

# **Separating signal from noise: Annotation errors and alternative interpretations**

**Marie-Catherine de Marneffe**

FNRS – UCLouvain – CENTAL

LAW XX, July 3 2026



LAW X 2016

The 10th Linguistic Annotation Workshop @ ACL 2016

Berlin, August 2016

Special theme: Evaluation of annotation quality

Invited talk. Marie-Catherine de Marneffe:

**Assessing the consistency of crowdsourcing  
and dependency annotations**



LAW X 2016

The 10th Linguistic Annotation Workshop @ ACL 2016

Berlin, August 2016

Special theme: Evaluation of annotation quality

Invited talk. Marie-Catherine de Marneffe:

**Assessing the consistency of crowdsourcing  
and dependency annotations**

**Variation in annotation is signal, not noise**

# Inferring veridicality of events

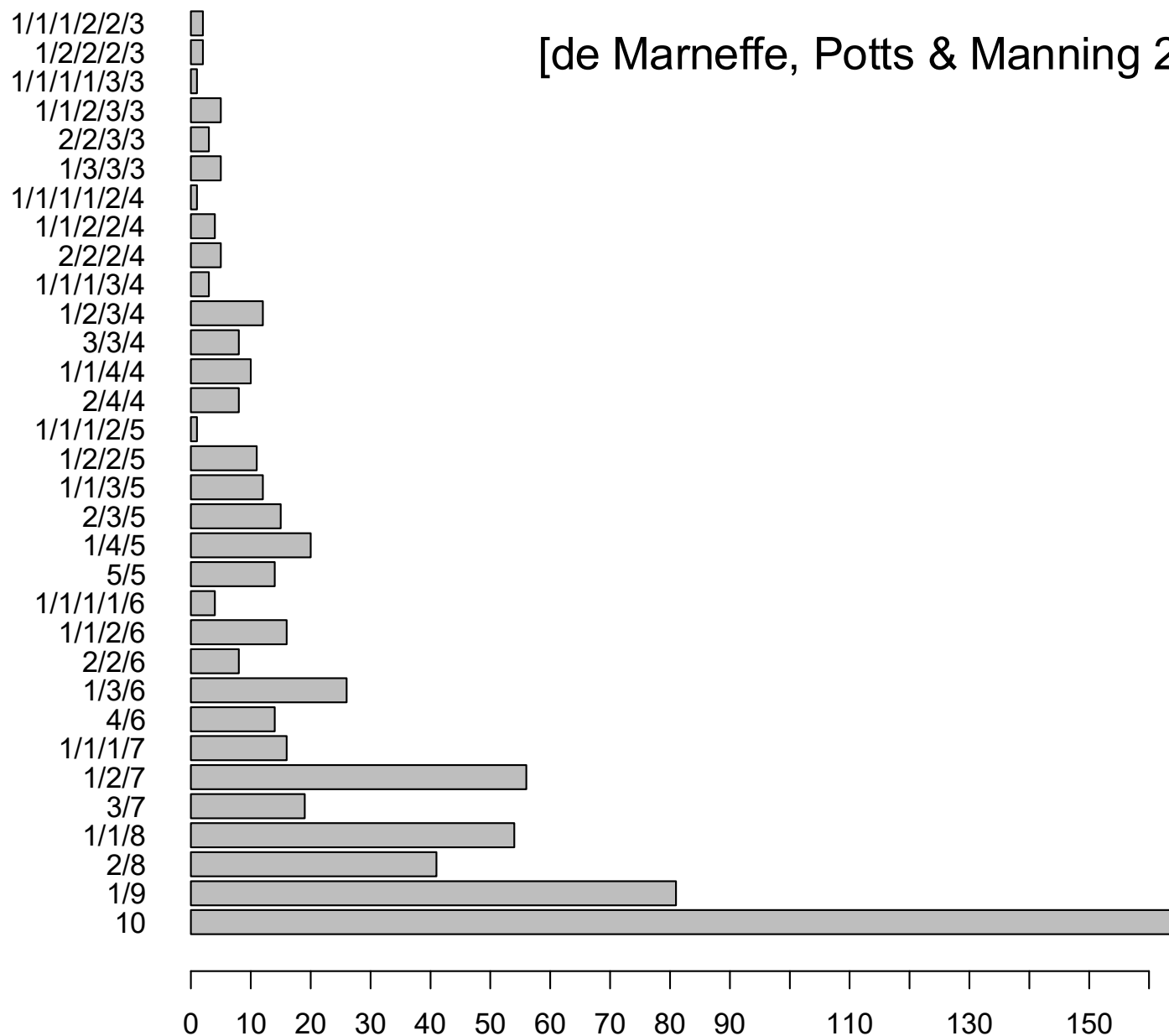
[de Marneffe, Potts & Manning 2012]

FBI agents alleged in court documents today that Zazi had admitted receiving weapons and explosives training from al Qaeda operatives in Pakistan last year.

Did Zazi receive weapons and explosive training?

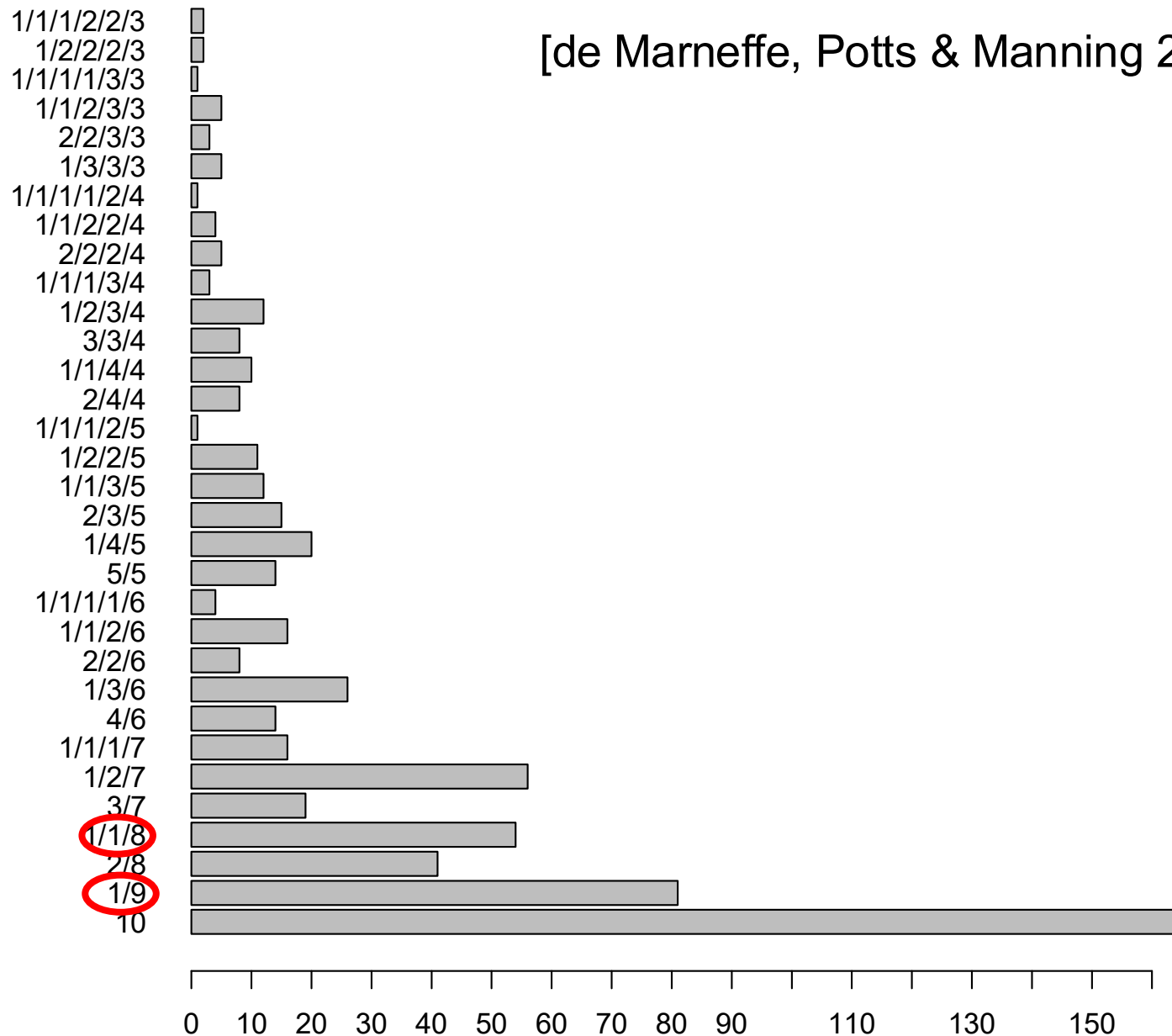
# Inference is graded and variable

[de Marneffe, Potts & Manning 2012]



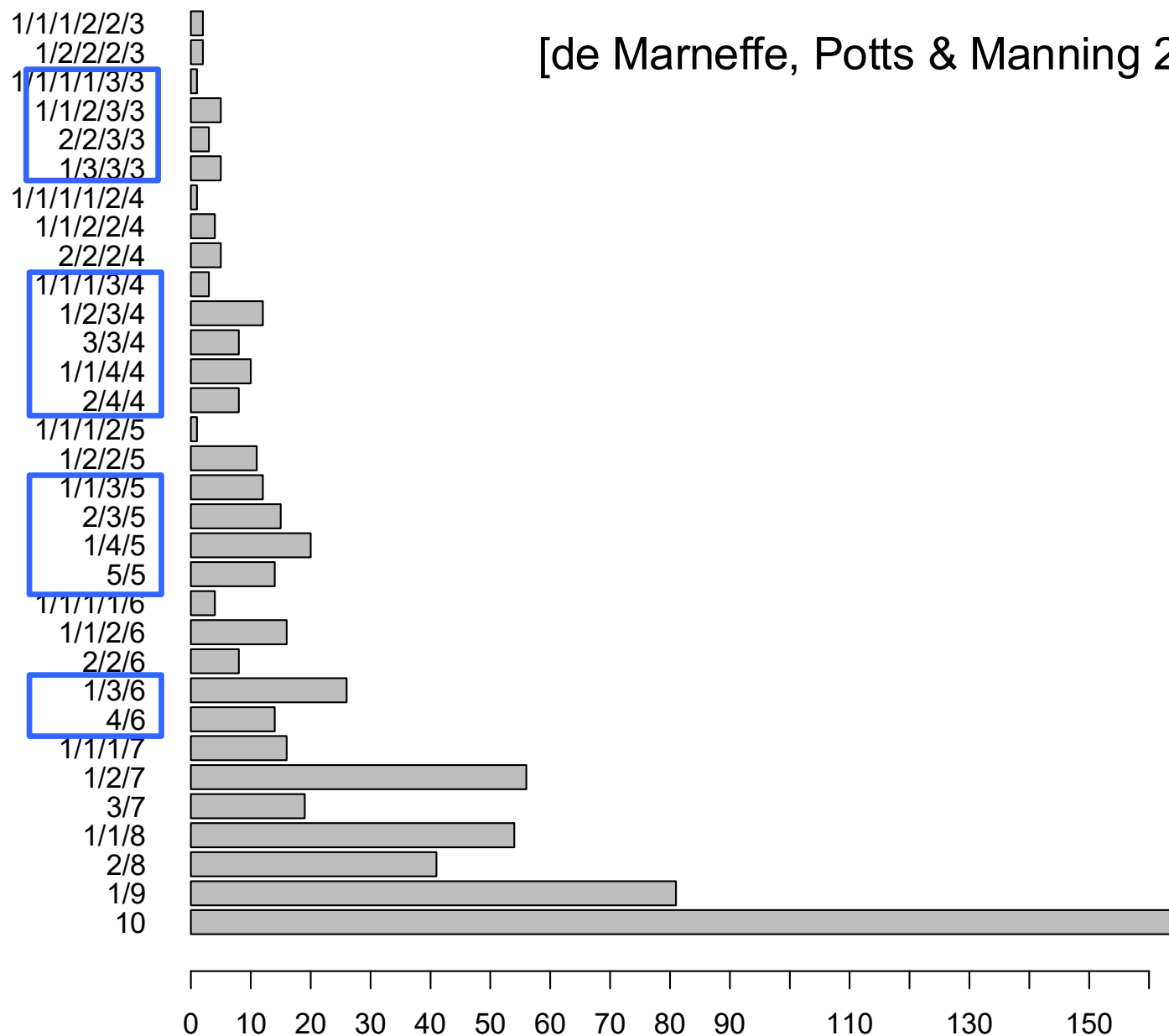
# Inference is graded and variable

[de Marneffe, Potts & Manning 2012]



# Inference is graded and variable

[de Marneffe, Potts & Manning 2012]



# Old problem, but put aside

## **Penn Treebank POS tagging** [Santorini 1990]

- a. Sampling data can be time-consuming.
- b. Sampling data can be full of errors.
- c. Sampling data can be fun.



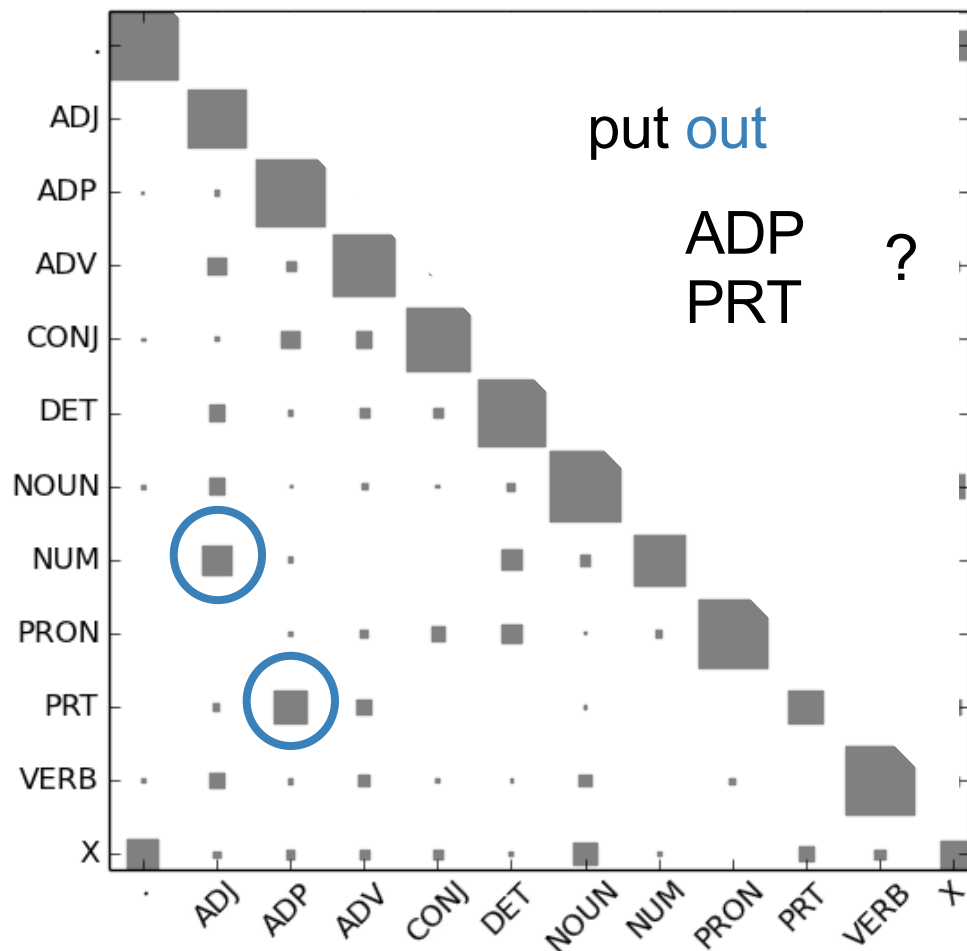
# Old problem, but put aside

## **Penn Treebank POS tagging** [Santorini 1990]

- a. Sampling/*VBG* data can be time-consuming.
- b. Sampling/*NN* data can be full of errors.
- c. Sampling/*NN|VBG* data can be fun.

“Nevertheless, uncertainties can arise. Rather than attempting to forcibly resolve such uncertainties, with the attendant risk of inconsistency, you should simply record them by separating the relevant tags by a vertical slash” (p. 31-32)

# Signal: Annotation variation helps POS tagging



[Plank et al. 2014]

# “Truth is a lie: Crowd truth and the 7 myths of human annotation”

[Aroyo & Welty AI Magazine 2015]

1. One truth
2. Disagreement is bad
3. Detailed guidelines help
4. One is enough
5. Experts are better
6. All examples are created equal
7. Once done, forever valid

# Human label variation

## **The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation**

**Barbara Plank**

EMNLP 2022



# The 20th Linguistic Annotation Workshop

Co-located with ACL 2026 in San Diego

Special theme: Errors in annotations



# The 20th Linguistic Annotation Workshop

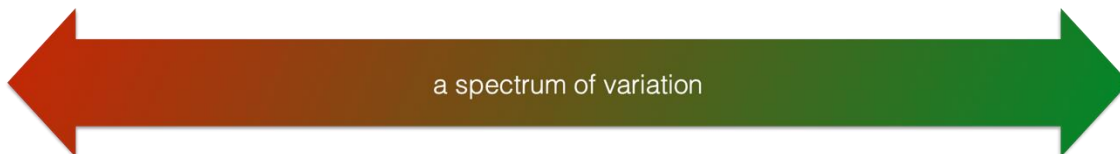
Co-located with ACL 2026 in San Diego

Special theme: Errors in annotations

## Separating signal from noise



**Error** vs.



plausible **Human Label Variation**



# The 20th Linguistic Annotation Workshop

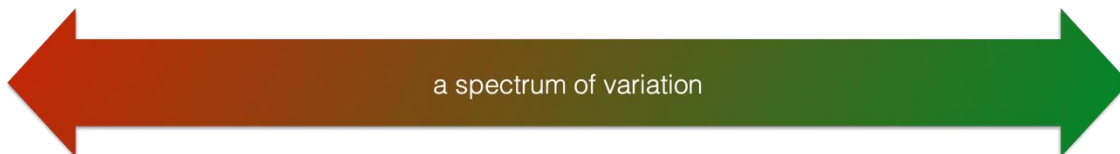
Co-located with ACL 2026 in San Diego

Special theme: Errors in annotations

## Separating signal from noise through explanations of labels



**Error** vs.



plausible **Human Label Variation**

# How to distinguish real errors?

VariErr NLI: Separating annotation error  
from human label variation

Leon Weber-Genzel, Siyao Peng, MC de Marneffe  
& Barbara Plank. ACL 2024



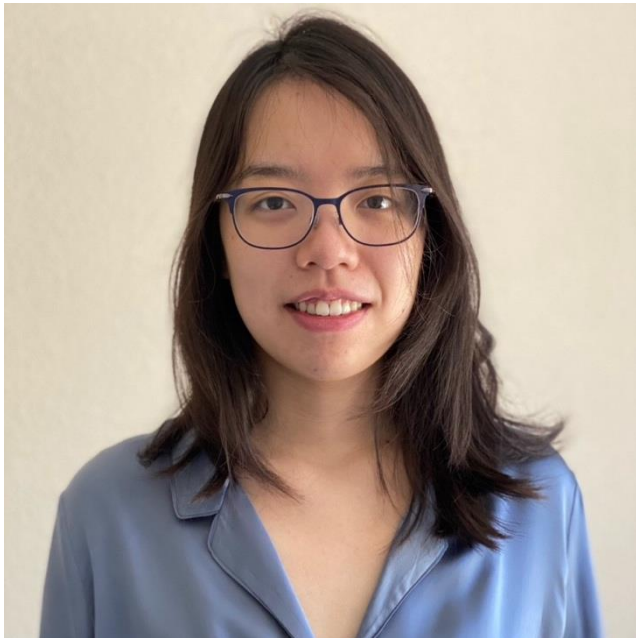


# Inspiration for VariErr: LiveNLI dataset

Ecologically valid explanations for label variation in NLI

Nan-Jiang Jiang, Chenhao Tan & MC de Marneffe

Findings of EMNLP 2023



# Inspiration for VariErr: LiveNLI dataset

P: for a change i i got i get sick of winter just looking everything so dead i hate that

H: I'm so sick of summer. [0, 0.35, 0.65]

**C** – The premise is stating how one is sick of winter, not summer, as the hypothesis describes.

**C** – The speaker hates winter because the foliage is dead, therefore he likely loves summer when everything is alive.

**N** – The context mentions being sick of winter while the statement mentions being sick of summer. These could both be true because the same person may still complain of summer's heat.

# VariErr: Annotations in two rounds

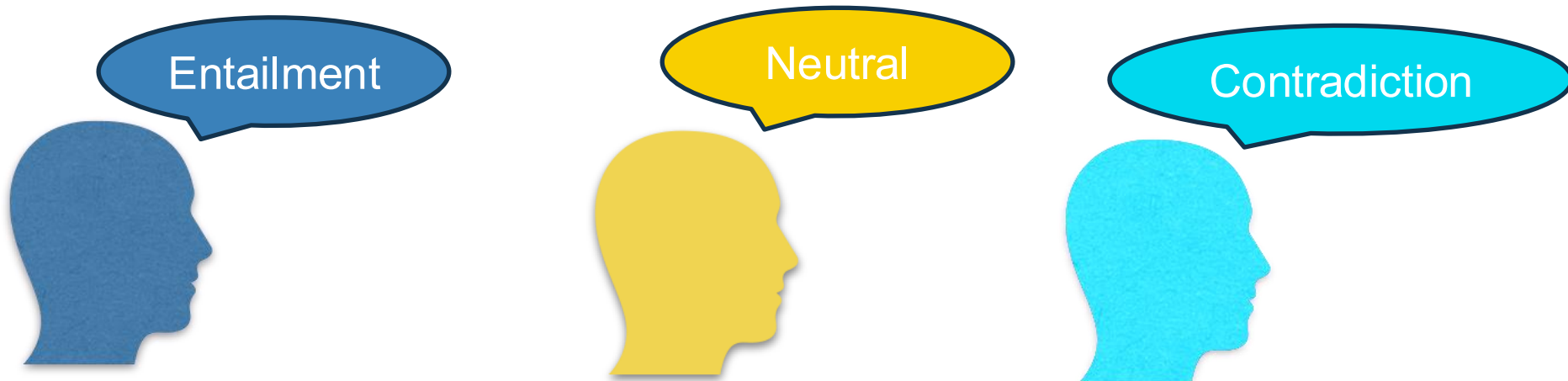
P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as \$5.

# VariErr: Annotations in two rounds

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as \$5.



# 1<sup>st</sup> round: Label and explanation

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as \$5.

## *Round 1: NLI Label & Explanation*

| L | Explanation                                                                                                                                         |
|---|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| E | 5 dollars for a piece of newspaper.                                                                                                                 |
| N | The context only mentions how low the price may be, not how high it may be.<br>The maximum cost of newspaper preprints is not given in the context. |
| C | The context says 5 or 6 cents, not \$5.                                                                                                             |

500 NLI items (from ChaosNLI): 1,933 label-explanation pairs

## 2<sup>nd</sup> round: Validation

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as \$5.

| Round 1: NLI Label & Explanation |                                                                             | Round 2: Validity |   |   |   |
|----------------------------------|-----------------------------------------------------------------------------|-------------------|---|---|---|
| L                                | Explanation                                                                 | 1                 | 2 | 3 | 4 |
| E                                | 5 dollars for a piece of newspaper.                                         | ×                 | × | × | × |
| N                                | The context only mentions how low the price may be, not how high it may be. | ✓                 | ✓ | ✓ | ✓ |
|                                  | The maximum cost of newspaper preprints is not given in the context.        | ✓                 | ✓ | ✓ | ✓ |
| C                                | ! The context says 5 or 6 cents, not \$5.                                   | ×                 | × | ✓ | ✓ |

500 NLI items: 1,933 label-explanation pairs

Strict definition of what counts as an error

# VariErr: Error identification

P: Because marginal costs are very low, a newspaper price for preprints might be as low as 5 or 6 cents per piece.

H: Newspaper preprints can cost as much as \$5.

| Round 1: NLI Label & Explanation |   |                                                                             | Round 2: Validity |   |   |   |
|----------------------------------|---|-----------------------------------------------------------------------------|-------------------|---|---|---|
| L                                | A | Explanation                                                                 | 1                 | 2 | 3 | 4 |
| E                                | 4 | 5 dollars for a piece of newspaper.                                         | ×                 | × | × | × |
| N                                | 1 | The context only mentions how low the price may be, not how high it may be. | ✓                 | ✓ | ✓ | ✓ |
|                                  | 3 | The maximum cost of newspaper preprints is not given in the context.        | ✓                 | ✓ | ✓ | ✓ |
| C                                | 2 | The context says 5 or 6 cents, not \$5.                                     | ×                 | × | ✓ | ✓ |

self-validation

500 NLI items: 1,933 label-explanation pairs

Strict definition of what counts as an error

88.57% self-validated, 82.82% peer-validated

37% of items for which there was a label error

# Can we identify errors automatically?

Automatic error detection (AED) as a ranking task

## **State-of-the-art AED models**

2 variants of DataMaps [Swayamdipta et al. 2020]

Metadata Archaeology [Siddiqui et al. 2023]



# We tested several methods including heuristics

## Automatic error detection (AED) as a ranking task

### State-of-the-art AED n 2 variants of DataMaps

### Metadata Archaeology

### GPT-4

System:

You are an expert linguistic annotator.

User:

We have collected annotations for a NLI instance together with reasons for the labels. Your task is to judge whether the reasons make sense for the label. Provide the probability (0.0 - 1.0) that the reason makes sense for the label. Give ONLY the reason and the probability, no other words or explanation. For example:

Reason: <The verbatim copy of the reason>

Probability: <the probability between 0.0 and 1.0 that the reason makes sense for the label, without any extra commentary whatsoever; just the probability!>.

Context: {CONTEXT}

Statement: {STATEMENT}

Reason for label {LABEL}: {REASON\_1}

Reason for label {LABEL}: {REASON\_2}

[...]

Reason for label {LABEL}: {REASON\_n}

Reason {REASON\_1}

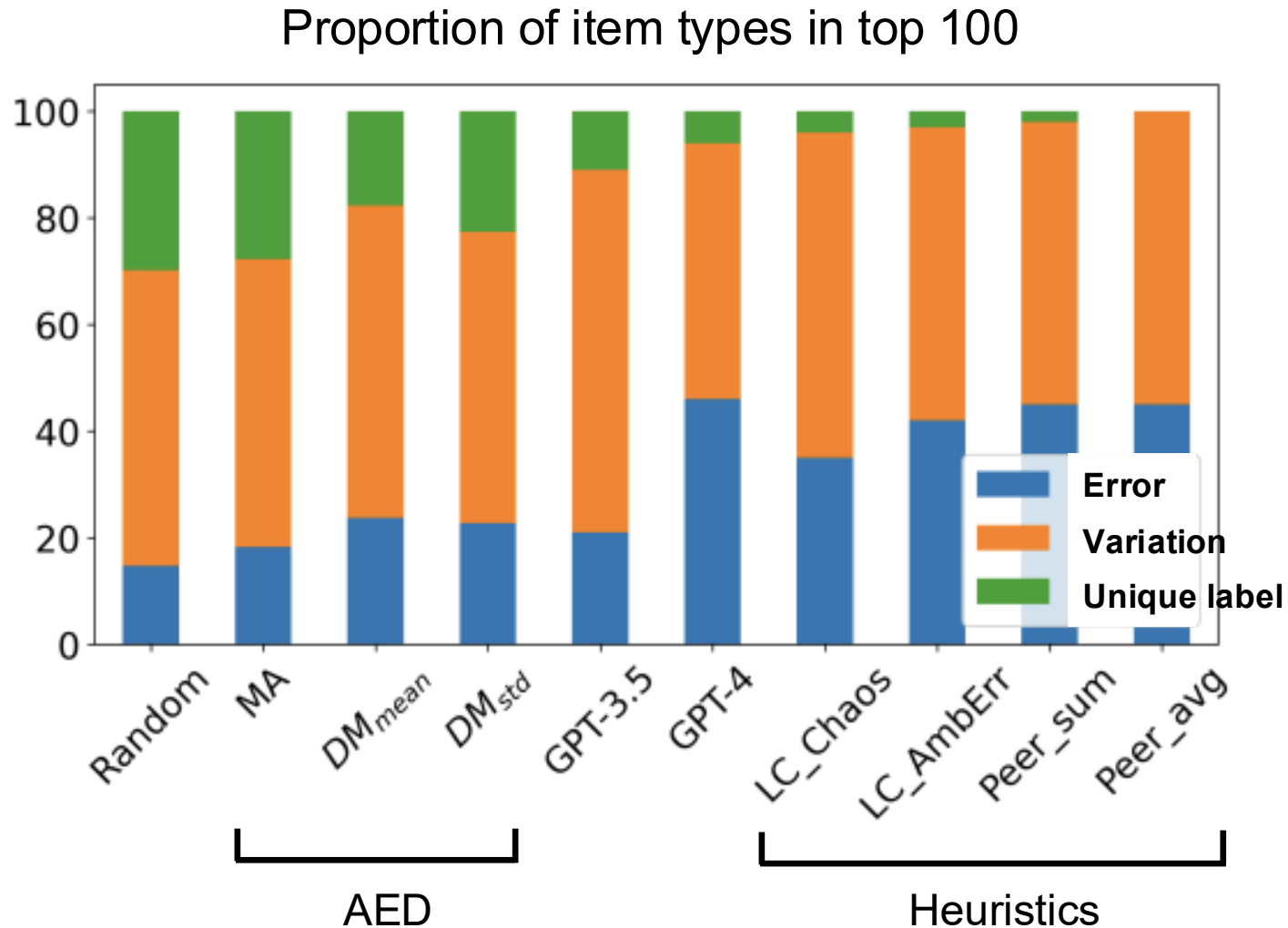
Probability:

# AED is challenging

Comparison of ranked lists to real errors

|                                  | <u>Average precision</u> |
|----------------------------------|--------------------------|
| DataMaps (mean)                  | 22.8                     |
| GPT-4                            | 31.3                     |
| Heuristics on ChaosNLI           | 32.5                     |
| Heuristics on VariErr (Peer_Sum) | 46.5                     |

# Where do AED methods make mistakes?



# Can LLMs help?

Evade: LLM-based explanations generation and validation for error detection in NLI

Longfei Zuo, Barbara Plank & Siyao Peng.  
Findings of ACL 2026.

Label and explanation variation in LLM-based annotation: a case study in Natural Language Inference

Artur Kulmizev, Erika Lombart, Patrick Watrin  
& MC de Marneffe. ACL 2026

**“Common sense” annotations without explanations are meaningless**

# Very different construals of the NLI task

P: It is, as you see, highly enlarged. [0.46, 0.27, 0.16, 0.11]

H: It is plain for you to see that it is amplified.

E – *Highly enlarged* can correspond to *amplified*.

E – Depending on the qualified object, *enlarged* and *amplified* can mean the same thing.

N – It all depends on what we're talking about, on the context. *Enlarged* and *amplified* are not strict synonyms.

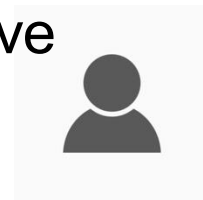
N – Depending on the meaning of "it", *enlarged* can or cannot be a synonym of *amplified*.

P: We were playing all sorts of sports, and you were not, so shut up and stop twitching,' the microbe's tone of voice changed, it was lower and more resounding.

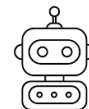
H: We have played football, basketball, baseball and soccer and you haven't played anything so shut your mouth.

Same distribution between the 20 LLMs and the humans [0.8, 0.2, 0]

N - The premise mentions sports without saying which ones, the hypothesis mentions specific sports that might or might not have been played.



N - The premise states 'all sorts of sports' without specifying which sports were played, while the statement lists particular sports like football, basketball, baseball, and soccer. Since the hypothesis does not confirm or deny the inclusion of these specific sports, hypothesis could be true if those sports were part of 'all sorts,' or false if other sports were played instead.

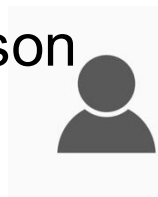


P: We were playing all sorts of sports, and you were not, so shut up and stop twitching,' the microbe's tone of voice changed, it was lower and more resounding.

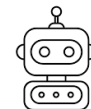
H: We have played football, basketball, baseball and soccer and you haven't played anything so shut your mouth.

Same distribution between the 20 LLMs and the humans [0.8, 0.2, 0]

E - They stated they played all kinds of sports so it could have been football, basketball, baseball or soccer and it's true that one person did not play.



E - The premise shows a conversation where one party is asserting dominance by claiming superior sports participation. The microbe seems to be directly telling another party that they have played multiple sports while the other person has not. The hypothesis precisely mirrors the premise, listing the same sports (football, basketball, baseball and soccer) and using the same confrontational language of telling someone to shut up.





P: We were playing all sorts of sports, and you were not, so shut up and stop twitching,' the microbe's tone of voice changed, it was lower and more resounding.

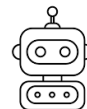
H: We have played football, basketball, baseball and soccer and you haven't played anything so shut your mouth.

Same distribution between the 20 LLMs and the humans [0.8, 0.2, 0]

E - They stated they played all kinds of sports so it could have been football, basketball, baseball or soccer and it's true that one person did not play.



E - The premise shows a conversation where one party is asserting dominance by claiming superior sports participation. The microbe seems to be directly telling another party that they have played multiple sports while the other person has not. The hypothesis precisely mirrors the premise, listing the same sports (football, basketball, baseball and soccer) and using the same confrontational language of telling someone to shut up.



P: Do you want to see historic sights, tour museums and art galleries?

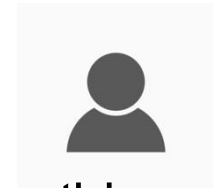
H: You wouldn't like to visit historic places, museums, and art galleries, would you?

[Groenendijk and Stokhof 1984]

|                                 |        |               |
|---------------------------------|--------|---------------|
| Different distributions between | LLMs   | [0, 0.6, 0.4] |
|                                 | humans | [0.5, 0.5, 0] |

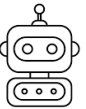
E - The premise and the hypothesis say basically the same thing, but one asks in a positive way and the other in a negative way.

E - The heart of both of these questions seems to match.



N - The premise is asking if someone would like to do these things, and there is no indication of whether or not they do.

N - The premise asks whether the person desires to experience historic sights, museums, and art galleries, but provides no information about their actual preference. The hypothesis asserts a negative preference with wouldn't like to visit and references equivalent attractions. Since the context only inquires without revealing the person's inclination, the hypothesis could be true if the person dislikes these activities or false if they enjoy them.



P: Do you want to see historic sights, tour museums and art galleries?

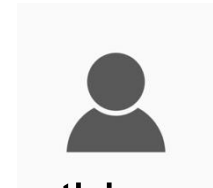
H: You wouldn't like to visit historic places, museums, and art galleries, would you?

[Groenendijk and Stokhof 1984]

|                                 |        |               |
|---------------------------------|--------|---------------|
| Different distributions between | LLMs   | [0, 0.6, 0.4] |
|                                 | humans | [0.5, 0.5, 0] |

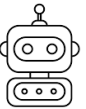
E - The premise and the hypothesis say basically the same thing, but one asks in a positive way and the other in a negative way.

E - The heart of both of these questions seems to match.



N - The premise is asking if someone would like to do these things, and there is no indication of whether or not they do.

N - The premise asks whether the person desires to experience historic sights, museums, and art galleries, but provides no information about their actual preference. The hypothesis asserts a negative preference with wouldn't like to visit and references equivalent attractions. Since the context only inquires without revealing the person's inclination, the hypothesis could be true if the person dislikes these activities or false if they enjoy them.

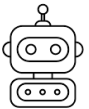


P: Do you want to see historic sights, tour museums and art galleries?

H: You wouldn't like to visit historic places, museums, and art galleries, would you?

Different distribution between LLMs [0, 0.6, 0.4]  
humans [0.5, 0.5, 0]

C - The premise asks if the person wants to see historic sights and tour museums and art galleries, which suggests the person has an interest in these activities. The hypothesis asserts the opposite—that the person wouldn't like to visit historic places, museums, and art galleries—which contradicts the interest expressed or implied in the premise. Therefore, if we assume the premise is true, the hypothesis is most likely false.



C - The premise expresses interest in visiting historic sights, museums, and art galleries, while the hypothesis suggests the opposite, implying disinterest. The key difference lies in the positive tone of the premise ('want to see') versus the negative tone of the hypothesis ('wouldn't like to visit'). This contradiction makes the hypothesis most likely false if the premise is true.

# Wrapping up: Pendulums swung too far?

LAW X:

Represent distribution in annotations as it is signal

LAW XX:

There is still noise in there that we should discard, especially in this new LLM era

Tapping into annotation decisions through explanations

LAW XXX?

Efficient collaboration with LLMs?

# Thank you!

[github.com/njjiang/LiveNLI](https://github.com/njjiang/LiveNLI)  
[github.com/mainlp/VariErr-NLP](https://github.com/mainlp/VariErr-NLP)

