COSI-230B: Natural Language Annotation for Machine Learning

Lecture 8: Inter-Annotator Agreement I

Jin Zhao

Brandeis University

February 11, 2026

Today's Agenda

1. Why measure agreement?
2. Observed agreement and its problems
3. Cohen's Kappa: formula and step-by-step calculation
4. Practice: binary and multi-class Kappa
5. When Kappa can be misleading (with examples)
6. Per-category Kappa

Why Measure Agreement?

IAA serves multiple purposes:

1. **Quality assessment:** How reliable is our data?
2. **Task validation:** Is the task well-defined?
3. **Guideline evaluation:** Are guidelines clear?
4. **Annotator calibration:** Are annotators consistent?
5. **Upper bound:** What's the best model can achieve?

Key insight: If humans can't agree, models can't learn reliably

Observed Agreement

Simplest measure: Proportion of items where annotators agree

$$A_o = \frac{\text{Number of agreements}}{\text{Total items}}$$

Example:

Item	Ann. 1	Ann. 2	Agree?
1	Positive	Positive	✓
2	Negative	Negative	✓
3	Positive	Neutral	×
4	Negative	Negative	✓
5	Neutral	Neutral	✓

$$A_o = \frac{4}{5} = 0.80$$

Problem: Chance Agreement

Observed agreement doesn't account for chance

Example: Binary task (Yes/No)

- Two annotators randomly guessing
- Each says "Yes" 50% of the time
- Expected agreement by chance: 50%

If we observe 70% agreement:

- Seems good, but...
- Only 20% above chance
- Is that meaningful?

Solution: Chance-corrected measures like Kappa

Cohen's Kappa: The Formula

Kappa corrects for chance agreement

$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

Where:

- A_o = Observed agreement
- A_e = Expected agreement by chance
- κ = Kappa coefficient

Intuition: Agreement beyond chance, normalized by maximum possible agreement beyond chance

Computing Expected Agreement

A_e based on marginal distributions

Confusion matrix:

	Ann. 2: Yes	Ann. 2: No	Total
Ann. 1: Yes	a	b	$a+b$
Ann. 1: No	c	d	$c+d$
Total	$a+c$	$b+d$	n

$$A_e = P(\text{both Yes}) + P(\text{both No})$$

$$A_e = \frac{(a+b)(a+c)}{n^2} + \frac{(c+d)(b+d)}{n^2}$$

Why Compute by Hand at Least Once?

Python can compute Kappa in one line. Why bother doing it manually?

1. **Build intuition for A_e:** The formula's power comes from expected agreement — you need to *feel* how marginals drive chance agreement
2. **Diagnose surprising results:** When Kappa seems too low or too high, you need to trace *which component* is responsible
3. **Understand the paradoxes:** Later we'll see cases where 94% agreement gives $\kappa = 0.37$ — this only makes sense if you understand the mechanics
4. **Sanity-check your code:** If you can't ballpark Kappa from a confusion matrix, you can't catch bugs in your pipeline

Goal: After today, you'll use Python for computation but *understand* what the numbers mean

Step-by-Step Kappa Calculation

Example: Sentiment annotation (Pos/Neg), 50 items

	Ann. 2: Pos	Ann. 2: Neg	Total
Ann. 1: Pos	20	5	25
Ann. 1: Neg	10	15	25
Total	30	20	50

Goal: Calculate Cohen's Kappa for this data

Three steps: Observed agreement $\rightarrow$ Expected agreement $\rightarrow$ Kappa

Step 1: Observed Agreement

	Ann. 2: Pos	Ann. 2: Neg	Total
Ann. 1: Pos	**20**	5	25
Ann. 1: Neg	10	**15**	25
Total	30	20	50

Observed agreement = items where both annotators chose the same label

$$A_o = \frac{\text{diagonal sum}}{\text{total}} = \frac{20 + 15}{50} = \frac{35}{50} = 0.70$$

Annotators agreed on 70% of items

Step 2: Expected Agreement

	Ann. 2: Pos	Ann. 2: Neg	**Total**
Ann. 1: Pos	20	5	**25**
Ann. 1: Neg	10	15	**25**
Total	**30**	**20**	50

Expected agreement = probability both choose same label by chance

- P(both Pos) $= \frac{25}{50} \times \frac{30}{50} = 0.50 \times 0.60 = 0.30$
- P(both Neg) $= \frac{25}{50} \times \frac{20}{50} = 0.50 \times 0.40 = 0.20$

$$A_e = 0.30 + 0.20 = 0.50$$

Step 3: Compute Kappa

From previous steps:

- Observed agreement: $A_o = 0.70$
- Expected agreement: $A_e = 0.50$

Apply the formula:

$$\kappa = \frac{A_o - A_e}{1 - A_e} = \frac{0.70 - 0.50}{1 - 0.50} = \frac{0.20}{0.50} = 0.40$$

Interpretation

$\kappa = 0.40 \rightarrow$ **Fair agreement** (Landis & Koch scale)

70% raw agreement, but half was expected by chance. Only moderate real agreement.

Your Turn: Compute Kappa

Task: NER annotation (Entity / Not Entity), 80 items

	Ann. 2: Entity	Ann. 2: Not	Total
Ann. 1: Entity	30	10	40
Ann. 1: Not	8	32	40
Total	38	42	80

Calculate on paper:

1. Observed agreement A_o
2. Expected agreement A_e
3. Cohen's Kappa κ
4. Interpret the result using the Landis & Koch scale

Solution

Step 1: $A_o = \frac{30+32}{80} = \frac{62}{80} = 0.775$

Step 2:

- P(both Entity) $= \frac{40}{80} \times \frac{38}{80} = 0.50 \times 0.475 = 0.2375$
- P(both Not) $= \frac{40}{80} \times \frac{42}{80} = 0.50 \times 0.525 = 0.2625$

$A_e = 0.2375 + 0.2625 = 0.50$

Step 3: $\kappa = \frac{0.775-0.50}{1-0.50} = \frac{0.275}{0.50} = 0.55$

Interpretation

$\kappa = 0.55 \rightarrow$ **Moderate agreement**

Better than the sentiment example ($\kappa = 0.40$). The more balanced marginals here help — both annotators labeled close to 50/50.

Interpreting Kappa Values

Common interpretation guidelines (Landis & Koch, 1977):

Kappa	Interpretation
< 0	Less than chance agreement
$0.00 - 0.20$	Slight agreement
$0.21 - 0.40$	Fair agreement
$0.41 - 0.60$	Moderate agreement
$0.61 - 0.80$	Substantial agreement
$0.81 - 1.00$	Almost perfect agreement

Caution: These are rough guidelines, not absolute thresholds

What Kappa Should You Aim For?

Depends on the task:

- **Objective tasks** (POS tagging): $\kappa > 0.80$
- **Standard annotation** (NER, sentiment): $\kappa > 0.70$
- **Subjective tasks** (quality, safety): $\kappa > 0.50$
- **Highly subjective** (sarcasm, humor): $\kappa > 0.40$

Context matters:

- Compare to published baselines for similar tasks
- Consider task difficulty
- Low Kappa may indicate unclear guidelines, not bad annotators

Kappa for Multi-Class

Same formula, larger confusion matrix

For k categories:

$$A_o = \frac{1}{n}\sum_{i=1}^{k} n_{ii} \qquad A_e = \sum_{i=1}^{k} \frac{n_{i\cdot} \times n_{\cdot i}}{n^2}$$

Where n_{ii} = diagonal count, $n_{i\cdot}$ = row total, $n_{\cdot i}$ = column total

Same intuition: sum agreements on the diagonal, compare to chance from marginals

Your Turn: Multi-Class Kappa

3-class sentiment: Positive / Neutral / Negative, 100 items

	Pos	Neu	Neg	Total
Pos	30	5	5	40
Neu	3	20	2	25
Neg	2	3	30	35
Total	35	28	37	100

Calculate on paper:

1. Observed agreement A_o (hint: sum the diagonal)
2. Expected agreement A_e (hint: one term per category)
3. Cohen's Kappa κ and interpretation

Solution: Multi-Class Kappa

Step 1: Observed agreement (diagonal: 30 + 20 + 30)

$A_o = \frac{30+20+30}{100} = 0.80$

Step 2: Expected agreement (one term per category)

$$A_e = \underbrace{\frac{40 \times 35}{10000}}_{Pos:0.14} + \underbrace{\frac{25 \times 28}{10000}}_{Neu:0.07} + \underbrace{\frac{35 \times 37}{10000}}_{Neg:0.13} = 0.339$$

Step 3: Kappa

$\kappa = \frac{0.80-0.339}{1-0.339} = \frac{0.461}{0.661} = 0.697$ (**Substantial** agreement)

Takeaway

The multi-class formula is the same as binary — just more terms in A_e. With 3 balanced categories, chance agreement is lower (0.34 vs. 0.50 for binary), so Kappa has more room.

When Kappa Can Be Misleading

Kappa paradoxes:

1. **High agreement, low Kappa:**
   - When one category dominates (class imbalance)
   - High A_e reduces Kappa
2. **Symmetric vs. asymmetric disagreement:**
   - Kappa treats all disagreements equally
   - Some disagreements may be more problematic
3. **Prevalence effect:**
   - Rare categories have outsized impact

Recommendation: Report Kappa AND confusion matrix

Example: High Agreement, Low Kappa

Toxicity detection: 100 comments, only 5% toxic

	Ann. 2: Toxic	Ann. 2: Safe	Total
Ann. 1: Toxic	2	3	5
Ann. 1: Safe	3	92	95
Total	5	95	100

Step 1: $A_o = \frac{2+92}{100} = 0.94$ (94% agreement — looks great!)

Step 2: $A_e = \frac{5\times5}{100^2} + \frac{95\times95}{100^2} = 0.0025 + 0.9025 = 0.905$

Step 3: $\kappa = \frac{0.94-0.905}{1-0.905} = \frac{0.035}{0.095} = 0.37$ (**Fair** agreement!)

The paradox

94% raw agreement but $\kappa = 0.37$. Both annotators say "Safe" 95% of the time, so chance agreement is already 90.5%. The high agreement is mostly base rate, not real consistency on the Toxic class.

Example: Symmetric vs. Asymmetric Disagreement

Both matrices: 100 items, 80% agreement, 20 disagreements

Symmetric (10 each way):

	Pos	Neg	Tot
Pos	40	**10**	50
Neg	**10**	40	50
Tot	50	50	100

$A_e = 0.50, \quad \kappa = 0.60$

Annotators are equally uncertain

Asymmetric (18 vs. 2):

	Pos	Neg	Tot
Pos	40	**18**	58
Neg	**2**	40	42
Tot	42	58	100

$A_e = 0.49, \quad \kappa = 0.61$

Ann. 1 has a strong Pos bias

The paradox

Nearly identical κ (0.60 vs. 0.61), but very different problems. Symmetric = genuine ambiguity. Asymmetric = one annotator is systematically biased. **Kappa can't tell the difference** — always check the confusion matrix.

Example: The Prevalence Effect

Same 85% agreement, different category balance

Balanced (50/50 split):

	Pos	Neg	Tot
Pos	43	7	50
Neg	8	42	50
Tot	51	49	100

$A_e = 0.50$

$\kappa = \frac{0.85-0.50}{1-0.50} = \mathbf{0.70}$ (Substantial)

Imbalanced (90/10 split):

	Pos	Neg	Tot
Pos	80	10	90
Neg	5	5	10
Tot	85	15	100

$A_e = 0.78$

$\kappa = \frac{0.85-0.78}{1-0.78} = \mathbf{0.32}$ (Fair)

The paradox

Identical 85% raw agreement, but $\kappa = 0.70$ vs. $\kappa = 0.32$. When one category dominates, A_e is high, leaving little room for Kappa. **Compare Kappa across tasks with similar prevalence**, or use per-category Kappa to diagnose specific labels.

Per-Category Kappa

Idea: Binarize each category ("this label vs. all others"), compute κ separately

3-class sentiment, 100 items:

	Pos	Neu	Neg	Tot
Pos	35	8	2	45
Neu	5	10	10	25
Neg	0	7	23	30
Tot	40	25	35	100

Overall $\kappa = 0.51$ (Moderate)

Per-category breakdown:

Category	κ	Interpretation
Positive	0.69	Substantial
Neutral	**0.20**	**Slight**
Negative	0.57	Moderate
Overall	0.51	Moderate

E.g., Neutral binarized: $A_o = 0.70$, $A_e = 0.625$, $\kappa = 0.20$

Why this matters

Overall $\kappa = 0.51$ looks acceptable. But Neutral has $\kappa = 0.20$ — annotators can barely agree on it. **Fix the Neutral guidelines** before collecting more data.

Using Python for Kappa

scikit-learn implementation:

```
from sklearn.metrics import cohen_kappa_score

ann1 = ['pos', 'neg', 'pos', 'neg', 'neu']
ann2 = ['pos', 'neg', 'neu', 'neg', 'neu']

kappa = cohen_kappa_score(ann1, ann2)
print(f"Kappa: {kappa:.3f}")
```

Output: `Kappa: 0.600`

Reporting Agreement

What to include in your report:

1. **Kappa value** with interpretation
2. **Observed agreement** for context
3. **Confusion matrix** for detailed view
4. **Per-class agreement** if relevant
5. **Sample size** (number of items annotated)
6. **Number of annotators**

Example: "Two annotators labeled 500 items for sentiment. Cohen's Kappa was 0.72 (substantial agreement) with observed agreement of 85%."

What If Agreement Is Low?

Diagnosis and remediation:

1. **Unclear guidelines:**
   - Add examples and edge cases
   - Clarify ambiguous definitions
2. **Insufficient training:**
   - More annotator calibration sessions
   - Practice on sample data
3. **Task too subjective:**
   - Simplify categories
   - Accept lower agreement threshold
4. **Bad annotators:**
   - Review individual performance
   - Remove or retrain outliers

Next Class: IAA II

Lecture 9: Inter-Annotator Agreement II

Topics:

- Fleiss' Kappa (multiple annotators)
- Krippendorff's Alpha
- Agreement for spans and rankings
- Human-LLM agreement measurement
- LLM self-consistency as quality signal

Reading: Artstein (2017) - Handbook chapter on IAA

Key Takeaways

1. **IAA** measures annotation quality and task clarity
2. **Observed agreement** doesn't account for chance
3. **Cohen's Kappa** corrects for chance agreement
4. $\kappa = \frac{A_o - A_e}{1 - A_e}$ measures agreement beyond chance
5. **Interpretation** depends on task type
6. **Low agreement** may indicate guideline or task problems

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu