agreement reported	12	18	14	6
C06	Label scale specified	72	46	60	30
C07	Library / version cited	22	16	26	8
C08	Per-label breakdown	34	24	40	2
C11	Session duration	28	8	16	4
C13	Data / labels released	50	30	68	36
Judgment-laden items					
C05	Adjudication described	82	82	56	16
C09	Annotator selection	54	38	46	8
C10	Compensation stated	52	24	18	8
C12	Expertise described	88	74	70	36
Mean (13 items)		**48.9**	**36.3**	**45.1**	**16.8**

Table 5: Checklist compliance rates (%) for the 50 most recent papers in the corpus (2025–2026), scored independently by two human coders (H1, H2) and two LLMs (Claude Opus 4.8, GPT-4o via aiXplain) on all 13 items. No annotator pair agrees well: human–human $\kappa = 0.26$ (95% CI: 0.20–0.33), LLM–LLM $\kappa = 0.36$, human–LLM 0.06–0.42. Estimates span a factor of three, so the audit supports a range rather than a point estimate, and shows that compliance at this granularity cannot be audited reliably by any annotator type under the present item definitions (§12.3).

Annotation setting	Recommended metric	Minimum reporting
2 annotators, nominal labels	Cohen's κ + raw agreement	Marginals, CI, double-coded n
>2 annotators, nominal	Krippendorff's α or Fleiss' κ	Rater count, missing-data handling
Ordinal labels	Weighted κ / α ordinal	Weighting scheme stated
Span / unitizing	γ + Span F1	Boundary-matching policy
Continuous ratings	ICC (specify variant) + CI	ICC type, rater model
Subjective / crowd	Distributional labels + MACE or CrowdTruth	Annotator metadata, label distributions
LLM as annotator	Human vs. LLM κ or α	Prompt version, temperature, model ID

Table 6: Operational reporting checklist indexed by annotation setting. For each setting, name the recommended metric and report all minimum items. Pair with Table 4 for the per-paper checklist.

every IAA estimate (Reichenheim, 2004). **(2) Reporting standard:** a four-annotator audit of the 50 most recent corpus papers puts mean checklist compliance at 17–49% (Table 5); adopt the metric name and parametrisation, annotator count, double-annotated item count, per-category breakdown, CI, and selection procedure (Hallgren, 2012; James, 2026), and define each item so that a reviewer can verify it from the surface text. **(3) One-size-fits-all κ:** 52% of metric-naming MT papers use κ where ICC suits continuous DA scores; justify metric choice against task assumptions and co-report percentage agreement. **(4) Discarded disagreement:** release all individual annotations, not only adjudicated gold labels. **(5) Unmeasured LLM agreement:** human–LLM agreement is an IAA problem and must be measured with chance-corrected coefficients rather than raw accuracy. **(6) Annotator metadata:** report annotator count, compensation, and ethics review status alongside every IAA statistic, since annotator effects are systematic (James, 2026). **(7) Cross-lingual framework:** develop IAA comparison across typologically distinct languages that accounts for annotation difficulty and guideline translation.

13 Conclusion

We presented a large-scale empirical survey of IAA reporting practice in NLP, covering 665 ACL Anthology papers (1999–2026). Full-text analysis of 549 papers confirms that Cohen's κ dominates (259 papers, 38.9%) and that 25.5% of metric-naming papers exhibit at least one detectable misuse, most often unweighted κ on ordinal labels (15.3%) and, in MT evaluation, κ on continuous DA scores (52% of metric-naming MT papers). We separate two reporting concerns: 34.0% of papers name no coefficient in the full text, an upper bound on the share reporting no reliability evidence of any kind, and a further 50.1% name a coefficient in the body but not the abstract, a discoverability concern. A four-annotator audit (two humans, two LLMs) of the 50 most recent papers puts mean adherence to the 13-item checklist between 17% and 49% depending on the annotator, and no pair agrees well (human–human $\kappa = 0.26$, LLM–LLM $\kappa = 0.36$), which shows that reporting compliance cannot be audited reliably at this granularity until checklist items are defined as surface-verifiable facts. The rise of disagreement modelling and LLM-as-annotator (18.6% of the corpus) signals that the classical gold-label paradigm is under empirical pressure. This survey complements the prescriptive guidance of James (2026) with field-scale evidence on how IAA is measured and reported in practice.

Limitations

Corpus boundary. The search ran over all 121,832 papers indexed in the ACL Anthology, which includes the *Computational Linguistics* journal and *TACL*; the corpus contains five CL journal papers and seven TACL papers. Excluded is work published outside the Anthology: IEEE and speech venues (Interspeech, ICASSP), computer vision (CVPR), clinical and health-informatics journals, and linguistics journals. The corpus may also underrepresent non-English-medium communities.

Search vocabulary and title/abstract matching. Two boundaries limit recall. First, the search vocabulary matches *inter-annotator* and *inter-rater* but not *inter-coder*, so studies using only that term, including the meta-analysis of Bayerl and Paul (2011), fall outside the corpus; we release an *inter-coder* pattern for future extensions but did not apply it retroactively, since doing so would change the corpus the reported figures describe. Second, matching on title and abstract rather than full text is a precision–recall trade-off: full-text matching returns thousands of papers that mention κ once in passing, the pattern our manual exclusion step removed 70 papers for, but it would also recover papers that analyse agreement without flagging it in the abstract. Some older Anthology entries lack abstract metadata and are invisible to a title+abstract search. The corpus is therefore best read as papers that *foreground* IAA, not as every Anthology paper that measures it.

Scope of the metric-appropriateness claims. Two claims in §12.1 have different exposure to venue selection. The misuse findings (25.5% of metric-naming papers; 52% of metric-naming MT papers applying κ to continuous DA scores) are within-paper mismatches between a paper's own data type and its own metric choice, so they cannot be produced by which venues were sampled. The *prevalence* figures are bounded by venue selection, and we scope them accordingly: ICC appears in 2.4% of full texts *in the ACL venues sampled*, not in NLP at large. ICC is standard in clinical reliability studies, and the corpus contains only seven clinical or biomedical papers (three Clinical NLP workshop papers plus BioNLP, SMM4H, and CLPsych entries), none of which names ICC. That this community is nearly absent from the Anthology is itself the point: clinical reliability work is published elsewhere, so we scope the prevalence claim rather than generalising it, and an extension to JAMIA or BMC Bioinformatics would likely raise ICC counts.

Clustering. The NMF parameters, including the choice of k=16 topics, were fixed by inspecting topic coherence; alternative settings yield different but equally valid taxonomies. **Full-text vs. abstract-level metric scan.** Table 2 reports metric frequency at two levels: *full text* (primary; 549/665 papers, 82.6%) and *abstract* (secondary; all 665 papers). The 116 fallback papers could not be parsed from PDF (scanned images, external venues, network errors) and contribute to the abstract column only; this is conservative. Abstract-level counts are additionally reported because that is what readers and reviewers encounter without the full paper; the 84.1% abstract gap is a distinct policy-relevant finding alongside the 34.0% full-text figure. **Sensitivity analysis.** Best case (all 116 fallbacks name a metric): unnamed falls to 18.2%; worst case (none do): rises to 35.6%. Parsed coverage is 18% (pre-2010), 55% (2010–2018), 100% (2019–2026). Readers seeking prescriptive metric-selection guidance should consult James (2026), which this survey is designed to complement. **Coder reliability for paper-type classification.** The *any metric named* dimension is algorithmic and reproducible. The *paper-type* dimension was assigned by keyword heuristics and validated by the independent double-coding reported in Appendix A.

Compliance audit. The audit covers 50 papers and four annotators, and its central result is negative: no annotator pair agrees well enough to license a point estimate of checklist compliance (§12.3). The compliance percentages should therefore be read as a range whose width reflects annotator disagreement rather than sampling error, and the item-level agreement figures as diagnostics of the checklist's wording rather than as measurements of the papers.

Ethics Statement

This work surveys existing published research and introduces no new annotation tasks or human subjects research. The corpus metadata (paper IDs, titles, abstracts, venue, year) is derived from the ACL Anthology and redistributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence. Code released with this paper is provided under the MIT Licence. No private or sensitive data were accessed. The only human con-

tributor besides the author is an independent second coder who double-coded a stratified 10% sample for the inter-coder reliability analysis (Appendix A) and was compensated at standard academic rates; the survey involves no other recruitment of human participants and no collection of personal information.

Acknowledgments

The author thanks the independent second annotator, a PhD-level NLP researcher, for the double-coding work underpinning the inter-coder reliability analysis and the regex extraction validation.

Use of AI assistants. Two large language models, Claude Opus 4.8 and GPT-4o, served as annotators in the reporting-compliance audit of §12.3; their agreement and limitations are described there, and their raw outputs are released with the corpus. The author also used AI assistants during manuscript preparation for copy-editing, for checking numerical consistency between the tables and the text, and for revising prose. The study design, corpus construction, analyses, and conclusions are the author's own, and the author reviewed every AI-assisted edit and verified all reported figures.

References

Željko Agić, Daša Berović, Danijela Merkler, and Marko Tadić. 2014. Croatian Dependency Treebank 2.0: New Annotation Guidelines for Improved Parsing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ACL Anthology: L14-1545.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2020. Identifying Annotator Bias: A new IRT-based method for bias identification. In *Proceedings of the 28th International Conference on Computational Linguistics*. ACL Anthology: 2020.coling-main.421.

Mariia Anisimova and Šárka Zikánová. 2024. Attitudes in Diplomatic Speeches: Introducing the CoDipA UNSC 1.0. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*. ACL Anthology: 2024.isa-1.3.

David Antunes, Jorge Baptista, and Nuno J. Mamede. 2025. A European Portuguese corpus annotated for verbal idioms. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*. ACL Anthology: 2025.mwe-1.7.

Claudio Aracena, Luis Miranda, Thomas Vakili, Fabián Villena, Tamara Quiroga, Fredy Núñez-Torres, Victor Rocco, and Jocelyn Dunstan. 2024. A Privacy-Preserving Corpus for Occupational Health in Spanish: Evaluation for NER and Classification Tasks. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*. ACL Anthology: 2024.clinicalnlp-1.11.

Rahul Aralikatte and Anders Søgaard. 2020. Model-based Annotation of Coreference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.9.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Varvara Arzt and Allan Hanbury. 2024. Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*. ACL Anthology: 2024.genbench-1.8.

Takuto Asakura, André Greiner-Petter, Akiko Aizawa, and Yusuke Miyao. 2020. Towards Grounding of Formulae. In *Proceedings of the First Workshop on Scholarly Document Processing*. ACL Anthology: 2020.sdp-1.16.

Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. 2020. Annotating for Hate Speech: The MaNeCo Corpus and Some Input from Critical Discourse Analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.626.

Anaëlle Baledent, Yann Mathet, Antoine Widlöcher, Christophe Couronne, and Jean-Luc Manguin. 2022. Validity, Agreement, Consensuality and Annotated Data Quality. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. ACL Anthology: 2022.lrec-1.315.

Lorena Barberia, Pedro Schmalz, Norton Trevisan Roman, Belinda Lombard, and Tatiane Moraes de Sousa. 2025. It's about What and How you say it: A Corpus with Stance and Sentiment Annotation for COVID-19 Vaccines Posts on X/Twitter by Brazilian Political Elites. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. ACL Anthology: 2025.nlp4dh-1.32.

Rachel Bawden, Marie-Amélie Botalla, Kim Gerdes, and Sylvain Kahane. 2014. Correcting and Validating Syntactic Dependency in the Spoken French Treebank Rhapsodie. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ACL Anthology: L14-1596.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Riyaz Ahmad Bhat, Shahid Mushtaq Bhat, and Dipti Misra Sharma. 2014. Towards building a Kashmiri Treebank: Setting up the Annotation Pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ACL Anthology: L14-1187.

Danielle Bitterman, Timothy Miller, David Harris, Chen Lin, Sean Finan, Jeremy Warner, Raymond Mak, and Guergana Savova. 2020. Extracting Relations between Radiotherapy Treatment Details. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. ACL Anthology: 2020.clinicalnlp-1.21.

Peter Bourgonje and Manfred Stede. 2020. The Potsdam Commentary Corpus 2.2: Extending Annotations for Shallow Discourse Parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.133.

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.562.

Thorsten Brants. 2000. Inter-annotator Agreement for a German Newspaper Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. ACL Anthology: L00-1247.

Dany Bregeon, Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Halftermeyer. 2019. Redonner du sens à l'accord interannotateur : vers une interprétation des mesures d'accord en termes de reproductibilité de l'annotation [Interpreting inter-annotator agreement measures : towards an interpretation in terms of annotation reproducibility]. In *Traitement Automatique des Langues, Volume 60, Numéro 2 : Corpus annotés [Annotated corpora]*. ACL Anthology: 2019.tal-2.3.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 Task 2: Multilingual and Cross-lingual Semantic Word Similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. ACL Anthology: S17-2002.

Sheila Castilho. 2020a. Document-Level Machine Translation Evaluation Project: Methodology, Effort and Inter-Annotator Agreement. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. ACL Anthology: 2020.eamt-1.49.

Sheila Castilho. 2020b. On the Same Page? Comparing Inter-Annotator Agreement in Sentence and Document Level Human Machine Translation Evaluation. In *Proceedings of the Fifth Conference on Machine Translation*. ACL Anthology: 2020.wmt-1.137.

Sheila Castilho. 2021. Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. ACL Anthology: 2021.humeval-1.4.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. `https://github.com/pln-fing-udelar/fast-krippendorff`.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*. ACL Anthology: 2020.peoples-1.5.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. 2020. Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. ACL Anthology: 2020.sltu-1.28.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Morena Danieli, Balamurali A R, Evgeny Stepanov, Benoit Favre, Frederic Bechet, and Giuseppe Riccardi. 2016. Summarizing Behaviours: An Experiment on the Annotation of Call-Centre Conversations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. ACL Anthology: L16-1701.

Swapnil Dhanwal, Hritwik Dutta, Hitesh Nankani, Nilay Shrivastava, Yaman Kumar, Junyi Jessy Li, Debanjan Mahata, Rakesh Gosangi, Haimin Zhang, Rajiv Ratn Shah, and Amanda Stent. 2020. An Annotated Dataset of Discourse Modes in Hindi Stories. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.149.

Barbara Di Eugenio and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Mónica Domínguez, Mireia Farrús, and Leo Wanner. 2016. An Automatic Prosody Tagger for Spontaneous Speech. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL Anthology: C16-1037.

Khalid N. Elmadani, Nizar Habash, and Hanada Taha-Thomure. 2025. A Large and Balanced Corpus for Fine-grained Arabic Readability Assessment. In

Findings of the Association for Computational Linguistics: ACL 2025. ACL Anthology: 2025.findings-acl.842.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling Dialogue in Conversational Cognitive Health Screening Interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.147.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.468.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL Anthology: 2020.acl-main.116.

Yingxue Fu, Mark-Jan Nederhof, and Anais Ollagnier. 2025. A Topicality-Driven QUD Model for Discourse Processing. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL Anthology: 2025.sigdial-1.17.

Gloria Gagliardi. 2018. Inter-Annotator Agreement in linguistica: una rassegna critica. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*. ACL Anthology: 2018.clicit-1.37.

Mingqi Gao, Jie Ruan, and Xiaojun Wan. 2023. A Reproduction Study of the Human Evaluation of Role-Oriented Dialogue Summarization Models. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*. ACL Anthology: 2023.humeval-1.10.

Marcello Gecchele, Hiroaki Yamada, Takenobu Tokunaga, Yasuyo Sawaki, and Mika Ishizuka. 2022. Automating Idea Unit Segmentation and Alignment for Assessing Reading Comprehension via Summary Protocol Analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. ACL Anthology: 2022.lrec-1.498.

Pranav Goel, Suhan Prabhu, Alok Debnath, Priyank Modi, and Manish Shrivastava. 2020. Hindi TimeBank: An ISO-TimeML Annotated Reference Corpus. In *Proceedings of the 16th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. ACL Anthology: 2020.isa-1.2.

Christian Gold, Ronja Laarmann-Quante, and Torsten Zesch. 2023. Preserving the Authenticity of Handwritten Learner Language: Annotation Guidelines for Creating Transcripts Retaining Orthographic Features. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*. ACL Anthology: 2023.cawl-1.3.

Gili Goldin, Shira Wigderson, Ella Rabinovich, and Shuly Wintner. 2025. An Annotation Scheme for Factuality and Its Application to Parliamentary Proceedings. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*. ACL Anthology: 2025.ranlp-1.49.

Rob van der Goot, Alan Ramponi, Tommaso Caselli, Michele Cafagna, and Lorenzo De Mattei. 2020. Norm It! Lexical Normalization for Italian and Its Downstream Effects for Dependency Parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.769.

Jonas Groschwitz, Shay Cohen, Lucia Donatelli, and Meaghan Fowlie. 2023. AMR Parsing is Far from Solved: GrAPES, the Granular AMR Parsing Evaluation Suite. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. ACL Anthology: 2023.emnlp-main.662.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, Christian Federmann, and Tom Kocmi. 2021. On User Interfaces for Large-Scale Document-Level Human Evaluation of Machine Translation Outputs. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*. ACL Anthology: 2021.humeval-1.11.

Kilem Gwet. 2001. *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, Gaithersburg, MD.

Nizar Habash, Hanada Taha-Thomure, Khalid N. Elmadani, Zeina Zeino, and Abdallah Abushmaes. 2025. Guidelines for Fine-grained Sentence-level Arabic Readability Annotation. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*. ACL Anthology: 2025.law-1.30.

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, Erik Faessler, Jenny Traumüller, Susann Schröder, and Kerstin Hornbostel. 2012. Iterative Refinement and Quality Checking of Annotation Guidelines — How to Deal Effectively with Semantically Sloppy Named Entity Types, such as Pathological Phenomena. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. ACL Anthology: L12-1441.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1):23–34.

Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*. ACL Anthology: 2021.crac-1.3.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. ACL Anthology: 2020.nlpcovid19-acl.6.

Uxoa Iñurrieta, Arantza Díaz de Ilarraza, Gorka Labaka, Kepa Sarasola, Itziar Aduriz, and John Carroll. 2016. Using Linguistic Data for English and Spanish Verb-Noun Combination Identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. ACL Anthology: C16-1082.

Joseph H. F. James. 2026. Counting on Consensus: Selecting the Right Inter-annotator Agreement Metric for NLP Annotation and Evaluation. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 4434–4446, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. NorNE: Annotating Named Entities for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.559.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing Dataset Annotation Quality Management in the Wild. In *Computational Linguistics, Volume 50, Issue 3 - September 2024*. ACL Anthology: 2024.cl-3.1.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*, 3rd edition. SAGE Publications.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for Human Evaluation of Faithfulness in Long-form Summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. ACL Anthology: 2023.eacl-main.121.

Ronja Laarmann-Quante, Leska Schwarz, Andrea Horbach, and Torsten Zesch. 2022. 'Meet me at the ribary' – Acceptability of spelling variants in free-text answers to listening comprehension prompts. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*. ACL Anthology: 2022.bea-1.22.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Meisin Lee, Lay-Ki Soon, Eu Gene Siew, and Ly Fie Sugianto. 2022. CrudeOilNews: An Annotated Crude Oil News Corpus for Event Extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. ACL Anthology: 2022.lrec-1.49.

Seunggun Lee, Alexandra DeLucia, Nikita Nangia, Praneeth Ganedi, Ryan Guan, Rubing Li, Britney Ngaw, Aditya Singhal, Shalaka Vaidya, Zijun Yuan, Lining Zhang, and João Sedoc. 2023. Common Law Annotations: Investigating the Stability of Dialog System Output Annotations. In *Findings of the Association for Computational Linguistics: ACL 2023*. ACL Anthology: 2023.findings-acl.780.

Yiru Li, Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Same Trends, Different Answers: Insights from a Replication Study of Human Plausibility Judgments on Narrative Continuations. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*. ACL Anthology: 2023.humeval-1.15.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL Anthology: 2023.acl-long.228.

Nikola Ljubešić, Tomislava Lauc, and Damir Boras. 2010. Building a Gold Standard for Event Detection in Croatian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. ACL Anthology: L10-1145.

Chi-kiu Lo and Dekai Wu. 2013. Human semantic MT evaluation with HMEANT for IWSLT 2013. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*. ACL Anthology: 2013.iwslt-evaluation.2.

Sharid Loáiciga, Simon Dobnik, and David Schlangen. 2022. Anaphoric Phenomena in Situated dialog: A First Round of Annotations. In *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*. ACL Anthology: 2022.crac-1.4.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A Broad-coverage Corpus for Finnish Named Entity Recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.567.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Michał Marcińczuk, Marcin Oleksy, and Jan Wieczorek. 2020. PST 2.0 – Corpus of Polish Spatial Texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.265.

Yann Mathet and Antoine Widlöcher. 2011a. Aligner : un outil d'alignement et de mesure d'accord inter-annotateurs (Aligner: a tool for aligning and measuring inter-annotator agreement). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Démonstrations*. ACL Anthology: 2011.jeptalnrecital-demonstration.15.

Yann Mathet and Antoine Widlöcher. 2011b. Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs (A holistic and unified approach to aligning and measuring inter-annotator agreement). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*. ACL Anthology: 2011.jeptalnrecital-long.16.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. In *Computational Linguistics, Volume 41, Issue 3 - September 2015*. ACL Anthology: J15-3003.

Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. 2010. Annotating Event Mentions in Text with Modality, Focus, and Source Information. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. ACL Anthology: L10-1469.

J. A. Meaney. 2020. Crossing the Line: Where do Demographic Variables Fit into Humor Detection? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. ACL Anthology: 2020.acl-srw.24.

Gaurav Mohanty, Pruthwik Mishra, and Radhika Mamidi. 2020. Annotated Corpus for Sentiment Analysis in Odia Language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.339.

Lori Moon, Christos Christodoulopoulos, Cynthia Fisher, Sandra Franco, and Dan Roth. 2018. Gold Standard Annotations for Preposition and Verb Sense with Semantic Role Labels in Adult-Child Interactions. In *Proceedings of the 27th International Conference on Computational Linguistics*. ACL Anthology: C18-1254.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. ACL Anthology: L12-1077.

Jiří Mírovský, Lucie Mladová, and Šárka Zikánová. 2010. Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT. In *Coling 2010: Posters*. ACL Anthology: C10-2089.

Quy Nguyen, Yusuke Miyao, Ha Le, and Ngan Nguyen. 2016. Challenges and Solutions for Consistent Annotation of Vietnamese Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. ACL Anthology: L16-1243.

Laura Occhipinti. 2024. Complex Word Identification for Italian Language: A Dictionary–based Approach. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*. ACL Anthology: 2024.clib-1.12.

Vaishnavi Pamulapati, Gayatri Purigilla, and Radhika Mamidi. 2020. A Novel Annotation Schema for Conversational Humor: Capturing the Cultural Nuances in Kanyasulkam. In *Proceedings of the 14th Linguistic Annotation Workshop*. ACL Anthology: 2020.law-1.4.

Feng Pan, Rutu Mulkar, and Jerry R. Hobbs. 2006. An Annotated Corpus of Typical Durations of Events. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. ACL Anthology: L06-1128.

Sean Papay and Sebastian Padó. 2020. RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.104.

Donnie Parent, Nina Georgiades, Charvi Mishra, Khaled Mohammed, and Sandra Kübler. 2025. Annotating Hate Speech towards Identity Groups. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*. ACL Anthology: 2025.ranlp-1.102.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 677–694.

Bolette Pedersen, Manex Agirrezabal, Sanni Nimb, Ida Olsen, and Sussi Olsen. 2018. Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In *Proceedings of the 9th Global Wordnet Conference*. ACL Anthology: 2018.gwc-1.21.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A Corpus of Adpositional Supersenses for Mandarin Chinese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.733.

Francesco Periti, Pierluigi Cassotti, Stefano Montanelli, Nina Tahmasebi, and Dominik Schlechtweg. 2024. TRoTR: A Framework for Evaluating the Re-contextualization of Text Reuse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. ACL Anthology: 2024.emnlp-main.774.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. ACL Anthology: E14-1078.

Maja Popović. 2021. Agree to Disagree: Analysis of Inter-Annotator Disagreements in Human Evaluation of Machine Translation Output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. ACL Anthology: 2021.conll-1.18.

Maja Popović and Anya Belz. 2021. A Reproduction Study of an Annotation-based Human Evaluation of MT Outputs. In *Proceedings of the 14th International Conference on Natural Language Generation*. ACL Anthology: 2021.inlg-1.31.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. 2022. Reproducing a Manual Evaluation of the Simplicity of Text Simplification System Outputs. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*. ACL Anthology: 2022.inlg-genchal.12.

Thomas Proisl, Natalie Dykes, Philipp Heinrich, Besim Kabashi, Andreas Blombach, and Stefan Evert. 2020. EmpiriST Corpus 2.0: Adding Manual Normalization, Lemmatization and Semantic Tagging to a German Web and CMC Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.754.

Laurent Prévot, Roxane Bertrand, and Julie Hunter. 2025. Segmenting a Large French Meeting Corpus into Elementary Discourse Units. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. ACL Anthology: 2025.sigdial-1.14.

Georgiana Puşcaşu and Verginica Barbu Mititelu. 2008. Annotation of WordNet Verbs with TimeML Event Classes. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. ACL Anthology: L08-1563.

Michael E Reichenheim. 2004. Confidence intervals for the kappa statistic. *The Stata Journal*, 4(4):421–428.

Josef Ruppenhofer, Ines Rehbein, and Carolina Flinz. 2020. Fine-grained Named Entity Annotations for German Biographic Interviews. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.566.

Marzieh Saeidi, Majid Yazdani, and Andreas Vlachos. 2021. Cross-Policy Compliance Detection via Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. ACL Anthology: 2021.emnlp-main.678.

Aleksandar Savkov, Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Anya Belz, and Ehud Reiter. 2022. Consultation Checklists: Standardising the Human Evaluation of Medical Note Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. ACL Anthology: 2022.emnlp-industry.10.

Robin Schaefer and Manfred Stede. 2022. GerCCT: An Annotated Corpus for Mining Arguments in German Tweets on Climate Change. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. ACL Anthology: 2022.lrec-1.658.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. ACL Anthology: N18-2027.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, pages 321–325.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Per Erik Solberg, Arne Skjærholt, Lilja Øvrelid, Kristin Hagen, and Janne Bondi Johannessen. 2014. The Norwegian Dependency Treebank. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ACL Anthology: L14-1273.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. ACL Anthology: 2021.findings-emnlp.101.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Regina Stodden and Laura Kallmeyer. 2022. TS-ANNO: An Annotation Tool to Build, Annotate and Evaluate Text Simplification Corpora. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. ACL Anthology: 2022.acl-demo.14.

Veselin Stoyanov and Claire Cardie. 2008. Annotating Topics of Opinions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. ACL Anthology: L08-1088.

Sara Stymne and Carin Östman. 2020. SLäNDa: An Annotated Corpus of Narrative and Dialogue in Swedish Literary Fiction. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.103.

Alec Sánchez-Montero, Gemma Bel-Enguix, and Sergio-Luis Ojeda-Trueba. 2024. Evaluating the Development of Linguistic Metaphor Annotation in Mexican Spanish Popular Science Tweets. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*. ACL Anthology: 2024.figlang-1.8.

Aniruddha Tammewar, Franziska Braun, Gabriel Roccabruna, Sebastian Bayerl, Korbinian Riedhammer, and Giuseppe Riccardi. 2022. Annotation of Valence Unfolding in Spoken Personal Narratives. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. ACL Anthology: 2022.lrec-1.757.

Ayla Rigouts Terryn and Miryam de Lhoneux. 2024. Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*. ACL Anthology: 2024.humeval-1.2.

Paul Thompson, Tim Yates, Emrah Inan, and Sophia Ananiadou. 2020. Semantic Annotation for Improved Safety in Construction Work. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.245.

Hadrien Titeux and Rachid Riad. 2021. pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in Python. *Journal of Open Source Software*, 6(62):2989.

Hadrien Titeux, Rachid Riad, Xuan-Nga Cao, Nicolas Hamilakis, Kris Madden, Alejandrina Cristia, Anne-Catherine Bachoud-Lévi, and Emmanuel Dupoux. 2020. Seshat: a Tool for Managing and Verifying Annotation Campaigns of Audio Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.861.

Sean Trott and Benjamin Bergen. 2021. RAW-C: Relatedness of Ambiguous Words in Context (A New Lexical Resource for English). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL Anthology: 2021.acl-long.550.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying Medical Self-Disclosure in Online Communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL Anthology: 2021.naacl-main.347.

Sai Vallurupalli, Sayontan Ghosh, Katrin Erk, Niranjan Balasubramanian, and Francis Ferraro. 2022. POQue: Asking Participant-specific Outcome Questions for a Deeper Understanding of Complex Events. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. ACL Anthology: 2022.emnlp-main.594.

Marc Verhagen, Robert Knippen, Inderjeet Mani, and James Pustejovsky. 2006. Annotation of Temporal Relations with Tango. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. ACL Anthology: L06-1394.

Elena Volodina, Yousuf Ali Mohammed, Sandra Derbring, Arild Matsson, and Beata Megyesi. 2020. Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays. In *Proceedings of the 28th International Conference on Computational Linguistics*. ACL Anthology: 2020.coling-main.32.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.

Wajdi Zaghouani and Md. Rafiul Biswas. 2025. An Annotated Corpus of Arabic Tweets for Hate Speech Analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*. ACL Anthology: 2025.ranlp-1.163.

Monika Zaśko-Zielińska and Maciej Piasecki. 2018. Towards Emotive Annotation in plWordNet 4.0. In

Proceedings of the 9th Global Wordnet Conference. ACL Anthology: 2018.gwc-1.18.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A Fine-grained Sentiment Dataset for Norwegian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. ACL Anthology: 2020.lrec-1.618.

Coding dimension	κ	95% CI		n	Strength
Algorithmic regex extraction (precision / recall / F1)					
Any metric named (full text)	P=0.96, R=0.91, F1=0.93 (n=50)				high
Per-family P (range)	0.93–1.00 across 15 metric families				high
Per-family R (range)	0.88–0.97 across 15 metric families				high
Human-coded dimensions (independent coder, n=50)					
Paper type (6 cat.)	0.634	0.465	0.789	50	substantial
Survey section (8 cat.)	0.78	0.66	0.88	50	substantial
Per-category κ (paper type): Corpus: 0.87 Disagreement: 0.83 Empirical: 0.77 LLM: 0.57 Methodology: 0.00					

Table 7: Reliability of all coding decisions in this survey. Algorithmic patterns are validated by precision/recall on an independent 50-paper sample; human-coded dimensions are validated by an independent second coder on a stratified 10% sample. The second coder is a PhD-level NLP researcher, compensated at standard academic rates. Bootstrap 95% CI (5,000 iterations). Raw agreement: paper type 74%, survey section 84%. Per-category κ for paper type is reported in the bottom row; the low Methodology score reflects n=3 cell size rather than genuine unreliability (see text).

A Coding Reliability and Algorithmic Extraction Validation

This appendix expands the reliability evidence summarised in §3. Table 7 reports the reliability of every coding decision in the survey: the algorithmic regex extractor is validated by precision and recall on an independent 50-paper sample, and the two human-coded dimensions are validated by an independent second coder on a stratified 10% sample.

Human-coded dimensions. *Paper type.* Cohen's $\kappa = 0.634$ (95% CI: 0.465–0.789), indicating substantial agreement (Landis and Koch, 1977); raw agreement 74%. Per-category κ: Corpus 0.87, Disagreement study 0.83, Empirical 0.77, LLM annotation 0.57, Methodology 0.00. The low Methodology score reflects the small cell size (n=3 papers) rather than genuine unreliability; the four disagreements involved boundary cases between Methodology and Empirical or Corpus, adjudicated by discussion. The most common confusion pattern was Corpus vs. Empirical (3 cases), arising when a paper both releases a dataset and reports model performance; the adjudication rule codes such a paper as Corpus when the corpus is its named primary contribution.

Survey section. Cohen's $\kappa = 0.78$ (95% CI: 0.66–0.88); raw agreement 84%. The higher agreement reflects the clearer thematic separation of NMF-anchored sections relative to paper-type categories. Disagreements concentrated on the boundary between Semantics (§7) and Disagreement & LLM annotation (§10) for papers that combine a semantic annotation scheme with explicit disagreement modelling.

Algorithmic extraction: precision and recall. The *any metric named* dimension is derived from 61 regular expressions grouped into 15 metric families (Appendix B), applied to title, abstract, and full text; a paper is flagged for a family if any of its patterns matches, and several patterns require the metric term within a short window of annotation vocabulary to exclude background-section false positives. We validated the patterns on a 50-paper stratified sample independent of the inter-coder reliability sample: two coders read each paper end-to-end and recorded every metric mention the patterns should have flagged, with disagreements adjudicated by discussion. Pooled across all 15 families, the regex extractor achieves precision = 0.96 (95% CI: 0.91–0.99), recall = 0.91 (95% CI: 0.84–0.96), and F1 = 0.93. False positives (n=3/76 flagged mentions) were dominated by metric terms in related-work discussion of prior corpora rather than the paper's own measurement; false negatives (n=9) concentrated on (i) abbreviated forms in tables (e.g., "κ" alone without "Cohen" or "kappa" in adjacent text; 6/9) and (ii) non-standard library names (e.g., `statsmodels.stats.inter_rater`; 2/9). Per-family precision is uniformly $\geq$0.93; per-family recall is lowest for Cohen's κ (0.88) due to the abbreviation pattern. Correcting the headline 34.0% "no coefficient named" rate for these error rates yields a 95% band of 28.4%–33.7%, so the headline finding is robust to extraction error.

B Metric-Extraction Patterns

Table 8 lists verbatim the regular expressions that produce the *any metric named* dimension and every per-metric count in this survey (§4). They are applied case-insensitively to the concatenated title and abstract (abstract scan) and to the extracted full text (full-text scan); a paper is flagged for a metric family if any pattern in that family matches, and `any_metric` is the disjunction over all families. Several patterns bound the distance between the metric term and annotation vocabulary, which suppresses matches in related-work discussion of prior corpora. The patterns, the extraction notebook, and the corpus they produce are all released, so every count in Tables 2 and 11 can be regenerated.

The corpus-selection stage that precedes extraction, the Anthology query described in §3, is documented in prose there; its search script is not part of the release.

#	Pattern (case-insensitive)
Cohen's κ	
1	`cohen.{0,8}kappa`
2	`cohen.{0,5}k`
3	`weighted kappa`
4	`kappa coefficient`
5	`kappa statistic`
6	`inter.?annotator kappa`
7	`k\s*=\s*0\.[0-9]`
8	`kappa\s*=\s*0\.[0-9]`
9	`cohen.{0,15}1960`
Fleiss' κ	
10	`fleiss.{0,8}kappa`
11	`fleiss.{0,5}k`
12	`multi.?rater kappa`
13	`fleiss.{0,15}1971`
Krippendorff's α	
14	`krippendorff`
15	`krippendorf`
16	`a\s*=\s*0\.[0-9].{0,20}(agreement|reliab)`
17	`reliability alpha`
Gamma (γ)	
18	`\bgamma\b.{0,50}(annotator|agreement|iaa|unitiz|mathet)`
19	`(agreement|iaa|unitiz).{0,50}\bgamma\b`
20	`mathet.{0,20}(2015|gamma)`
21	`pygamma`
22	`g.measure`
Percentage agreement	
23	`percentage agreement`
24	`percent.{0,6}agreement`
25	`raw agreement`
26	`observed agreement`
27	`proportion.{0,10}agree`
28	`pairwise agreement`
Span F1	
29	`\bf[_\-]?1\b.{0,50}(span|overlap|entity|ner\b|srl\b|iaa)`
30	`(span|overlap|entity|ner\b|srl\b).{0,50}\bf[_\-]?1\b`
31	`exact match.{0,20}(annotator|agreement)`

#	Pattern (case-insensitive)
ICC	
32	`\bicc\b`
33	`intraclass correlation`
34	`intra.class correlation`
35	`shrout.{0,20}fleiss`
36	`icc\s*\([1-3],[1-2]\)`
Gwet's AC1/AC2	
37	`gwet`
38	`\bac1\b.{0,30}(agreement|reliab)`
39	`\bac2\b.{0,30}(agreement|reliab)`
40	`kappa.paradox`
Correlation	
41	`spearman.{0,15}(rho|r).{0,30}(agree|reliab|annotati)`
42	`kendall.{0,15}(tau|t).{0,30}(agree|reliab|annotati)`
43	`pearson.{0,15}(r\b|correlation).{0,30}(agree|annotator|iaa)`
Scott's π	
44	`scott.{0,8}(pi|p)\b`
45	`scott.{0,15}reliability`
MACE	
46	`\bmace\b.{0,30}(annotati|reliab|agreement)`
47	`multi.annotator competence`
48	`hovy.{0,10}(2013|mace)`
Dawid–Skene	
49	`dawid.{0,10}skene`
50	`\bEM\b.{0,15}annotator`
51	`expectation.maximization.{0,20}annotati`
IRT / Rasch	
52	`\birt\b.{0,30}(annotati|reliab|agreement)`
53	`item response theory.{0,30}annotati`
54	`rasch.{0,20}annotati`
WindowDiff	
55	`windowdiff`
56	`window.diff`
57	`p_k\b.{0,20}(segmentation|agreement)`
CrowdTruth	
58	`crowdtruth`
59	`crowd.truth`
60	`aroyo.{0,15}(2014|2015)`
61	`cosine.agreement.{0,20}annotati`

Table 8: The 61 metric-extraction regular expressions, grouped into the 15 metric families reported in Table 2, quoted verbatim from the released extraction notebook. Greek letters appear italicised here for typesetting; the released file preserves the literal characters.

C Supplementary Tables and Figures

Table 9 lists the dimensions coded for each corpus paper (§3); Figure 2 shows the corpus by year and venue; Table 10 is the metric-appropriateness framework underlying the misuse analysis of §12.1.

D Metric × Section Cross-Tabulation

Table 11 provides per-section metric counts from the full-text scan, making all claims in the main text (including the 34% Cohen's κ rate in §11 and the 16% CrowdTruth rate in §10) directly verifiable.

E IAA Metric Overview

Table 12 is a quick-reference companion to the metric descriptions in §2. For each coefficient it records the data type the metric supports, whether it accommodates more than two raters, whether it tolerates missing annotations, and whether it applies a chance correction, grouped into categorical, structured/span-based, and continuous families.

F IAA Metric × Task-Type Matrix

Table 13 cross-references each IAA coefficient against the annotation task types of the eight-section taxonomy. For every (metric, task) pair it records whether the metric is a recommended primary choice, a usable secondary measure, or rarely appropriate for that task, operationalising the metric-appropriateness analysis of §12.1 as a practitioner look-up table.

G NMF Topic Top Terms

Table 14 shows the six highest-weighted terms for each of the 16 NMF topics and their mapping to survey sections. This table allows readers to verify the topic-to-section assignment and identify potential confusion cases.

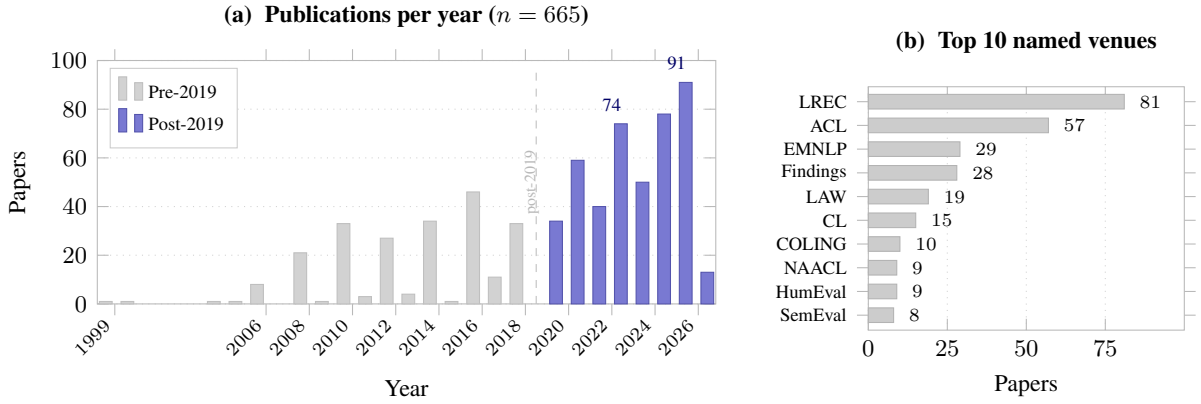


Figure 2: Corpus statistics for the 665-paper ACL Anthology corpus. **(a)** Annual publication counts. The post-2019 surge comprises 439 papers (66%) of the corpus, peaking at n=91 in 2025. **(b)** Top 10 *named* venues; papers indexed only under the Anthology's generic "Workshop" grouping are excluded from this panel. LREC/LREC-COLING accounts for 81 papers; ACL contributes 57; the Linguistic Annotation Workshop (LAW) contributes 19 dedicated papers.

Dimension	Values / description
Year	Integer (1999–2026)
Venue	ACL, EMNLP, LREC, LAW, Findings, Workshop, …
Paper type	Empirical, Corpus, Disagreement study, LLM annotation, Metric proposal, Methodology
Survey section	§4–§11 (NMF k=16 topic assignment)
Metric (abstract)	Cohen κ, Fleiss κ, Krippendorff α, …, None
Metric (full text)	Same + ICC, Gwet, Scott π, MACE, Dawid–Skene, …
CI reported	Binary: yes / no
IAA value	Numeric: reported coefficient value if stated
Annotator count	Numeric: n annotators if stated
Gold released	Binary: individual labels released beyond adjudicated gold

Table 9: Annotation dimensions used to characterise each corpus paper. NMF topic assignments were verified by the author through iterative review of topic coherence and keyword profiles.

Setting	Appropriate	Problematic use observed
2 annotators, nominal	Cohen κ	Raw % only
>2 annotators, nom.	Fleiss κ, α	Cohen κ (pairwise)
Ordinal labels	Weighted κ	Unweighted κ
Span / unitizing	γ, Span F1	Fixed-span κ only
Continuous ratings	ICC (specify type)	Cohen κ
Ranking tasks	Kendall τ, Spearman ρ	Nominal κ
Subjective / crowd	Distributional + MACE	Adjudicated gold only
LLM as annotator	Human vs. LLM κ/α	Raw accuracy

Table 10: Metric appropriateness by annotation setting. Column 3 lists patterns observed in the corpus where a less-appropriate metric was used. Frequencies are quantified in Table 3.

Section	Cohen κ	Kripp. α	Fleiss κ	% agree	Span F1	Crowd-Truth	MACE	ICC	Any metric	*N*
§4 IAA Metrics	28 (60%)	19 (40%)	12 (26%)	9 (19%)	1 (2%)	1 (2%)	0 (0%)	3 (6%)	33 (70%)	47
§5 Corpus Construction	45 (43%)	23 (22%)	15 (14%)	12 (12%)	6 (6%)	9 (9%)	7 (7%)	4 (4%)	68 (65%)	104
§6 Sequence Labelling	13 (23%)	5 (9%)	8 (14%)	3 (5%)	9 (16%)	0 (0%)	0 (0%)	0 (0%)	28 (49%)	57
§7 Semantic Annotation	52 (40%)	30 (23%)	15 (12%)	24 (18%)	7 (5%)	2 (2%)	3 (2%)	2 (2%)	78 (60%)	130
§8 Temporal	19 (32%)	3 (5%)	11 (19%)	2 (3%)	6 (10%)	1 (2%)	0 (0%)	2 (3%)	36 (61%)	59
§9 Sentiment & Hate Speech	42 (45%)	29 (31%)	30 (32%)	12 (13%)	2 (2%)	4 (4%)	5 (5%)	1 (1%)	72 (77%)	94
§10 Disagreement & LLMs	43 (35%)	33 (27%)	16 (13%)	11 (9%)	4 (3%)	20 (16%)	12 (10%)	3 (2%)	93 (75%)	124
§11 MT Evaluation	17 (34%)	15 (30%)	7 (14%)	4 (8%)	1 (2%)	0 (0%)	0 (0%)	1 (2%)	31 (62%)	50
Total (corpus)	**259 (38.9%)**	**157 (23.6%)**	**114 (17.1%)**	**77 (11.6%)**	**36 (5.4%)**	**37 (5.6%)**	**27 (4.1%)**	**16 (2.4%)**	**439 (66.0%)**	**665**

Table 11: Full-text metric counts by survey section (N = papers with a given metric named anywhere in the full text; % = fraction of section total). Counts are lower bounds (549/665 papers parsed; see Table 2 footnote). Notable patterns: Cohen's κ is used in 34% of §11 MT papers despite ICC being the appropriate choice for continuous DA scores; CrowdTruth and MACE are concentrated in §10, confirming their role as disagreement-modelling tools.

Metric	Data type	≥2 raters	Missing data?	Chance-corr.?
Categorical				
% agreement	Nominal	pair	–	–
Bennett's S	Nominal	2+	–	✓
Cohen's κ	Nominal	2	–	✓
Weighted κ	Ordinal	2	–	✓
Fleiss' κ	Nominal	3+	–	✓
Krippendorff α	Any	2+	✓	✓
Gwet AC1/AC2	Nom./Ord.	2+	✓	✓
Scott's π	Nominal	2	–	✓
Structured / span-based				
Span F1	Span	pair	–	–
P_k / WindowDiff	Segmentation	pair	–	–
Gamma (γ)	Unitising	2+	✓	✓
Continuous / interval-scale				
ICC	Continuous	2+	✓	✓
Cronbach α	Continuous	2+	✓	✓
Spearman ρ / Kendall τ	Ordinal	pair	–	–
Pearson r	Continuous	pair	–	–

Table 12: IAA metric overview. Adapted from James (2026).

Metric	Data type	Survey section / NLP task type							
		§4 IAA Theory n=47	§5 Corpus Constr. n=104	§6 NER & Treebanks n=57	§7 Disc. / SRL Lexical n=130	§8 Temporal Events n=59	§9 Sent. & Hate n=94	§10 Disagree. LLMs n=124	§11 MT Eval. n=50
Categorical data									
Percentage agreement	Nominal	○	○	○	○	○	●	●	·
Bennett's S	Nominal	○	·	·	·	·	·	·	·
Cohen's κ	Nominal	●	●	●	●	●	●	●	○
Weighted κ	Ordinal	○	·	·	·	·	○	○	·
Fleiss' κ	Nominal	○	○	·	○	○	○	●	·
Krippendorff's α	Any	●	○	·	○	○	○	○	○
Gwet's AC1/AC2	Nom. / Ord.	○	·	·	·	·	○	·	·
Scott's π	Nominal	○	·	·	·	·	·	·	·
Structured and span-based data									
Span F1 / Dice	Span	·	○	●	●	○	·	·	·
LAS (treebanks)	Syntactic	·	○	●	·	·	·	·	·
P_k / WindowDiff	Segmentation	○	·	·	●	·	·	·	·
Boundary Edit Dist.	Segmentation	○	·	·	○	·	·	·	·
Gamma (γ)	Unitising	●	○	·	●	·	·	·	·
Continuous and interval-scale data									
ICC[†]	Continuous	○	·	·	·	·	○	·	●
Cronbach's α	Continuous	○	·	·	·	·	·	·	○
CCC[‡]	Continuous	○	·	·	·	·	○	·	·
Spearman ρ / Kendall τ	Ordinal	○	·	·	·	·	○	○	●
Pearson r	Continuous	○	·	·	·	·	○	·	○
Probabilistic and aggregation models									
MACE (Hovy et al., 2013)	Nominal	·	·	·	·	·	○	●	·
CrowdTruth (Aroyo and Welty, 2015)	Nominal	·	·	·	·	·	○	●	·
IRT-based measures	Nominal	○	·	·	·	·	·	●	·

Symbols: ● primary (commonly used / expected for this task); ○ secondary (attested but task-atypical); · absent from corpus for this task. primary cell; secondary cell.

[†] ICC = Intraclass Correlation Coefficient (six variants; Shrout and Fleiss, 1979).

[‡] CCC = Concordance Correlation Coefficient; LAS = Labelled Attachment Score.

Table 13: **IAA metric suitability and observed usage across NLP annotation task types.** Section columns correspond to the eight survey dimensions (§4–§11); n gives the paper count per section. *Key empirical finding:* Cohen's κ appears as the primary metric across *all* categorical task columns, regardless of whether the task-specific assumptions hold, a one-size-fits-all reporting norm that this survey documents empirically for the first time across 665 papers (see §12.2). Probabilistic aggregation methods (MACE, CrowdTruth, IRT) are concentrated in §10, reflecting the emerging disagreement-as-signal paradigm. Qualitative coding (● primary, ○ secondary, · absent) reflects expert assessment of metric suitability per task type, validated against empirical usage counts in Appendix D. Darker shading ( ) marks primary cells; lighter shading marks secondary usage.

T	Section	N	Top 6 terms
0	§5	57	corpus, annotated, annotated corpus, corpora, guidelines, annotation
1	§5	47	natural, natural language, language, processing, language processing
2	§11	29	translation, machine translation, document level, document, evaluation
3	§10	60	disagreement, annotator disagreement, label, annotator variation
4	§9	36	hate, speech, hate speech, detection, social media, offensive
5	§7	46	sense, word sense, word, disambiguation, sense disambiguation
6	§6	36	entity, named, named entity, NER, entity recognition, recognition
7	§10	64	LLMs, LLM, human, models, language models, large language
8	§9	58	sentiment, sentiment analysis, analysis, dataset, opinion
9	§6	21	treebank, dependency, dependency treebank, parsing, syntactic
10	§4	14	bias, 2024, shared task, annotation quality, inter-annotator
11	§11	21	reproduction, original, human evaluation, MT, simplification
12	§7	29	discourse, relations, discourse relations, RST, coherence
13	§4	33	annotator agreement, inter annotator, agreement, metric
14	§7	55	annotation scheme, argument, argumentative, argument mining
15	§8	59	event, temporal, events, relation, extraction, time

Potential overlap

T0 vs T1	Both §5; T0 centres on annotated-corpus construction, T1 on broader language-resource papers
T8 vs T4	Both §9; T8 is opinion/SA, T4 is hate speech specifically
T10 vs T13	Both §4; T10 has recent bias/shared-task work, T13 IAA theory

Table 14: NMF (k=16) topic top terms and survey-section mapping. Overlap cases and confusion cases are noted.