Universiteit Utrecht

From annotation reliability to interpretive diversity:

Rethinking ground truth in discourse research

Merel Scholman @ HLV

What discourse relation holds between these two sentences?

Ryan was decorating the house for a party.
He was making a balloon arch.

1. Manner
2. Purpose
3. Other

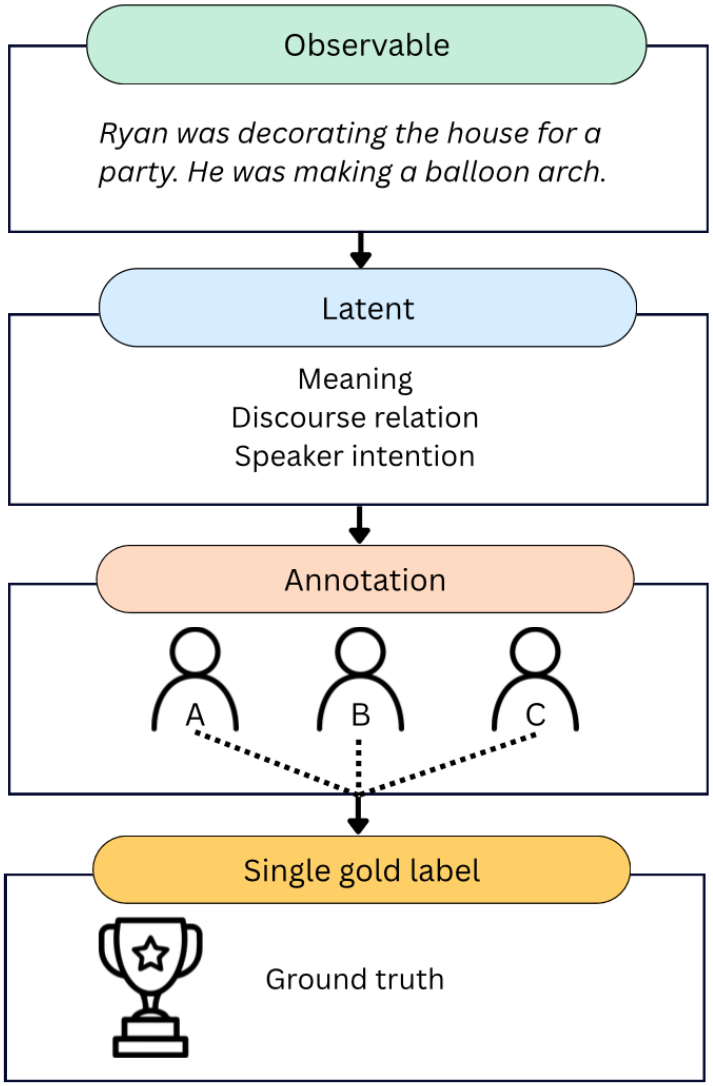
- If two annotators disagree, what should happen?

What is the goal of annotation?

Traditionally:

- recover a single "ground truth"
- minimise disagreement
- adjudicate annotations into one gold label

➢ Disagreement is treated as noise.



What is good enough agreement?

Artstein & Poesio (2008):

5.3 Interpreting the Values

We view the lack of consensus on how to interpret the values of agreement coefficients as a serious problem with current practice in reliability testing, and as one of the main reasons for the reluctance of many in CL to embark on reliability studies. Unlike significance values which report a probability (that an observed effect is due to chance), agreement coefficients report a magnitude, and it is less clear how to interpret such magnitudes. Our own experience is consistent with that of Krippendorff: Both in our earlier work (Poesio and Vieira 1998; Poesio 2004a) and in the more recent efforts (Poesio and Artstein 2005) we found that only values above 0.8 ensured an annotation of reasonable quality (Poesio 2004a). We therefore feel that if a threshold needs to be set, 0.8 is a good value.

***Caveat for discourse coherence**

Spooren & Degand (2010):

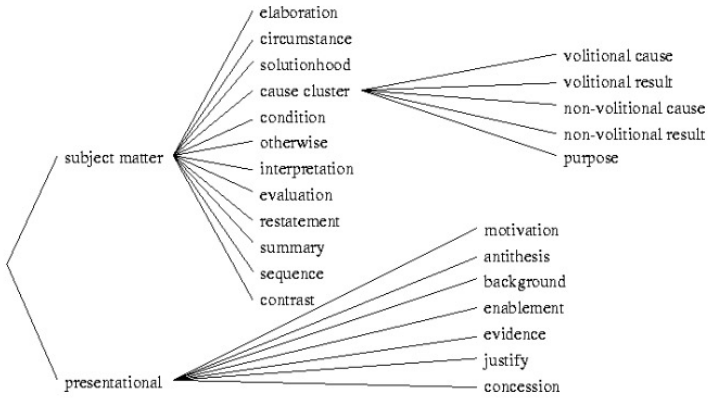
- Our theories contain many categories

Temporal	
Synchronous	–
Asynchronous	Precedence Succession

Contingency	
Cause	Reason Result Negative-result*
Condition	Arg1-as-cond Arg2-as-cond
Negativecondition	Arg1-as-negcond Arg2-as-negcond
Purpose	Arg1-as-goal Arg2-as-goal Arg2-as-negGoal

Comparison	
Contrast	–
Similarity	–
Concession	Arg1-as-denier* Arg2-as-denier

Expansion	
Conjunction	-
Disjunction	–
Equivalence	–
Instantiation	Arg1-as-instance Arg2-as-instance
Level-of-detail	Arg1-as-detail Arg2-as-detail
Substitution	Arg1-as-subst Arg2-as-subst
Exception	Arg1-as-excpt Arg2-as-excpt
Manner	Arg1-as-manner Arg2-as-manner



***Caveat for discourse coherence**

Spooren & Degand (2010):

- Our theories contain many categories
- Discourse is underdetermined and expressed in contextual information

Jet arrived at the party. Merel left.

[Common ground: Jet & Merel are best friends]

***Caveat for discourse coherence**

Spooren & Degand (2010):

- Our theories contain many categories
- Discourse is underdetermined and expressed in contextual information
- Coherence is created in the mind of the comprehender

Hamsters turn into cannibals when they lack food.

***Caveat for discourse coherence**

Spooren & Degand (2010):

- Our theories contain many categories
- Discourse is underdetermined and expressed in contextual information
- Coherence is created in the mind of the comprehender

Suggested standard: kappa > 0.7, or explain why it is lower

Today

From annotation reliability…

- Disagreement = bad!
- How can we minimize disagreement and obtain the ground truth?

…to interpretive diversity

- What can disagreement tell us?
- How can we exploit disagreement to obtain meaningful data from diverse perspectives?

Today

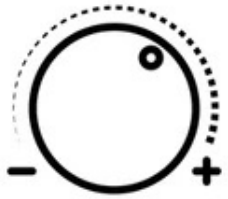
From annotation reliability…

- Disagreement = bad!
- How can we minimize disagreement and obtain the ground truth?

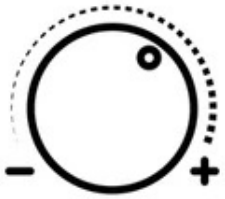
…to interpretive diversity

- What can disagreement tell us?
- How can we exploit disagreement to obtain meaningful data from diverse perspectives?

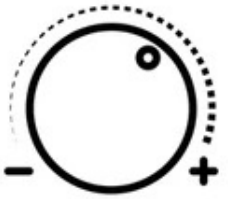
Traditional annotation: how can we improve agreement?



Improve **input**:
more context



Improve **task**:
use linguistic tests



Improve **annotators**:
more training

Increase agreement: provide context

Spooren & Degand (2010):

> The analyst who wants to interpret coherence relations in a text should make use of all of the contextual information available. In the case of writing, the contextual information is usually encoded into language: writers expect their readers to infer their communicative intentions on the basis of the text only. In case of spontaneous speech, this is radically different: much of the communicative content is available in the paralinguistic and extralinguistic context. Ideally then, the analyst should have witnessed the original conversation; in corpora of spoken language the sound recording is nowadays usually provided (sometimes even video recordings) as an approximation. In our case, however,

How much context?

No guidelines on how much context is needed...

- entire text? (e.g., Rehbein et al., 2016; Zufferey et al., 2012)
- certain amount of context preceding and following the relation? (e.g., Hoek & Zufferey, 2015; Scholman et al., 2016)

How much context? -> Scholman & Demberg (2017, LAW)

Items: 234 relations from Penn Discourse Treebank

2 conditions: with and without context

Context:

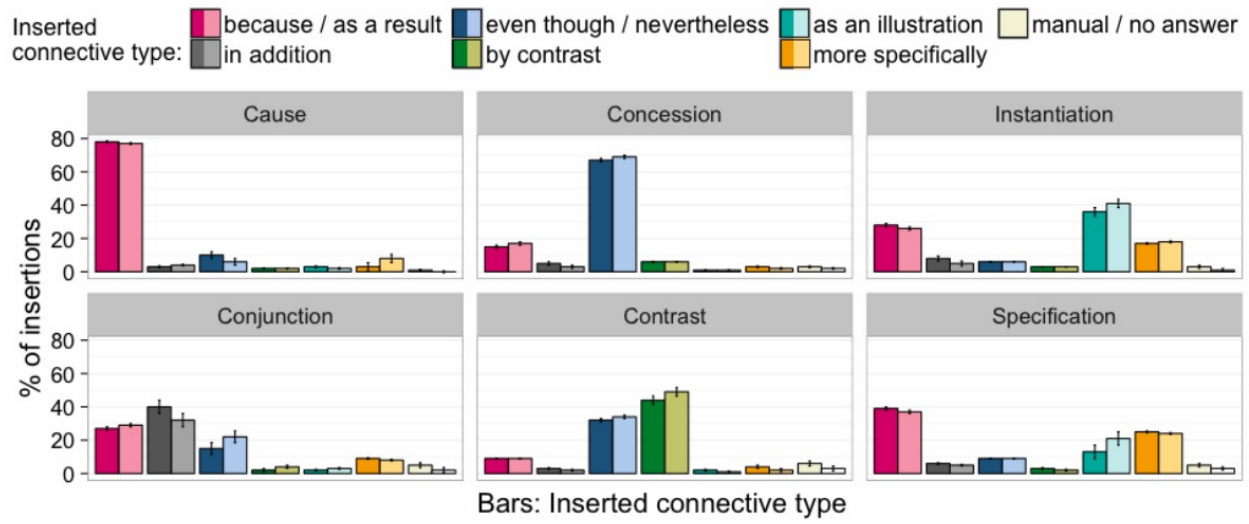
- 2 sentences preceding
- 1 sentence following relation

How much context? -> Scholman & Demberg (2017)

Darker colours = context condition; lighter colours = no-context condition

How much context? -> Scholman & Demberg (2017)

Darker colours = context condition; lighter colours = no-context condition



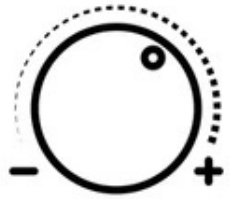
So really no context effect? -> Atwell et al (2022)

Annotator accuracy increased from 0.350 to 0.414 (+6.4%)

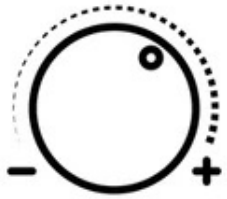
Discourse sense	Ground Truth	Without context	With context	Context effect
Temp.Asynchronous	6	4	5	**better**
Temp.Synchronous	0	1	1	neutral
Cont.Cause	14	11	9	*worse*
Cont.Cause+Belief	2	3	3	neutral
Cont.Cause+SpeechAct	2	1	1	neutral
Cont.Purpose	4	0	0	neutral
Comp.Contrast	3	17	13	**better**
Comp.Concession	8	2	3	**better**
Comp.Concession+SpeechAct	1	0	0	neutral
Comp.Similarity	2	0	1	**better**
Exp.Conjunction	20	31	25	**better**
Exp.Instantiation	5	14	15	*worse*
Exp.Equivalence	4	7	8	*worse*
Exp.Exception	1	0	0	neutral
Exp.Level-of-detail	30	14	21	**better**
Exp.Manner	0	1	1	neutral
Exp.Substitution	4	0	0	neutral

Table 11: Frequency of discourse senses in our annotated set with respect to ground truth, annotations without context, and annotations with context. We can see that at a label distribution level, a context window usually adds a better or neutral effect.

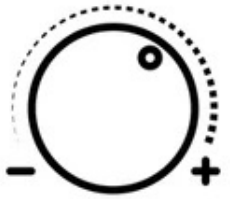
Traditional annotation: how can we improve agreement?



Improve input:
more context



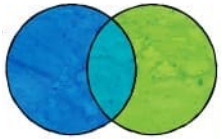
Improve task:
use linguistic tests



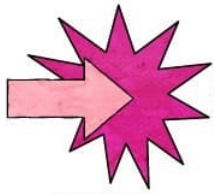
Improve annotators:
more training

Increase agreement: focus on connectives and cue phrases

And then → temporal



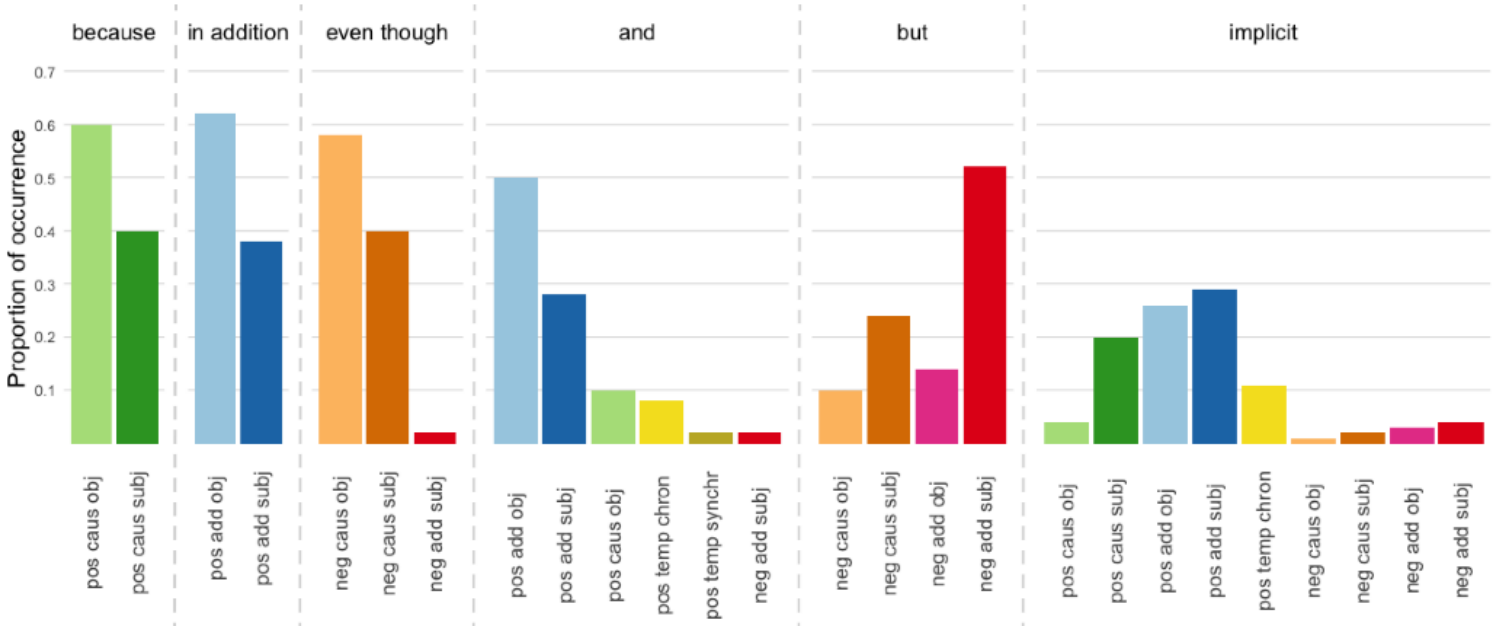
Whereas → contrastive



Because → causal

How do connectives impact agreement?

Hoek, Scholman & Sanders (2021):



How do connectives impact agreement?

Hoek, Scholman & Sanders (2021):

Connective		%	κ	AC_1
explicit	*because*	84	.68	.68
	in addition	82	.57	.69
	even though	78	.58	.74
underspecified	*and*	74	.58	.71
	but	58	.39	.46
implicit	Ø	66	.58	.64

Table 1: IAA per connective type and connective

How do connectives impact agreement?

Hoek, Scholman & Sanders (2021):

Connective		%	κ	AC_1
explicit	*because*	84	.68	.68
	in addition	82	.57	.69
	even though	78	.58	.74
underspecified	*and*	74	.58	.71
	but	58	.39	.46
implicit	Ø	66	.58	.64

Table 1: IAA per connective type and connective

Zikánová et al (2019):

	Discourse implicit	Discourse explicit
F1 – presence of a relation	0.54	0.84
Agreement on types	0.58	0.77
Cohen's κ on types	0.47	0.71

Solution: connective insertion or paraphrase tests

Knott & Dale (1996)

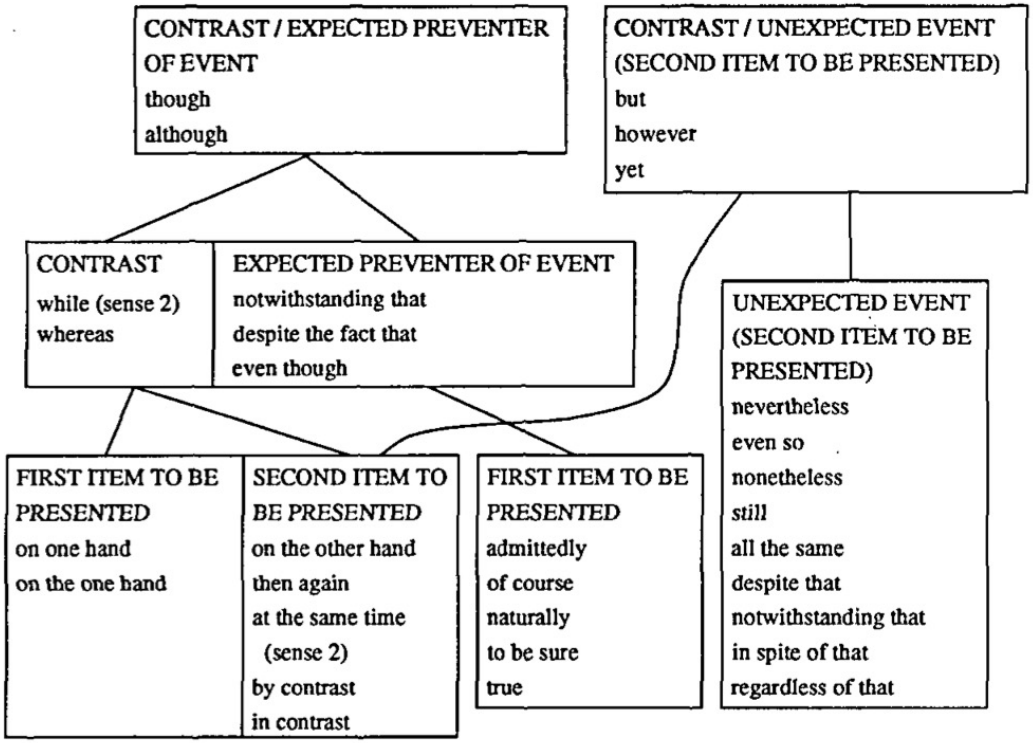


Figure B4. Contrast/violated expectation/choice.

Solution: connective insertion or paraphrase tests

Degand (1998); Sanders (1997)

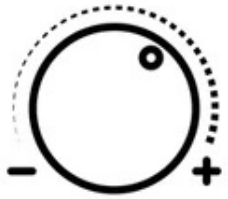
Ideational, volitional: this action is the consequence of the following fact :
Ideational, non-volitional: this situation is the consequence of the following fact :
Interpersonal, epistemic: one may conclude this on the basis of the following situation :
Interpersonal, speech-act: this utterance is motivated by the following situation :

Reconstruct the causal basic operation between the propositions *P* and *Q*, which correspond roughly to the propositions underlying S1 en S2 (they are the propositions in the basic operation, cf. Sanders et al., 1992, section 2.1.). Paraphrase it by making use of the formulations below and consider which formulation is the best expression of the meaning of the CR in this context.

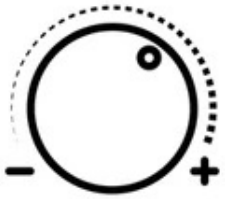
(i) a. the fact that P *causes S.'s claim / advice / conclusion* that *Q*
(i) b. the fact that *Q causes S.'s claim / advice / conclusion* that *P*
(ii) a. the fact that *P causes the fact* that *Q*
(ii) b. the fact that *Q causes the fact* that *P*

Quite meta-linguistic… What type of annotators do we need?

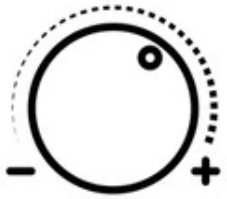
Traditional annotation: how can we improve agreement?



Improve input:
more context



Improve task:
use linguistic tests



Improve annotators:
more training

Increase agreement: improve the annotator

Who annotates?

- Trained linguistic students / researchers
- Read annotation guidelines, practice on examples
- Receive feedback and resolve ambiguities

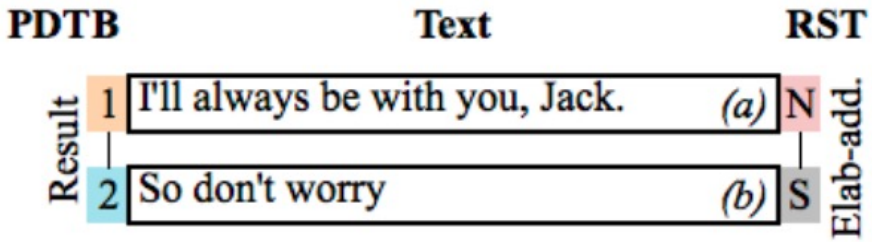
→ Better-trained annotators should converge on the intended interpretation.

→ But what if disagreement remains after training?

When experts disagree…

- Is one expert wrong?

→ Demberg, Scholman & Asr (2019, D&D):
PDTB and RST-DT annotations partly on the same text.
Do these experts agree?



When experts disagree: explicit relations

RST-DT label \ PDTB label	Synch.	Asynch.	Cause	Condition	Contrast	Concession	Conjunction	Instantiation	List	Total
Temp.-same-time	**63**	1	1	1	4		5			75
Temporal-after		**48**			1		2			54
Sequence	2	**29**	1		1		19			52
Circumstance	**89**	79	29	21	7	4	**10**			259
Result	1	3	**30**		2		5			41
Consequence	5	6	**45**		4	1	16		1	95
Explanation-arg.	6		**39**		7	2	1			57
Reason		2	**67**							72
Condition	5	13	1	**104**	2	1	1			182
Contrast	1	1			**160**	23	17			208
Concession	5	4		3	101	**53**	4			182
Antithesis	1	3			**170**	**37**	10			243
Comparison	2	1			26	2	**9**			41
Elaboration-add.	4	3	1		30	8	**122**	3	3	192
Example	1				1		3	**29**		35
List	2	4	1		17	1	303		**47**	377
Total	187	197	215	129	533	132	527	32	51	

When experts disagree: implicit relations

RST-DT label \ PDTB label	Asynch.	Cause	Contrast	Conjunction	Instantiation	Specification	List	EntRel	Total
Background	**7**	12	10	16	2	4		21	72
Circumstance	**1**	10	8	**11**		6		15	51
Consequence	4	**16**	7	11	2	2	1	4	49
Evidence		**12**	2	12	29	28		6	92
Explanation-arg.	2	**114**	16	16	35	51	1	22	267
Reason	1	**20**	1	3	1	2		3	31
Evaluation		**7**	9	11	2	**1**		12	46
Interpretation		**16**	3	9	2	8		8	49
Contrast	1	2	**35**	2			2	3	56
Comparison		1	24	14			2	2	44
Elaboration-add.	29	168	73	**221**	36	151	7	266	991
Example		9	1	6	**64**	21		2	106
Elab.-gen.-spec.		8		11	17	**44**		18	99
List	6	24	29	120	6	13	**74**	30	311
Comment		16	9	**8**		**4**		8	49
Total	75	499	277	511	202	406	88	463	

Is PDTB or RST-DT right?

Goosebumps were once very useful to our ancestors. The bumps used to help their hair stand on end to make them appear larger in a threatening situation.

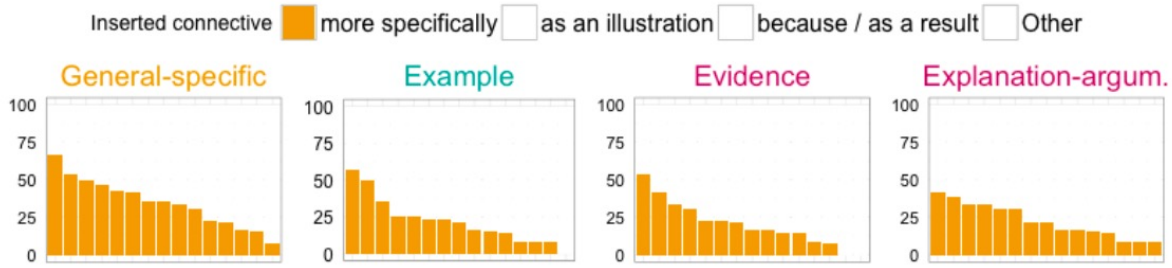
- Specification -> *more specifically*
- Example -> *for example*
- Reason / evidence -> *because*

How do non-experts interpret these relations?

Scholman & Demberg (2017, D&D): crowdsource interpretations

SPECIFICATIONS:

Distribution (%) of inserted connectives per RST-DT class and item:

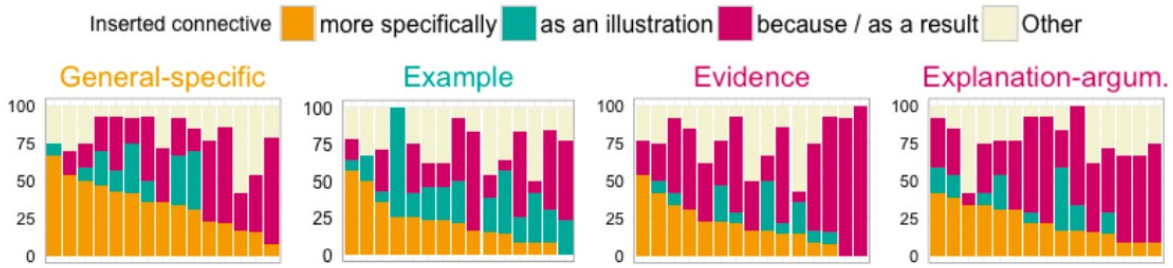


How do non-experts interpret these relations?

Scholman & Demberg (2017, D&D): crowdsource interpretations

SPECIFICATIONS:

Distribution (%) of inserted connectives per RST-DT class and item:



Variation is not merely a defect of annotation;
it reflects something about discourse interpretation itself.

Today

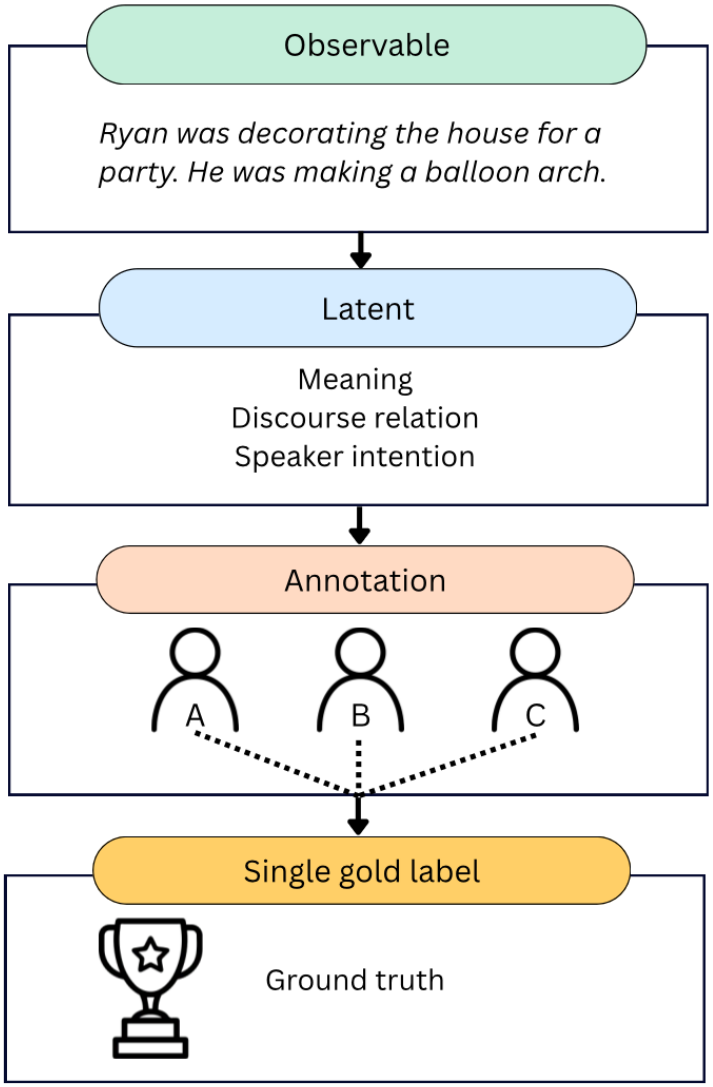
From annotation reliability...

- Disagreement = bad!
- How can we minimize disagreement and obtain the ground truth?

...to interpretive diversity

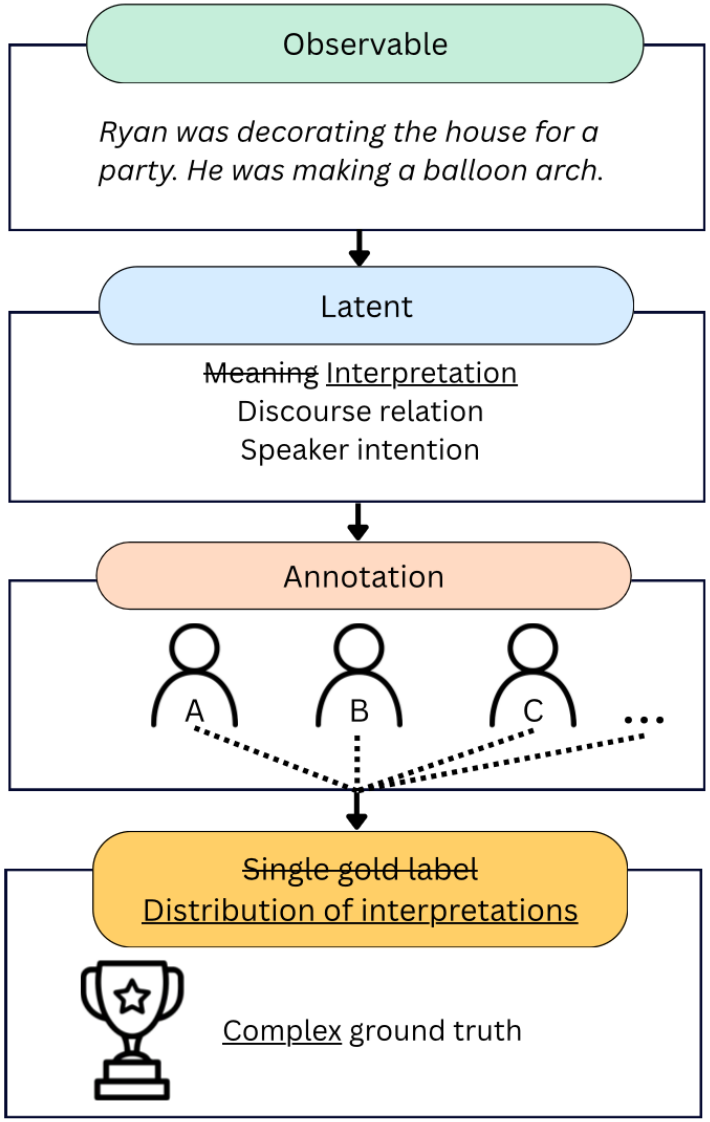
- What can disagreement tell us?
- How can we exploit disagreement to obtain meaningful data from diverse perspectives?

What is the goal of annotation?



What is the goal of annotation?

- One gold label -> multiple plausible interpretations
- Annotation as measurement -> as sampling human interpretations
- Disagreement = error -> information
- Maximize agreement -> Characterize variation



Is disagreement bad?

We should distinguish disagreement we want to eliminate from disagreement we want to preserve.

Eliminate:	Preserve:
Error	Ambiguity
Uncertainty	Interpretive diversity

Challenge: reliably obtain multiple perspectives and distinguish undesirable from informative disagreement.

Capturing interpretive diversity:
Who should annotate?

Traditionally:

- best annotators = expertise + agreement

HLV perspective:

- right annotators = representative + diverse

→ Expertise is not always an advantage when the goal is to capture human interpretation.

Capturing interpretative diversity:
Why crowdsourcing?

Scientific

✓ Capture population-level variation
✓ Avoid convergence due to shared theoretical training
✓ Obtain distributions rather than consensus

Practical

✓ Parallel annotation
✓ Large samples
✓ Easy recruitment

Capturing interpretative diversity:
Different task design

1. Freely insert connective to express relation

> I merely repeat, remember always your duty of enmity towards Man and all his ways. [type here]
> Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

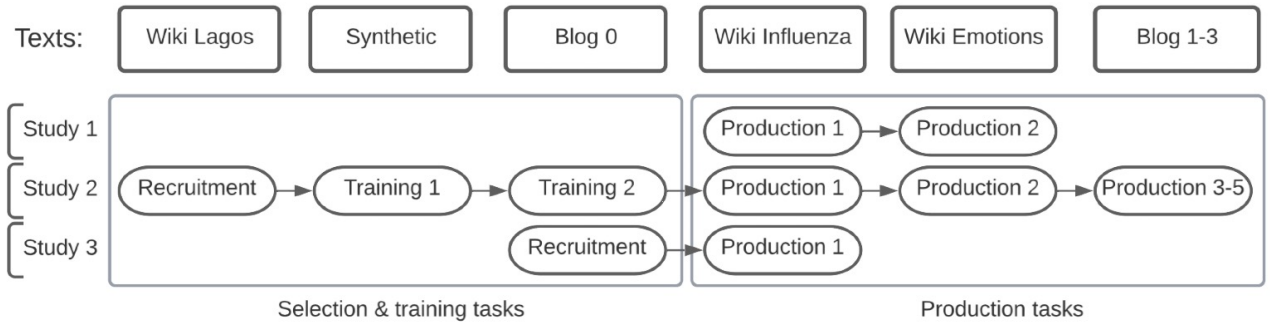
2. Choose from automatically provided list to disambiguate

> [the reason(s) is/are that] [in more detail,] [considering the fact that] [by means of]
>
> I merely repeat, remember always your duty of enmity towards Man and all his ways. [ ]
> Whatever goes upon two legs is an enemy. Whatever goes upon four legs, or has wings, is a friend.

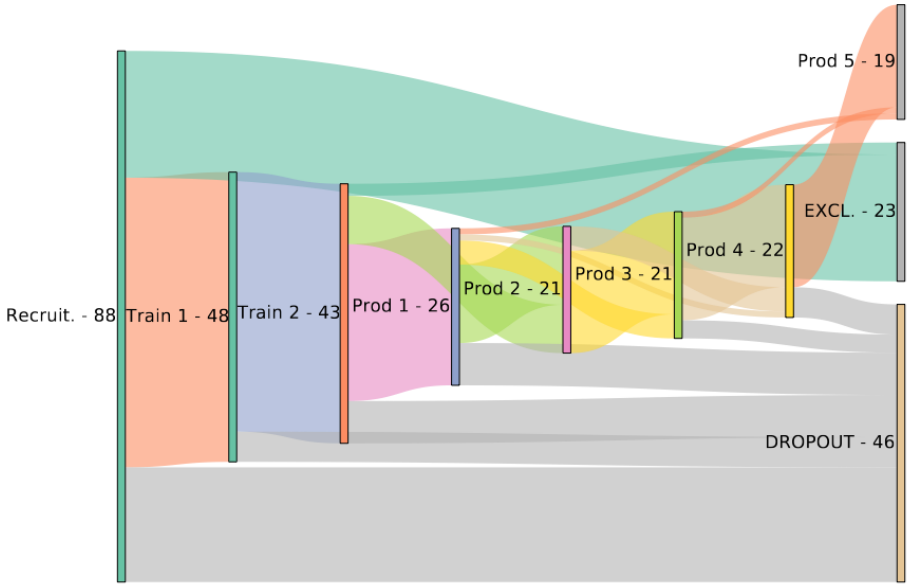
Yung, Scholman & Demberg (2019), *LAW*.

Capturing interpretative diversity:
Not every person fits the crowd…

Scholman, Pyatkin et al. (2022, LREC):
Studied how to ensure the right annotators for your task



Capturing interpretive diversity:
Study 2: Losing my crowd :(



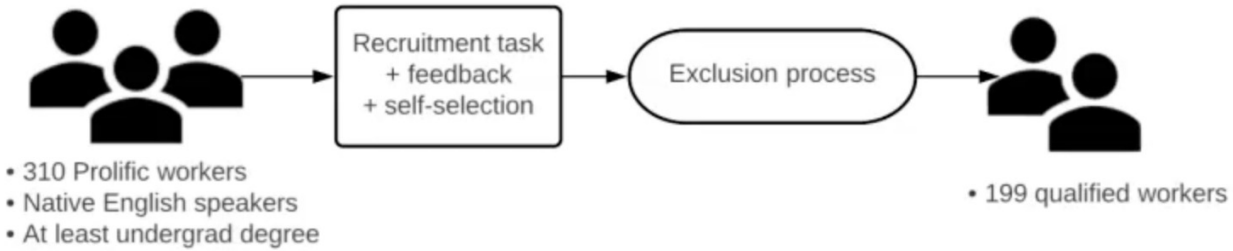
Capturing interpretive diversity:
Selecting your crowd

Study 3: Recruitment task to include only best performers

- Pre-screening: university education
- Feedback during selection task
- Post-screening: self-selection
- Decent: >50% agreement on gold items

Study	Participant type	Invested cost GBP	κ	% Agree gold-maj
1	Untrained	0	.27	45
2	Trained	11.98	.61	73
3	All selected	1.88	.41	68
3	Decent selected	1.88	.58	77

Capturing interpretive diversity:
Selection applied to create corpora, e.g. DiscoGeM 1.0



	EP	Lit	Wiki	Total
No. DRs	2,800	3,060	645	6,505

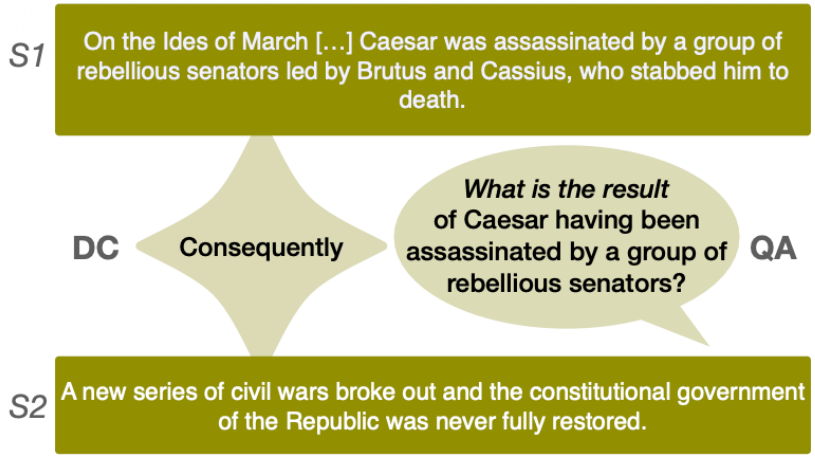
Capturing interpretive diversity:
DiscoGeM 1.0 (Scholman et al., 2022, LREC)

- **Majority-single**: sense with majority agreement
- **IRT-single**: highest probability sense, based on Dawid-Skene model (Passonneau & Carpenter, 2014)
  - Uses unsupervised learning to estimate the most likely sense for every item
  - Assumes single ground truth
- **CrowdTruth-distribution**: all senses that reached threshold of 20% probability based on CrowdTruth 2.0 (Dumitrache et al., 2018)
  - Doesn't enforce agreement between annotators
  - Assumes there is no single ground truth

Aggregated measure	κ	%	P	R	F1
Majority - single	.55	67	.67	.49	.55
IRT - single	.53	64	.64	.46	.52
CrowdTruth - distribution	**.75**	82	.69	.66	**.59**

Capturing interpretive diversity:
The influence of task design

Pyatkin, Scholman, et al (2023, TACL):
What biases do task design introduce?



Capturing interpretive diversity:
The influence of task design

- Similar sets of labels are produced; each method has unique biases
- Merging data distributions coming from different task designs can help boost performance on data coming from a third source

(3) *He had made an arrangement with one of the cockerels to call him in the mornings half an hour earlier than anyone else, and would put in some volunteer labour at whatever seemed to be most needed, before the regular day's work began.* **His answer to every problem, every setback, was "I will work harder!" - which he had adopted as his personal motto.** [QA:ARG1-AS-INSTANCE; DC:RESULT]

label	FN_{QA}	FN_{DC}	FP_{QA}	FP_{DC}
conjunction	43	46	203	167
arg2-as-detail	42	62	167	152
precedence	19	18	18	37
arg2-as-denier	38	20	15	47
result	10	5	110	187
contrast	8	17	84	39
arg2-as-instance	10	7	44	57
reason	12	17	54	37
synchronous	20	27	11	5
arg2-as-subst	21	13	1	0
equivalence	22	22	2	1

Table 4: FN and FP counts of each method grouped by the reference sub-labels.

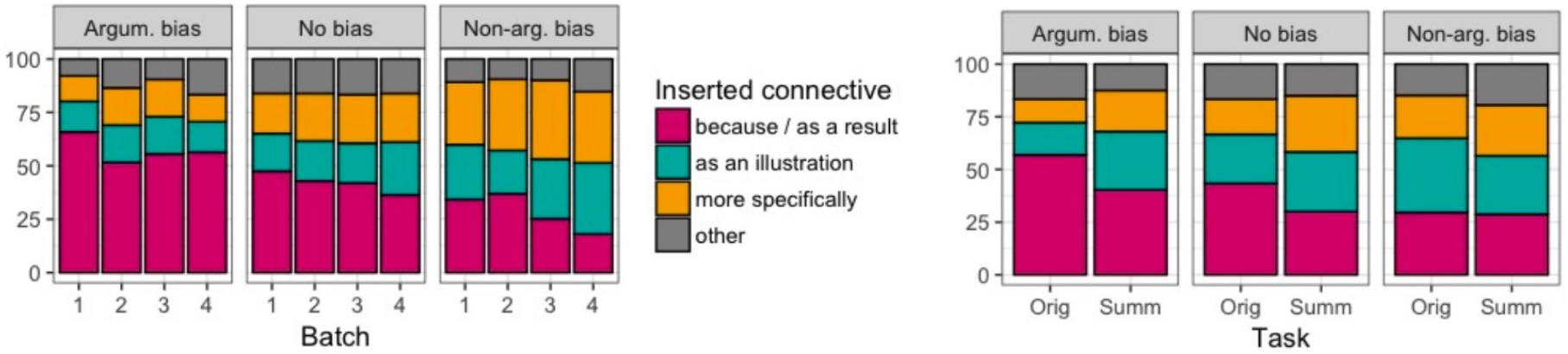
Capturing interpretive diversity:
Within or between comprehenders?

Goosebumps were once very useful to our ancestors. The bumps used to help their hair stand on end to make them appear larger in a threatening situation.

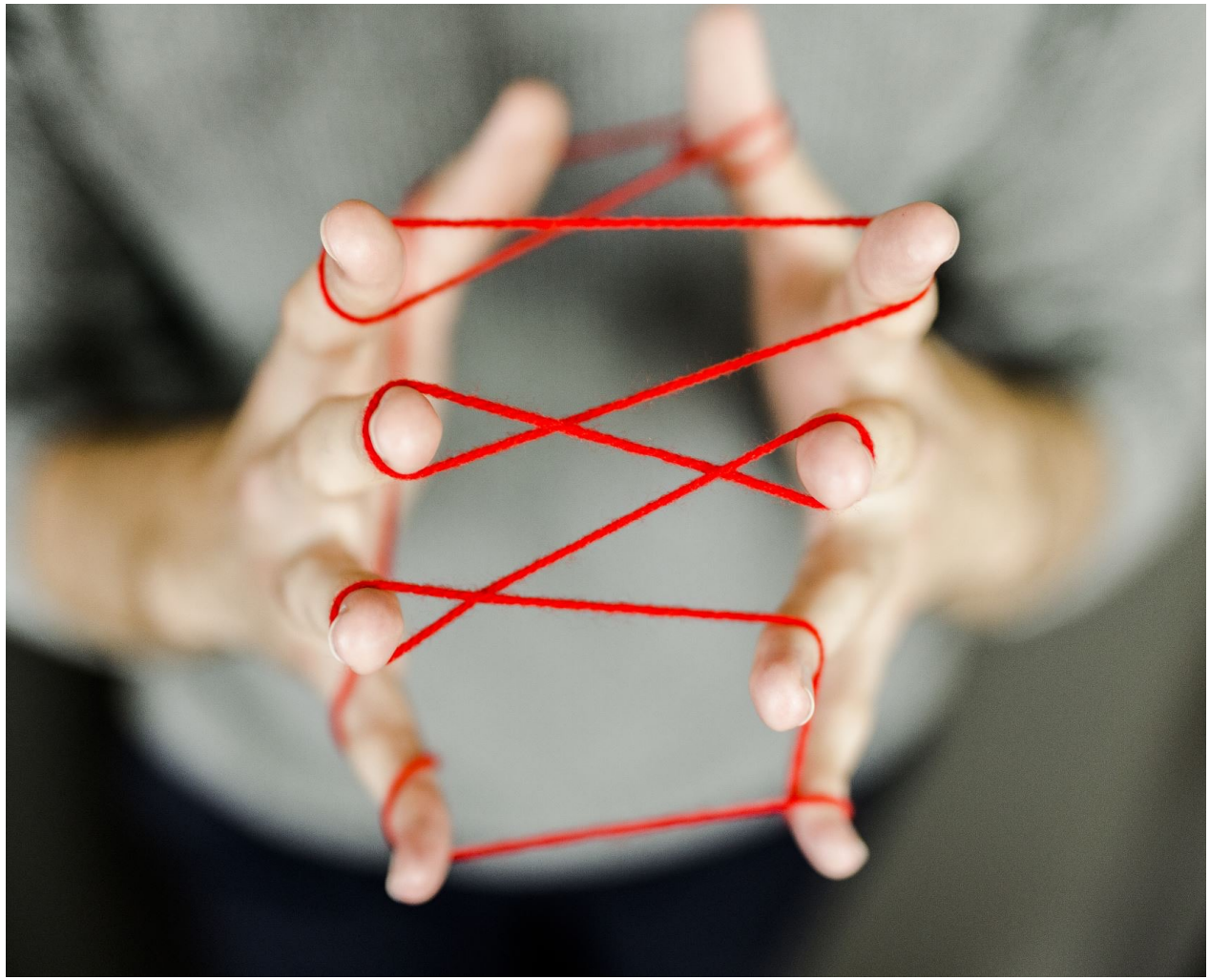
Capturing interpretive diversity:
Within or between comprehenders?

- 4% of our crowdsourced annotations consisted of double labels from single annotators (method effect?)

Capturing interpretive diversity:
Interpretation biases? (Scholman, 2019)



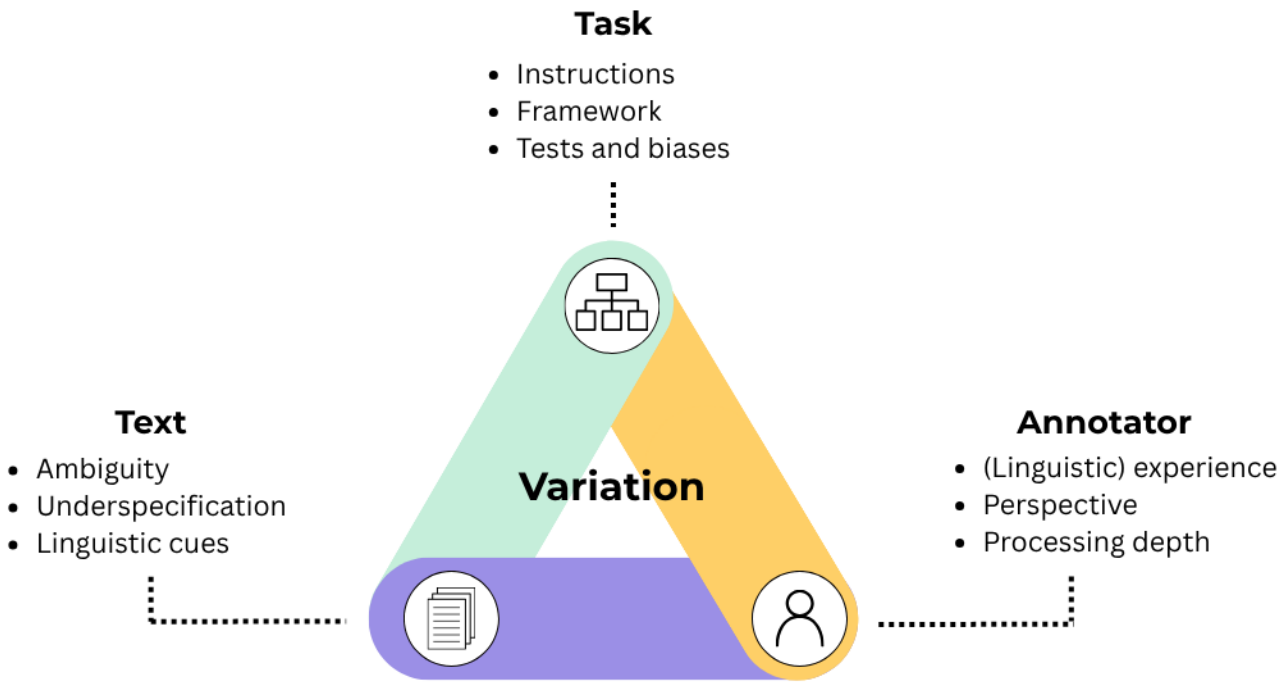
Readers have preferences for interpretating relations argumentatively or not.
These preferences are correlated with depth of processing.



Conclusion

- Discourse coherence is created in the mind of the comprehender, and so there is no single ground truth.
- Disagreement is inevitable and, in fact, informative.
- Challenge: how to capture meaningful interpretations versus noise in modeling approaches?

Human label variation



Variation can come from different sources:
(1) Ambiguity in the input

Texts do not determine their own interpretation.

Examples:
- implicit discourse relations
- underspecification
- context dependence

→ Some disagreement reflects the richness of discourse meaning.

Variation can come from different sources: (2) Annotator factors

Annotators are part of the measurement.

- Discourse expectations
- World knowledge and beliefs
- Individual differences (depth of processing, linguistic experience)

→ Different readers reveal different aspects of meaning.

Variation can come from different sources: (3) Annotation artifacts

Annotation is an experiment.

Annotation outcomes depend on:

- Task design and framework
- Task instructions
- Worker recruitment

→ Small design choices manipulate interpretation.

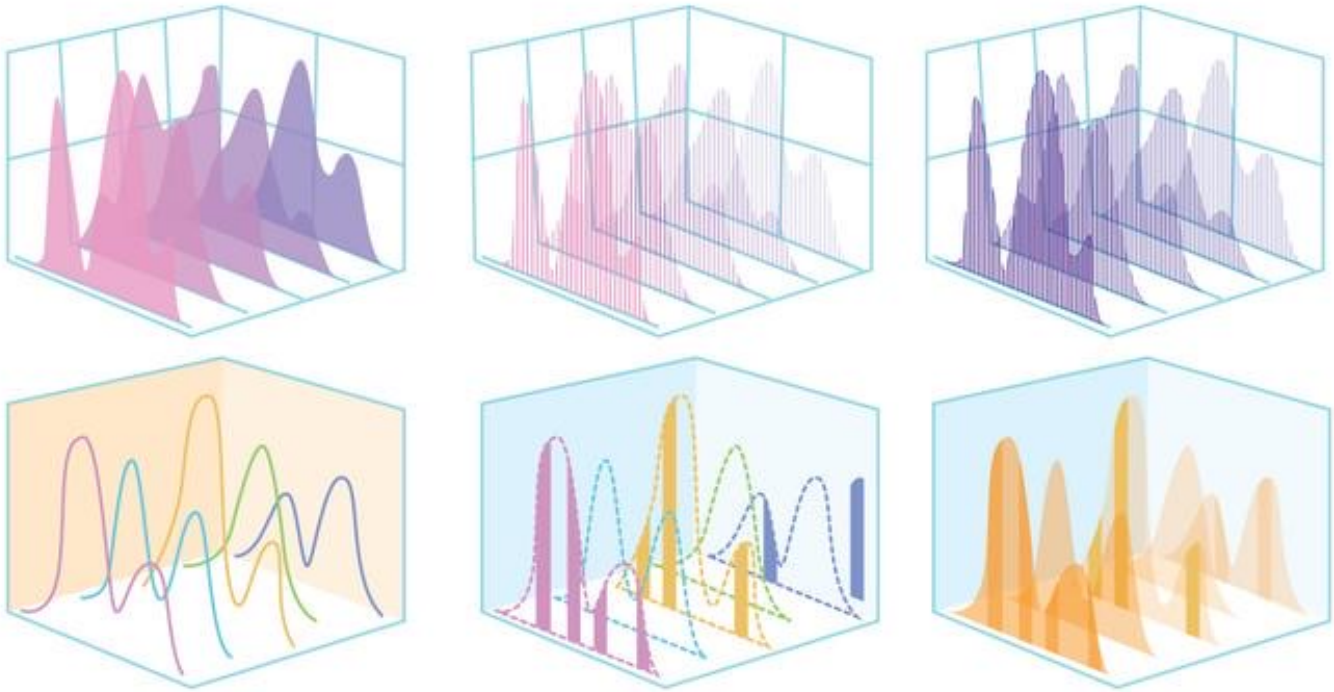
What is ground truth?

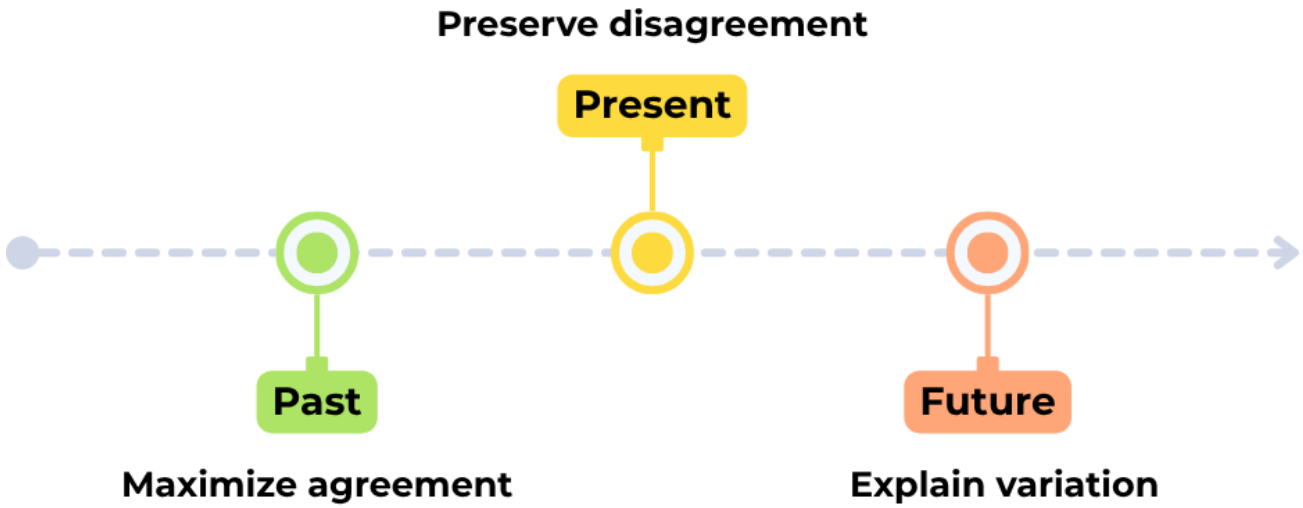
It’s complex!

Represented as a space of possible interpretations rather than a single label.

Possible representations

- Observed label distribution / soft labels
- Perspectivist annotation
- Underspecification





The label distribution is an observation; what is the mechanism?

Text:
Which linguistic properties systematically give rise to multiple interpretations?

Annotator:
Which reader characteristics influence interpretation?

Task:
How can annotation tasks distinguish methodological artifacts from genuine interpretive diversity?

Universiteit
Utrecht

Sharing science,
shaping tomorrow