Consistency is Key: Disentangling Label Variation in Natural Language Processing with Intra-Annotator Agreement

Gavin Abercrombie[1] and **Tanvi Dinkar**[1,] and **Amanda Cercas Curry**[2]
and **Verena Rieser**[1*] and **Dirk Hovy**[3]
[1]Heriot-Watt University [2]CENTAI Institute [3]Bocconi University
g.abercrombie@hw.ac.uk

Abstract

We commonly use agreement measures to assess the utility of judgements made by human annotators in Natural Language Processing (NLP) tasks. While *inter*-annotator agreement is frequently used as an indication of label *reliability* by measuring consistency between annotators, we argue for the additional use of *intra*-annotator agreement to measure label *stability* (and annotator *consistency*) over time. However, in a systematic review, we find that the latter is rarely reported in this field. Calculating these measures can act as important quality control and could provide insights into *why* annotators disagree. We conduct exploratory annotation experiments to investigate the relationships between these measures and perceptions of subjectivity and ambiguity in text items, finding that annotators provide inconsistent responses around 25% of the time across four different NLP tasks.

1 Introduction

Agreement measures are commonly used to assess the utility of judgements made by human annotators for Natural Language Processing (NLP) tasks. Indeed, the reporting of *inter*-annotator agreement (or inter-rater reliability) has long been the standard to indicate dataset quality (Carletta, 1996) and frequently serves as an upper bound for model performance on a task (Boguslav and Cohen, 2017).

While inter-annotator agreement is frequently used in NLP to determine the *reliability* of labels or the processes used to produce them (Artstein, 2017), *intra*-annotator agreement is rarely, if ever, reported. However, we can use it to measure the temporal *consistency* of the annotators who chose the labels and, hence, the *stability* of the labels and data that they generate.[1] Consistency and label stability are important because, without them, annotation schemes are unlikely to be repeatable or reproducible (Teufel et al., 1999).[2]

Such measures of intra-rater agreement are frequently reported in areas of medicine such as physiotherapy (e.g. Bennell et al., 1998; Meseguer-Henarejos et al., 2018), and speech pathology (e.g. Capilouto et al., 2005; Rose and Douglas, 2003). Intra-rater measures are also reported in other fields as diverse as economics (Hodgson, 2008), software engineering (Grimstad and Jørgensen, 2007), and psychology (Ashton, 2000).

However, reporting intra-annotator agreement is so far extremely uncommon in NLP, as we show in a systematic review in Section 2.

Disagreement and label variation in NLP In addition, we argue that the use of inter- and intra-annotator agreement allows us to distinguish and measure different sources of observed label variation (Rottger et al., 2022; Plank, 2022). This is important as NLP researchers have increasingly recognised that, for many tasks, different points of view may be equally valid (Aroyo and Welty, 2015; Basile et al., 2021a; Plank, 2022; Rottger et al., 2022), and that their aggregation can erase minority perspectives (Basile et al., 2021a; Blodgett, 2021).

One of the main challenges in implementing this new paradigm is the interpretation of disagreement. Disagreement between annotators may be due to two sources: 1) genuine differences in their subjective beliefs/perspectives, which can be desirable under this paradigm, or 2) task difficulty, ambiguity, or annotator error, all of which are undesirable. While agreement measures *between* annotators can give us an idea of task **subjectivity**, they provide little insight as to its **difficulty**, **ambiguity,** or the quality and attentiveness of the annotators themselves (Rottger et al., 2022).

*Now at Google DeepMind

[1]We apply the term *consistency* to annotator behaviour and *stability* to labels and datasets.

[2]Although there may be situations in which annotation consistency is not expected, such as longitudinal studies of attitudinal change.

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 63–74
November 8, 2025

In the following, we propose the use of *intra*-annotator agreement as a measure of subjectivity.

The reliability-stability agreement matrix What then, does it mean when individual annotators' interpretations are not stable, i.e., internally inconsistent? In addition to providing an additional layer of quality control, we suggest that measurement of label stability can help to interpret potential causes of *inter*-annotator disagreement. To this end, we propose the reliability-stability matrix, a framework for mapping and interpreting the relationship between inter- and intra-annotator agreement in labelled datasets (Table 1).

		Reliability (between annotators)	
		Low *inter*	High *inter*
Stability (temporal within annotator)	High *intra*	Variable perspectives/ High subjectivity	Straight-forward/ Good quality
	Low *intra*	Ambiguous or difficult/ Poor quality	Systematic errors/ Value changes

Table 1: The reliability-stability matrix for *inter*- and *intra*- annotator agreement.

Under this framework, *inter*-annotator agreement and *intra*-annotator agreement, taken together, indicate the task's ambiguity or complexity and its subjectivity level. *Inter*-annotator agreement measures reliability, while *intra*-annotator agreement measures stability. The resulting axes form a confusion matrix that describes four cases.

If both measures are high, we assume the task is unambiguous and simple, and the annotator group relatively homogonous. Presumably, the quality of the guidelines and textual data is also good (Ide and Pustejovsky, 2017). In this scenario, the task or item should be relatively straightforward.

Where both agreement measures are low, we are likely to be faced with a highly ambiguous or difficult task or item–perhaps with multiple equally valid responses–or the annotation quality is poor.

If reliability is low, but consistency is high, the labels likely reflect the annotators' varied but potentially equally valid subjective perspectives.

We do not foresee many situations where reliability is high yet stability/consistency is low. Any agreement between inconsistent annotators would presumably be purely by chance or mass random spamming, i.e., systematic errors. Exceptions could include population-level value shifts over longer time intervals arising from awareness-raising events such as the #MeToo (Szekeres et al., 2020) and #BLM (Sawyer and Gampa, 2018) movements.

Our framework can be applied at the dataset- or item-level by computing any standard agreement metrics.We illustrate this in exploratory annotation experiments described in Section 3.

Our contributions 1) We conduct a systematic review, finding that **a tiny fraction of NLP publications report intra-annotator agreement**; (2) we suggest addition of intra-annotator agreement as a standard measure, and show how **measuring annotator stability could complement existing reliability measures to distinguish reasons for label variation**; and (3) we conduct exploratory longitudinal annotation experiments across four NLP tasks, finding that **annotators provide inconsistent responses for more than 25% of items**, calling into question the implicit assumption that differences in annotation behaviour are seen only between and not within individuals.[3]

2 Intra-Annotator Agreement in the NLP Community

To get a snapshot of the extent to which intra-annotator agreement is reported in the NLP community, we conducted a systematic review of papers published in the *Anthology* of the Association for Computational Linguistics (ACL).[4] Here, we wish to discover for which tasks and what purposes NLP researchers collect and report on repeat annotations and evidence for how and when repeat items should be presented to annotators. Full details of the review methodology are available in Appendix A.

To what extent and why is intra-annotator agreement reported in NLP? When we conducted our study, the search and filtering process returned only 56 relevant publications out of more than 80,000 papers listed in the Anthology. In other words, a tiny fraction (less than 0.07%) of computational linguistics and NLP publications in the repository report measurement of intra-annotator agreement.[5]

The only area of NLP in which intra-annotator agreement is somewhat regularly reported is machine translation (MT), which accounts for

[3] Data available at `https://github.com/HWU-NLP/consistency`.

[4] `https://aclanthology.org/`

[5] We acknowledge that intra-annotator agreement is irrelevant to many papers, but highlight that the number of publications which report it is nevertheless extremely low.

more than half of the included publications. Most of these were agreement measures on human evaluation of translation quality, with one on word alignment annotation for MT (Li et al., 2010). Several other publications on evaluating natural language generation also report measurement on human evaluation tasks (e.g. Belz and Kow, 2011; Belz et al., 2016, 2018; Jovanovic et al., 2005). Other included fields are semantics (e.g. Cao et al., 2022; Hengchen and Tahmasebi, 2021), syntax (e.g. Baldridge and Palmer, 2009; Lameris and Stymne, 2021), affective computing (including sentiment analysis (Kiritchenko and Mohammad, 2017) and emotion detection (Vaassen and Daelemans, 2011)), and automatic text grading (Cleuren et al., 2008; Downey et al., 2011). There is also one paper on abusive language detection (Cercas Curry et al., 2021).[6]

Where the authors motivate the collection of repeat annotations, they usually mention quality control or annotator consistency. Notably, no papers mention the possibility that intra-annotator inconsistency could be valid or informative beyond these factors, as we propose.

Best practice for measuring intra-annotator agreement: how long should the label-relabel interval be? When designing annotation tasks (such as ours in Section 3), it would be helpful to know when to present repeated items, thus avoiding annotators labelling from memory, which may not be an actual test of their consistency.

Over a quarter of the papers (15/56) do not provide enough information to determine the interval between initial and repeat annotations. In most other cases, either it can be inferred, or the authors explicitly state that re-annotations are conducted in the same session as the original annotation. Those that report more extended time before re-annotation leave intervals varying from a few minutes (Kiritchenko and Mohammad, 2017) to a year (Cleuren et al., 2008; Hamon, 2010).

Two papers do specifically investigate the effects of time on annotator consistency. Li et al. (2010) experimented with intervals of one week, two weeks, and one month, comparing intra-annotator agreement for these and finding that consistency on their word alignment annotation degraded steadily over time. Kiritchenko and Mohammad (2017) performed a similar study, comparing intra-annotator agreement on ratings (on a scale) that were conducted with intervals from a few minutes to a few days between the initial and repeat judgements. They too found that inconsistencies increased as a function of increase in interval.

[6]We provide a full list of included papers in Appendix B.

3 Exploratory annotation experiments

We conduct an exploratory annotation experiment to investigate the relationships between agreement measures and the possible reasons for disagreements and inconsistencies. We also investigate whether, as is commonly believed, specific task types are generally more subjective than others.

Hypotheses

At the individual annotation item level, for a given task and dataset:

H1.1 *Subjective* annotation items have lower *inter*-annotator agreement than *straightforward* items, but higher *intra*-annotator agreement than *ambiguous* items.

H1.2 *Ambiguous* annotation items have lower *inter*- and *intra*-annotator agreement than both *straightforward* and *ambiguous* items.

At the dataset/task level:

H2 *Social* tasks—such as offensive language detection and sentiment analysis—are more *subjective* than *linguistic* tasks, like textual entailment or anaphora resolution. That is, stability is higher for social tasks than linguistic tasks.

	Task	Dataset	Labels
Social	Offensive language detection	Leonardelli et al. (2021)	*Offensive/not offensive*
	Sentiment analysis	Kenyon-Dean et al. (2018)	*Positive/negative/objective*
Linguistic	Natural language inference/ textual entailment	Williams et al. (2018)	*Entailment/contradiction/ neutral*
	Anaphora resolution	Poesio et al. (2019)	*Referring/non-referring*

Table 2: Datasets used in the annotation experiments.

Data We use subsets of four English language datasets, see Table 2: two social tasks that are commonly assumed to be subjective, and two linguistic tasks, thought of as objective (Basile et al., 2021b). These were selected because they (1) have limited label sets (of two or three classes), allowing for comparison across tasks; and (2) have been published with non-aggregated (i.e. annotator specific) labels, allowing us to include items with known inter-annotator disagreement in our subsamples. From each dataset, we selected 50 items with high disagreement in the original label sets for re-annotation.

Methodology We recruited crowdworkers from Prolific[7] to annotate a subset of fifty items from each of the tasks/datasets. As much of the text data is primarily sourced from the United States of America and, in some cases,[8] concerns American news stories such as the controversy surrounding the killing of George Floyd,[9] we recruited only annotators located in the US. To obtain high quality annotations, we prescreened participants to ensure that (1) their first language was English, and that (2) they had a 100% approval rate on Prolific.

Based on the evidence of our review (Li et al., 2010; Kiritchenko and Mohammad, 2017), and of more recent work by Abercrombie et al. (2023), we left an interval of two weeks before we recall the annotators to collect a second round of annotations in order to measure their consistency. Of 30 annotators that began the first task, 16 completed both rounds of all four tasks, and we base our results on the labels they provided. All annotators were L1 English speakers; nine were male and eight female; 11 identified as 'White', four as 'Black', one 'Asian', and one 'Mixed'; and ages ranged from 20 to 67; ($\mu = 43.9; s = 14.0$). Annotators were provided with the original instructions pertaining to each task.

We then recruited a second set of expert annotators to annotate the examples that demonstrate internal and or external disagreement with rationalisations for these disagreements, using the labels *ambiguous*. *subjective*, or *straightforward*.

4 Results

We report agreement for each task, and examine differences between the groups of items labelled as *subjective*, *ambiguous*, and *straightforward*.

Overall agreement As *intra*-annotator agreement is typically assumed to be 100% (i.e. by omitting to consider it (Abercrombie et al., 2023)), we measure and raw report percentage agreement as a primary metric to examine whether this holds. For *inter*-annotator agreement, we calculate these pairwise across annotators and report the means. For completeness, we also report Cohen's kappa scores in Appendix C.

	Reliability (*Inter-*) %		**Stability** (*Intra-*) %	
	μ	σ	μ	σ
Offence	68.3	15.4	74.4	15.0
Sentiment	63.6	21.7	69.2	19.5
Entailment	58.6	21.4	72.6	15.1
Anaphora	76.2	14.3	80.5	13.0
Overall	66.7	19.6	74.2	16.3

Table 3: Pairwise reliability and stability of the collected labels measured with mean (μ) and standard deviations (σ) across items for raw percentage inter- and intra-annotator agreement scores.

Agreement scores are presented in Table 8. As expected, agreement is higher for stability than reliability for all tasks, although considerably lower than perfect agreement—just 74.2% overall, and no higher than 80.5% for any task. Individual annotators all have very similar levels of stability: $\mu = 74.2\%; \sigma = 4.3\%; max = 81.5\%; min = 67.5\%$. These results are also remarkably similar to those of Abercrombie et al. (2023), who reported mean intra-annotator agreement of 74.5% on a hate speech identification task conducted over a comparable time frame and on the same recruitment and annotation platforms.

Agreement by task The distribution of annotation items on the reliability-stability matrix is shown in Figure 1. A multivariate Kruskal-Wallis test indicates statistically significant differences between tasks for both variables: for inter-annotator agreement, $H{-}statistic : 12.42, p{-}value : 0.01$; and for intra-annotator agreement, $H - statistic : 10.76, p - value : 0.01$.[10]

[7]https://www.prolific.co/

[8]Particularly in the offensive language dataset.

[9]*The Guardian* April 20 2021 (McGreal, 2021).

[10]Post-hoc pairwise Dunn's tests with Bonferroni correction reveal that only *sentiment-anaphora* and *entailment anaphora* have significantly different distributions for reliability, and only *sentiment-anaphora* for stability.

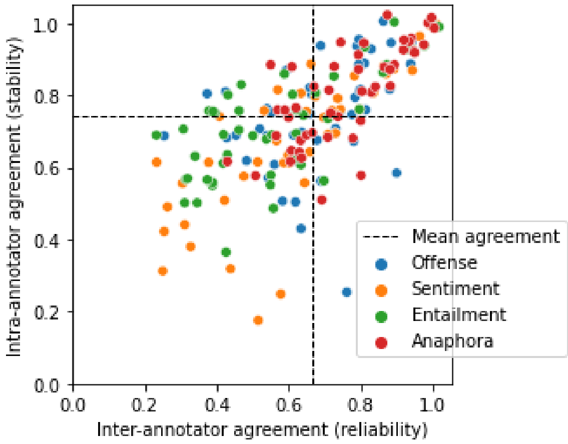

Figure 1: By task raw percentage agreement on individual items for reliability (pairwise) and stability.

However, these differences do not confirm the view that social tasks are more subjective than linguistic tasks (H2). Rather, the *offense* and *anaphora* tasks obtain higher agreement (both *inter* and *intra*) than the *sentiment* and *entailment tasks*, suggesting that, for the particular items in these data samples, the former are simply easier to agree and be consistent on than the latter.

	Bottom-left (*Amb.*)	Top-left (*Subj.*)	Top-right (*Straight.*)	Bottom-right (*Errors*)
Offense	30.0	18.0	38.0	14.0
Sentiment	46.0	10.0	38.0	6.0
Entailment	48.0	20.0	28.0	4.0
Anaphora	22.0	10.0	56.0	12.0
Overall	36.5	14.5	40.0	9.0

Table 4: Percentage of annotation items in each quadrant of the plot in Figure 1.

As Figure 1 and Table 4 show, while the annotation items are predominantly distributed across the bottom-left and top-right quadrants, *sentiment* and *entailment* are skewed to the bottom left, indicating greater ambiguity, and offensive language and entailment tend towards the top-right (*subjectivity*). With 68% of items on the left-hand side, *entailment* is the least, and *anaphora*, with 56% in the top-right, the most *straightforward* task.

Anaphora resolution seems to be the most *straightforward* task, with most items in the upper-right quadrant, while sentiment analysis and entailment are the most ambiguous/difficult, both having almost 50% of examples fall in the bottom left quadrant. As expected, the lowest number of items fall in the bottom right section of the plot.

Rationalisation In an attempt to validate the reliability-stability matrix and to test H1.1 and H1.2, rationalisation labels were applied by two postdoctoral researchers with backgrounds in NLP and computational linguistics. They were asked to read the annotation instructions and items and provide each example with a label: *subjective*, *ambiguous*, or *straightforward*. Disagreements were resolved by discussion between these and a third author. Inter-annotator agreement (before resolution) is shown in Table 5, indicating that this in itself was a very difficult task to reach agreement on.

Offence	Sentiment	Entailment	Anaphora
0.26	0.11	0.47	0.02

Table 5: Inter-annotator agreement on the rationalisation labelling task, measured with Cohen's *kappa*.

To quantitatively examine the relationship between the perceived reason for agreement/disagreement and the reliability and stability measurements, we applied a multivariate Kruskal-Wallis test to the independent categorical variable *rationale* (*straightforward*, *subjective*, and *ambiguous*) and the two dependent continuous variables *inter-* and *intra-annotator agreement*.

The test showed that there is only a very small and non-significant difference in the dependent vectors between the different groups, with an H-statistic of $2.734, p = 0.26$, indicating that the assigned rationale labels do not explain the inter- and intra annotator agreement rates.

5 Discussion and conclusion

We have examined the role and use of intra-annotator agreement measures in NLP research. Calculation of such measures can act as an important quality control and could potentially provide insights into the reasons for disagreements between annotators. However, in a systematic review, we found that they are rarely reported in this field.

We have proposed a framework for the interpretation of inter- and intra-annotator agreement, the *reliabilty-stability agreement matrix*. Exploratory annotation experiments failed to validate our theory that this framework can be used to tease apart subjectivity and ambiguity, and it proved to be very hard to recognise or agree on these, even for trained annotators. However, we have shown how comparing both *inter-* and *intra-* annotator agreement enables quantification of the difficulty of particular tasks and/or annotation items. Strikingly, we found that, across four different tasks, crowdsourced annotators were consistently inconsistent, calling into question the implicit assumption that labels provided by individual annotators are stable, and reinforcing the need to collect within-annotator labels for NLP tasks, including those typically considered to be ‘objective’.

Limitations

We acknowledge that the scope of our exploratory experiments is quite small at 50 items per task and 16 annotators, and that larger studies may produce different results. While we took some measures to ensure the quality of recruited annotators (section 3), there are known issues with crowdworker quality for annotation (e.g. Hovy et al., 2013; Weber-Genzel et al., 2024), and some annotator inconsistency may due to inattention—another factor that should be considered and further reason to measure and report intra-annotator agreement.

Ethical considerations

Because we recruit humans to work on data labelling, we obtained approval to undertake this study from the Institutional Review Board (IRB) of the School of Mathematics & Computer Science at Heriot-Watt University, reference 2023-4926-7368. Additionally, we took the following measures:

Compensation We paid the annotators above the Living Wage in our jurisdiction (higher than the legal minimum wage, as recommended (as a minimum) by Shmueli et al. (2021).

Welfare As some of the data to be labelled included offensive language, we:

- avoided recruiting members of vulnerable groups by restricting annotators to those aged over 18, provided them with comprehensive warnings prior to consenting to participate, and asked them to self-declare that they would not be adversely affected by participating;
- allowed annotators to leave the study at any time and informed them that they would be paid for their time regardless;
- kept the annotation task short to avoid lengthy exposure to material which may exceed ‘*minimal risk*’ (Shmueli et al., 2021).

Privacy All personal data of recruited annotators was collected anonymously.

Acknowledgements

We would like to thank the reviewers for their insightful comments which we have tried to incorporate into this version of the paper.

Gavin Abercrombie and Tanvi Dinkar were supported by the EPSRC project ‘Equally Safe Online’ (EP/W025493/1). Dirk Hovy was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR). He is a member of the MilaNLP group and the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA).

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Abdulrahman Alosaimy and Eric Atwell. 2018. Web-based annotation tool for inflectional language resources. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. 2017. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, Dordrecht.

Robert H. Ashton. 2000. A review and analysis of research on the test–retest reliability of professional judgment. *Journal of Behavioral Decision Making*, 13(3):277–294.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021a. Toward a perspectivist turn in ground truthing for predictive computing. In *Conference of the Italian Chapter of the Association for Intelligent Systems (ItAIS 2021)*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021b. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235, Portland, Oregon, USA. Association for Computational Linguistics.

Anja Belz, Adrian Muscat, Pierre Anguill, Mouhamadou Sow, Gaétan Vincent, and Yassine Zinessabah. 2018. SpatialVOC2K: A multilingual dataset of images with annotations and features for spatial relations between objects. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 140–145, Tilburg University, The Netherlands. Association for Computational Linguistics.

Anja Belz, Adrian Muscat, Brandon Birmingham, Jessie Levacher, Julie Pain, and Adam Quinquenel. 2016. Effect of data annotation, feature selection and model choice on spatial description generation in French. In *Proceedings of the 9th International Natural Language Generation conference*, pages 237–241, Edinburgh, UK. Association for Computational Linguistics.

Kim Bennell, Richard Talbot, Henry Wajswelner, Wassana Techovanich, David Kelly, and AJ Hall. 1998. Intra-rater and inter-rater reliability of a weight-bearing lunge measure of ankle dorsiflexion. *Australian Journal of Physiotherapy*, 44(3):175–180.

Luisa Bentivogli, Marcello Federico, Giovanni Moretti, and Michael Paul. 2011. Getting expert quality from the crowd for machine translation evaluation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Ph.D. thesis, University of Massachusetts Amherst.

Mayla Boguslav and Kevin Bretonnel Cohen. 2017. Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing. *Studies in health technology and informatics*, 245:298–302.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics-MATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Yang Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Gilson Capilouto, Heather Harris Wright, and Stacy A. Wagovich. 2005. CIU and main event analyses of the structured discourse of older and younger adults. *Journal of Communication Disorders*, 38(6):431–444.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leen Cleuren, Jacques Duchateau, Pol Ghesquière, and Hugo Van hamme. 2008. Children's oral reading corpus (CHOREC): Description and assessment of annotator agreement. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 67–78, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Ryan Downey, David Rubin, Jian Cheng, and Jared Bernstein. 2011. Performance of automated scoring for children's oral reading. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 46–55, Portland, Oregon. Association for Computational Linguistics.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Annemarie Friedrich and Alexis Palmer. 2014. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop* *and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Stein Grimstad and Magne Jørgensen. 2007. Inconsistency of expert judgment-based estimates of software development effort. *Journal of Systems and Software*, 80(11):1770–1777.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, and Edward Gillian. 2015. Human evaluation of grammatical error correction systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 461–470, Lisbon, Portugal. Association for Computational Linguistics.

Olivier Hamon. 2010. Is my judge a good one? In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Yifan He, Yanjun Ma, Johann Roturier, Andy Way, and Josef van Genabith. 2010. Improving the post-editing experience using translation recommendation: A user study. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Simon Hengchen and Nina Tahmasebi. 2021. SuperSim: a test set for word similarity and relatedness in Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 268–275, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Aurelie Herbelot and Ann Copestake. 2010. Annotating underquantification. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 73–81, Uppsala, Sweden. Association for Computational Linguistics.

Limor Hochberg, Cecilia Ovesdotter Alm, Esa M. Rantanen, Caroline M. DeLong, and Anne Haake. 2014a. Decision style in a clinical reasoning corpus. In *Proceedings of BioNLP 2014*, pages 83–87, Baltimore, Maryland. Association for Computational Linguistics.

Limor Hochberg, Cecilia Ovesdotter Alm, Esa M. Rantanen, Qi Yu, Caroline M. DeLong, and Anne Haake. 2014b. Towards automatic annotation of clinical decision-making style. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 129–138, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Robert T. Hodgson. 2008. An examination of judge reliability at a major U.S. wine competition. *Journal of Wine Economics*, 3(2):105–113.

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, Atlanta, Georgia. Association for Computational Linguistics.

Nancy Ide and James Pustejovsky. 2017. *Handbook of linguistic annotation*, volume 1. Springer.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2005. A corpus for studying addressing behavior in multi-party dialogues. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, pages 107–116, Lisbon, Portugal. Special Interest Group on Discourse and Dialogue (SIGdial).

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 135–144, Lisboa, Portugal. European Association for Machine Translation.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia. Association for Computational Linguistics.

Ivana Kruijff-Korbayová, Klára Chvátalová, and Oana Postolache. 2006. Annotation guidelines for Czech-English word alignment. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Harm Lameris and Sara Stymne. 2021. Whit's the richt pair o speech: PoS tagging for Scots. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–48, Kiyv, Ukraine. Association for Computational Linguistics.

Samuel Läubli, Mark Fishel, Manuela Weibel, and Martin Volk. 2013. Statistical machine translation for automobile marketing texts. In *Proceedings of Machine Translation Summit XIV: Posters*, Nice, France.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, and Kazuaki Maeda. 2010. Enriching word alignment with linguistic tags. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Wanqiu Long, Xinyi Cai, James Reid, Bonnie Webber, and Deyi Xiong. 2020. Shallow discourse annotation for Chinese TED talks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1025–1032, Marseille, France. European Language Resources Association.

Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Rui Li, Jeff B. Pelz, Pengcheng Shi, and Anne Haake. 2012. Annotation schemes to encode domain knowledge in medical narratives. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 95–103, Jeju, Republic of Korea. Association for Computational Linguistics.

Chris McGreal. 2021. Derek Chauvin found guilty of George Floyd's murder. *The Guardian*.

Ana-Belén Meseguer-Henarejos, Julio Sánchez-Meca, José-Antonio López-Pina, and Ricardo Carles-Hernández. 2018. Inter- and intra-rater reliability of the Modified Ashworth Scale: a systematic review and meta-analysis. *European journal of physical and rehabilitation medicine*, 54(4):576—590.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and the PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *Annals of Internal Medicine*, 151(4):264–269.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of EMNLP*. ACL.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Miranda Rose and Jacinta Douglas. 2003. Limb apraxia, pantomine, and lexical gesture in aphasic speakers: Preliminary findings. *Aphasiology*, 17(5):453–464.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Dana Ruiter, Thomas Kleinbauer, Cristina España-Bonet, Josef van Genabith, and Dietrich Klakow. 2022. Exploiting social media content for self-supervised style transfer. In *Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media*, pages 11–34, Seattle, Washington. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, and Juan Antonio Pérez-Ortiz. 2012. Source-language dictionaries help non-expert users to enlarge target-language dictionaries for machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3422–3429, Istanbul, Turkey. European Language Resources Association (ELRA).

Jeremy Sawyer and Anup Gampa. 2018. Implicit and explicit racial attitudes changed during Black Lives Matter. *Personality and Social Psychology Bulletin*, 44.

Claudia Schulz, Christian M. Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych. 2019. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2761–2772, Florence, Italy. Association for Computational Linguistics.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3758–3769, Online. Association for Computational Linguistics.

Hanna Szekeres, Eric Shuman, and Tamar Saguy. 2020. Views of sexual assault following #MeToo: The role of gender and individual differences. *Personality and Individual Differences*, 166:110203.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Frederik Vaassen and Walter Daelemans. 2011. Automatic emotion classification for interpersonal communication. In *Proceedings of the 2nd Workshop on*

Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011), pages 104–110, Portland, Oregon. Association for Computational Linguistics.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Mihaela Vela and Josef van Genabith. 2015. Reassessing the WMT2013 human evaluation with professional translators trainees. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 161–168, Antalya, Turkey.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2020. Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online. Association for Computational Linguistics.

Chaojun Wang, Christian Hardmeier, and Rico Sennrich. 2021. Exploring the importance of source text in automatic post-editing for context-aware machine translation. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 326–335, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Lu Wang, Hema Raghavan, Claire Cardie, and Vittorio Castelli. 2014. Query-focused opinion summarization for user-generated content. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1660–1669, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Systematic review methodology

For this review, we followed the established systematic review guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement (Moher et al., 2009), as recommended by van Miltenburg et al. (2021):

1. Develop search query terms
2. Conduct search
3. Apply inclusion/exclusion criteria
4. Code included publications
5. Measure inter- and intra-annotator agreement (re-code subset of publications)
6. Synthesise results

The review covers all results retrieved from the Anthology's search facility. The searches were conducted on September 14 2022. Following retrieval of the resulting publications, we applied the inclusion/exclusion criteria shown in Table 6.

Include	**Exclude**
Human annotation studies	No human annotation study is conducted (e.g. surveys/reviews of other work)
Repeated annotations are collected	Repeated annotations are not collected
Intra-annotator measurement is reported	Intra-annotator measurement not reported
Measurement conducted on manual labels applied by human annotators	Labelling is performed automatically
'Intra-' refers to repeat annotations of the same items by the same annotator	Term 'intra-' is used, but refers to agreement measurements between different items and/or annotators
Publication is a full paper	Posters, proceedings, proposals, technical system descriptions etc.

Table 6: Criteria for in/exclusion in/from the review.

The searches returned 138 publications. After removing duplicates, and applying the inclusion criteria we were left with 56 relevant publications in the Anthology.

Publication	**NLP sub-field**	**Publication**	**NLP sub-field**
Akhbardeh et al. (2021)	Machine Translation	Graham et al. (2013)	Machine Translation
Alosaimy and Atwell (2018)	Syntax	Grundkiewicz et al. (2015)	Syntax
Baldridge and Palmer (2009)	Machine Translation	Hamon (2010)	Machine Translation
Belz and Kow (2011)	NLG	He et al. (2010)	Machine Translation
Belz et al. (2016)	NLG	Hengchen and Tahmasebi (2021)	Semantics
Belz et al. (2018)	NLG	Herbelot and Copestake (2010)	Semantics
Bentivogli et al. (2011)	Machine Translation	Hochberg et al. (2014a)	Cognitive psychology
Berka et al. (2011)	Machine Translation	Hochberg et al. (2014b)	Cognitive psychology
Bojar et al. (2013)	Machine Translation	Jovanovic et al. (2005)	NLG
Bojar et al. (2014)	Machine Translation	Kiritchenko and Mohammad (2017)	Affective computing
Bojar et al. (2015)	Machine Translation	Kreutzer et al. (2020)	Machine Translation
Bojar et al. (2016)	Machine Translation	Kreutzer et al. (2018)	Machine Translation
Bojar et al. (2017)	Machine Translation	Kruijff-Korbayová et al. (2006)	Machine Translationn
Bojar et al. (2018)	Machine Translation	Lameris and Stymne (2021)	Syntax
Bouamor et al. (2014)	Machine Translation	Läubli et al. (2013)	Machine Translation
Callison-Burch et al. (2007)	Machine Translation	Li et al. (2010)	Machine Translation
Callison-Burch et al. (2010)	Machine Translation	Long et al. (2020)	Semantics
Callison-Burch et al. (2011)	Machine Translation	McCoy et al. (2012)	Cognitive psychology
Callison-Burch et al. (2012)	Machine Translation	Ruiter et al. (2022)	NLG
Callison-Burch et al. (2009)	Machine Translation	Sánchez-Cartagena et al. (2012)	Machine Translation
Callison-Burch et al. (2008)	Machine Translation	Schulz et al. (2019)	Semantics
Cao et al. (2022)	Semantics	Vaassen and Daelemans (2011)	Affective computing
Cercas Curry et al. (2021)	Abuse detection	Vela and van Genabith (2015)	Machine Translation
Cleuren et al. (2008)	Automatic text grading	Walsh et al. (2020)	Syntax
D'Souza et al. (2021)	Sematics	Wang and Sennrich (2020)	Machine Translation
Deshpande et al. (2022)	Semantics	Wang et al. (2021)	Machine Translation
Downey et al. (2011)	Automatic text grading	Wang et al. (2014)	NLG
Friedrich and Palmer (2014)	Semantics	Zeyrek et al. (2018)	Semantics

Table 7: Publications in the ACL Anthology in which intra-annotator agreement is reported.

B Included papers

A list of included publications from the ACL Anthology that report intra-annotator agreement is presented in Table 7.

C Cohen's kappa scores

	Reliability (*Inter-*) κ		**Stability** (*Intra-*) κ	
	μ	σ	μ	σ
Offence	0.05	0.28	0.27	0.28
Sentiment	0.02	0.25	0.17	0.29
Entailment	0.02	0.25	0.28	0.28
Anaphora	0.07	0.31	0.22	0.35
Overall	0.04	0.28	0.23	0.30

Table 8: Pairwise reliability and stability of the collected labels measured with mean (μ) and standard deviations (σ) across items for inter- and intra-annotator agreement scores measured with Cohen's kappa (κ).