Human Label Variation in Discourse and Pragmatics Phenomena
ESSLLI 2026 Workshop

Massimo Poesio and Manfred Stede

Utrecht University and Queen Mary University of London / University of Potsdam

Prague, August 10-15, 2026

Disagreement in Human Judgments

- We have known for at least two decades that the assumption that a single preferred interpretation exists in human interpretation tasks is an idealization at best, both in NLP and in vision (Poesio and Vieira, 1998; Chklovsky and Mihalcea, 2003; Poesio et al, 2001, 2005, 2006; Stede, 2008; Tsoumakas and Katakis, 2007; Sheng et al, 2008; Plank et al, 2014; Aroyo and Welty, 2015).
- Every annotation project since then, has encountered frequent cases on which humans **disagree**, particularly for complex tasks (Passonneau et al, 2006, 2012; Beigman-Klebanov et al, 2008; Poesio et al, 2008, 2019, 2020; Versley, 2008; Recasens et al, 2011; Pradhan et al, 2012; Aroyo et al, 2023;).

Disagreements due to ambiguity

3.1	M:	can we .. kindly hook up
3.2	:	uh
3.3	:	engine E2 to the boxcar at .. Elmira
4.1	S:	ok
5.1	M:	+and+ send **it** to Corning
5.2	:	as soon as possible please
6.1	S:	okay
		[2sec]
7.1	M:	do let me know when it gets there
8.1	S :	okay it'll /
8.2	:	it should get there at 2 AM
9.1	M:	great
9.2	:	uh can you give the
9.3	:	manager at Corning instructions that
9.4	:	as soon as it arrives
9.5	:	it should be filled with oranges
10.1	S :	okay
10.2	:	then we can get that filled

From the TRAINS Corpus

Variation is inescapable in subjective tasks

Variation is inescapable in subjective tasks

- The issue has acquired much greater prominence with the increased focus on SUBJECTIVE tasks such as sentiment analysis (Kenyon-Dean et al, 2018), humour / sarcasm detection (Casola et al, 2024), or unsafe language detection (Akthar et al, 2019; Kennedy et al, 2020; Leonardelli et al, 2021; Almanea and Poesio, 2022; Sap et al, 2022; Aroyo et al, 2023; Frenda et al, 2024; Jiang et al, 2024)
- The field has woken up to the challenge, developing new approaches to data collection, and new methods for training and evaluating AI models without relying on that assumption (Aroyo and Welty, 2015; Pavlick & Kwiatkovski, 2019; Uma et al, 2021; Basile, 2022; Davani et al, 2022; Plank, 2022; Sap et al, 2022; Jiang and de Marneffe, 2023)
- With the rise of large language models, the problem has been recast as pluralistic alignment (Sorensen et al, 2024)

An explosion in activity

- A number of projects and teams working on the topic
- Three Learning With Disagreement shared tasks have now been organized (Uma et al, 2021; Leonardelli et al, 2023; Leonardelli et al, 2025),
- Several series of workshops on the topic (e.g., NLPerspectives, UncertaiNLP, Pluralistic Alignment).

Open Questions

- But there are still plenty of open questions, such as:
  - What types of variation should be preserved?
    - E.g., what about noise?
  - What should models aim for?
    - E.g., model the distribution of judgments (SOFT LABEL) or the views of individuals / groups of individuals (PERSPECTIVISM) ?
  - How should models be evaluated?

The Return of Discourse and Pragmatic Interpretation

- Although much of the recent focus of work on dealing with variation in NLP has been on subjective tasks, discourse and pragmatic interpretation tasks raise many additional questions that are still open, such as
  - How to handle learning and evaluating with disagreement in cases where the 'label' is complex, such as coreference and discourse structure (see Manfred's intro)
  - Which of these disagreements are data-driven (= due to ambiguity) and which to subjective biases?

DeMeVa and Variability

- **Dealing with Meaning Variation in NLP (DeMeVa)** is a five-years project at the University of Utrecht.
  - The project investigates variation arising from a variety of discourse and pragmatics interpretation issues, such as reference to plural, reference in conversations, and discourse structure recognition .

DeMeVa and Variability

- **Disagreements and their limits in discourse structure annotation**
  - M. Stede / T. Scheffler
  - 4-year project in SFB 1287 "Limits of Variability in Language: Cognitive, Computational, and Grammatical Aspects" (U Potsdam)

RST annotation

- Annotators don't agree very much with each other.
  - not a problem
- Annotators don't agree very much with themselves.
  - A problem!
- RST tree is supposed to capture the subjective interpretation by the analyst - but it should do so as transparently and unambiguously as possible.

The HLV-DP Workshop

- A workshop on variation, disagreement, and perspectivist annotation in discourse and pragmatics
- Presenting ongoing research on variation in the interpretation of coreference, reference, discourse structure interpretation, and more