Corpus Linguistics: Inter-Annotation Agreements

Karën Fort

December 15, 2011
Sources

Most of this course is largely inspired by:

- THE reference article: **Inter-Coder Agreement for Computational Linguistics** [Artstein and Poesio, 2008]
- Massimo Poesio’s presentation at LREC on the same subject
- Gemma Boleda and Stefan Evert’s course on the same subject (ESSLLI 2009) [http://esslli2009.labri.fr/course.php?id=103]
- Cyril Grouin’s course on the measures used in evaluation protocols [http://perso.limsi.fr/grouin/inalco/1011/]
Introduction

Crucial issue: **Are the annotations correct?**

- ML learns to make same mistakes as human annotator (noise ≠ patterns in errors [Reidsma and Carletta, 2008])
- Misleading evaluation
- Inconclusive and misleading results from linguistic analysis and hand-crafted systems
We are interested in the **validity** of the manual annotation
- i.e. whether the annotated categories are correct

But there is no “ground truth”
- Linguistic categories are determined by human judgment
- Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation
- i.e. whether human annotators consistently make same decisions ⇒ they have internalized the scheme
- Assumption: high reliability implies validity

How can reliability be determined?
We are interested in the **validity** of the manual annotation

- i.e. whether the annotated categories are correct

But there is no “ground truth”

- Linguistic categories are determined by human judgment
- Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation

- i.e. whether human annotators consistently make same decisions
  - they have internalized the scheme
- Assumption: high reliability implies validity

How can reliability be determined?
We are interested in the **validity** of the manual annotation
i.e. whether the annotated categories are correct

But there is no “ground truth”
- Linguistic categories are determined by human judgment
- Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation
i.e. whether human annotators consistently make same decisions
they have internalized the scheme
- Assumption: high reliability implies validity

How can reliability be determined?
We are interested in the **validity** of the manual annotation
  i.e. whether the annotated categories are correct

But there is no “ground truth”
  Linguistic categories are determined by human judgment
  Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation
  i.e. whether human annotators consistently make same decisions \( \Rightarrow \)
  they have internalized the scheme
  Assumption: high reliability implies validity

How can reliability be determined?
We are interested in the **validity** of the manual annotation

i.e. whether the annotated categories are correct

But there is no “ground truth”

- Linguistic categories are determined by human judgment
- Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation

i.e. whether human annotators consistently make same decisions
they have internalized the scheme

Assumption: high reliability implies validity

How can reliability be determined?
We are interested in the **validity** of the manual annotation
- i.e. whether the annotated categories are correct

But there is no “ground truth”
- Linguistic categories are determined by human judgment
- Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation
- i.e. whether human annotators **consistently** make same decisions ⇔ they have internalized the scheme
- Assumption: high reliability implies validity

How can reliability be determined?
Motivations

Validity vs. Reliability [Artstein and Poesio, 2008]

- We are interested in the **validity** of the manual annotation
  - i.e. whether the annotated categories are correct
- But there is no “ground truth”
  - Linguistic categories are determined by human judgment
  - Consequence: we cannot measure correctness directly
- Instead measure **reliability** of annotation
  - i.e. whether human annotators **consistently** make same decisions ⇒ they have internalized the scheme
  - Assumption: high reliability implies validity
- How can reliability be determined?
We are interested in the **validity** of the manual annotation
  - i.e. whether the annotated categories are correct

But there is no “ground truth”
  - Linguistic categories are determined by human judgment
  - Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation
  - i.e. whether human annotators **consistently** make same decisions \(\Rightarrow\) they have internalized the scheme
  - Assumption: high reliability implies validity

How can reliability be determined?
We are interested in the **validity** of the manual annotation
- i.e. whether the annotated categories are correct

But there is no “ground truth”
- Linguistic categories are determined by human judgment
- Consequence: we cannot measure correctness directly

Instead measure **reliability** of annotation
- i.e. whether human annotators **consistently** make same decisions ⇒ they have internalized the scheme
- Assumption: high reliability implies validity

How can reliability be determined?
Achieving Reliability (consistency)

- each item is annotated by a single annotator, with random checks ($\approx$ second annotation)
- some of the items are annotated by two or more annotators
- each item is annotated by two or more annotators - followed by reconciliation
- each item is annotated by two or more annotators - followed by final decision by superannotator (expert)

In all cases, measure of reliability: coefficients of agreement
In some (rare and often artificial) cases, there exists a “reference”: the corpus was annotated, at least partly, and this annotation is considered “perfect”, a reference [Fort and Sagot, 2010].

In those cases, another, complementary measure, can be used:

Which one?
In some (rare and often artificial) cases, there exists a “reference”: the corpus was annotated, at least partly, and this annotation is considered “perfect”, a reference [Fort and Sagot, 2010].

In those cases, another, complementary measure, can be used:

F-measure
Precision/Recall: back to basics

- Recall:

- Silence:

- Precision:

- Noise:
Precision/Recall: back to basics

- **Recall**: measures the quantity of found annotations
  \[
  \text{Recall} = \frac{\text{Nb of correct found annotations}}{\text{Nb of correct expected annotations}}
  \]

- **Silence**:

- **Precision**:

- **Noise**:
Precision/Recall: back to basics

- **Recall**: measures the quantity of found annotations
  \[
  \text{Recall} = \frac{\text{Nb of correct found annotations}}{\text{Nb of correct expected annotations}}
  \]
- **Silence**: complement of recall (correct annotations not found)
- **Precision**:

- **Noise**:
Precision/Recall: back to basics

- **Recall**: measures the quantity of found annotations

  \[
  \text{Recall} = \frac{\text{Nb of correct found annotations}}{\text{Nb of correct expected annotations}}
  \]

- **Silence**: complement of recall (correct annotations not found)

- **Precision**: measures the quality of found annotations

  \[
  \text{Precision} = \frac{\text{Nb of correct found annotations}}{\text{Total nb of found annotations}}
  \]

- **Noise:**
Precision/Recall: back to basics

- **Recall**: measures the quantity of found annotations
  
  \[
  \text{Recall} = \frac{\text{Nb of correct found annotations}}{\text{Nb of correct expected annotations}}
  \]

- **Silence**: complement of recall (correct annotations not found)

- **Precision**: measures the quality of found annotations
  
  \[
  \text{Precision} = \frac{\text{Nb of correct found annotations}}{\text{Total nb of found annotations}}
  \]

- **Noise**: complement of precision (incorrect annotations found)
F-measure: back to basics (Wikipedia Dec. 10, 2010)

Harmonic mean of precision and recall or balanced F-score

\[ F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]

... aka the F1 measure, because recall and precision are evenly weighted.

It is a special case of the general F_\beta measure:

\[ F_\beta = (1 + \beta^2) \times \frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \]

The value of \( \beta \) allows to favor:

- recall (\( \beta = 2 \))
- precision (\( \beta = 0.5 \))
**A little more from biology and medicine**

True and false, positive and negative:

<table>
<thead>
<tr>
<th></th>
<th>Disease is present</th>
<th>Disease is absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative test</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**A little more from biology and medicine**

**True and false, positive and negative:**

<table>
<thead>
<tr>
<th>Disease is present</th>
<th>Disease is absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td><strong>TP</strong></td>
</tr>
<tr>
<td>Negative test</td>
<td><strong>TN</strong></td>
</tr>
</tbody>
</table>
### A little more from biology and medicine

True and false, positive and negative:

<table>
<thead>
<tr>
<th></th>
<th>Disease is present</th>
<th>Disease is absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive test</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Negative test</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>
A little more from biology and medicine

- **sensitivity**: corresponds to recall
  \[ SE = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \]

- **specificity**: rate of true negatives
  \[ SP = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \]

- **selectivity**: corresponds to precision
  \[ SEL = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \]

- **accuracy**: nb of correct predictions over the total nb of predictions
  \[ ACC = \frac{\text{true positives} + \text{true negatives}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \]
A little more from biology and medicine

- **sensitivity**: corresponds to recall
  \[
  SE = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}
  \]

- **specificity**: rate of true negatives
  \[
  SP = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}
  \]

- **selectivity**: corresponds to precision
  \[
  SEL = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}
  \]

- **accuracy**: nb of correct predictions over the total nb of predictions
  \[
  ACC = \frac{\text{true positives} + \text{true negatives}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}
  \]
Motivations

A little more from biology and medicine

- **sensitivity**: corresponds to recall
  \[ SE = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \]

- **specificity**: rate of true negatives
  \[ SP = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \]

- **selectivity**: corresponds to precision
  \[ SEL = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \]

- **accuracy**: nb of correct predictions over the total nb of predictions
  \[ ACC = \frac{\text{true positives} + \text{true negatives}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \]
A little more from biology and medicine

- **sensitivity**: corresponds to recall

  \[ SE = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \]

- **specificity**: rate of true negatives

  \[ SP = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} \]

- **selectivity**: corresponds to precision

  \[ SEL = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \]

- **accuracy**: nb of correct predictions over the total nb of predictions

  \[ ACC = \frac{\text{true positives} + \text{true negatives}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \]
Does a “Gold-standard” exist?

- reference rarely pre-exists
- can it be “perfect”? [Fort and Sagot, 2010]
  - can we use F-measure in other cases? Reading for next class!
  ⇒ Back to coefficients of agreement.
Easy and Hard Tasks


Objective tasks
- Decision rules, linguistic tests
- Annotation guidelines with discussion of boundary cases
- POS tagging, syntactic annotation, segmentation, phonetic transcription, . . .

Subjective tasks
- Based on speaker intuitions
- Short annotation instructions
- Lexical semantics (subjective interpretation!), discourse annotation & pragmatics, subjectivity analysis, . . .
Easy and Hard Tasks


Objective tasks

- Decision rules, linguistic tests
- Annotation guidelines with discussion of boundary cases
- POS tagging, syntactic annotation, segmentation, phonetic transcription, . . .

$\rightarrow$ IAA $= 98.5\%$ (POS tagging)
IAA $\approx 93.0\%$ (syntax)

Subjective tasks

- Based on speaker intuitions
- Short annotation instructions
- Lexical semantics (subjective interpretation!), discourse annotation & pragmatics, subjectivity analysis, . . .
Easy and Hard Tasks


Objective tasks

- Decision rules, linguistic tests
- Annotation guidelines with discussion of boundary cases
- POS tagging, syntactic annotation, segmentation, phonetic transcription, . . .

→ IAA = 98.5% (POS tagging)
  IAA ≈ 93.0% (syntax)

Subjective tasks

- Based on speaker intuitions
- Short annotation instructions
- Lexical semantics (subjective interpretation!), discourse annotation & pragmatics, subjectivity analysis, . . .

→ IAA = 68.6% (HW)
  IAA ≈ 70% (word senses)
## Example

<table>
<thead>
<tr>
<th>Sentence</th>
<th>A</th>
<th>B</th>
<th>Agree?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Put tea in a heat-resistant jug and add the boiling water.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Where are the batteries kept in a phone?</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Vinegar’s usefulness doesn’t stop inside the house.</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>How do I recognize a room that contains radioactive materials?</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>A letterbox is a plastic, screw-top bottle that contains a small notebook and a unique rubber stamp.</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

→ Agreement?
Contingency Table and Observed Agreement

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>No</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

Observed Agreement ($A_o$)

proportion of items on which 2 annotators agree.

Here:
### Observed Agreement (A₀)

Proportion of items on which 2 annotators agree.

Here: \( A₀ = \frac{4+2}{10} = 0.6 \)
Chance Agreement

Some agreement is expected by chance alone:

*In our case, what proportion of agreement is expected by chance?*
Some agreement is expected by chance alone:

- Two annotators randomly assigning “Yes” and ”No” labels will agree half of the time (0.5 can be obtained purely by chance: what does it mean for our result?).
- The amount expected by chance varies depending on the annotation scheme and on the annotated data.

Meaningful agreement is the agreement above chance.
→ Similar to the concept of “baseline“ for system evaluation.
Taking Chance into Account

Expected Agreement ($A_e$)

- expected value of observed agreement.

Amount of agreement above chance: $A_o - A_e$

Maximum possible agreement above chance: $1 - A_e$

Proportion of agreement above chance attained: $\frac{A_o - A_e}{1 - A_e}$

Perfect agreement: $\frac{1 - A_e}{1 - A_e}$

Perfect disagreement: $\frac{-A_e}{1 - A_e}$
Expected Agreement

How to compute the amount of agreement expected by chance ($A_e$)?
S [Bennett et al., 1954]

Same chance for all annotators and categories.

Number of category labels: $q$
Probability of one annotator picking a particular category $q_a$: $\frac{1}{q}$
Probability of both annotators picking a particular category $q_a$: $\left(\frac{1}{q}\right)^2$

Probability of both annotators picking the same category:

$$A_e^S = q \cdot \left(\frac{1}{q}\right)^2 = \frac{1}{q}$$
All the categories are equally likely: consequences

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>
All the categories are equally likely: consequences

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

\[ A_o = \frac{20 + 20}{50} = 0.8 \]
\[ A_e = \frac{1}{2} = 0.5 \]
\[ S = \frac{0.8 - 0.5}{1 - 0.5} = 0.6 \]
All the categories are equally likely: consequences

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
A_o = \frac{20 + 20}{50} = 0.8 \\
A_e = \frac{1}{2} = 0.5 \\
S = \frac{0.8 - 0.5}{1 - 0.5} = 0.6
\]
All the categories are equally likely: consequences

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
A_o = \frac{20 + 20}{50} = 0.8
\]
\[
A_e = \frac{1}{2} = 0.5
\]
\[
S = \frac{0.8 - 0.5}{1 - 0.5} = 0.6
\]

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>C</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
A_o = \frac{20 + 20}{50} = 0.8
\]
\[
A_e = \frac{1}{4} = 0.25
\]
\[
S = \frac{0.8 - 0.25}{1 - 0.25} = 0.73
\]
Different chance for different categories.

Total number of judgments: $N$
Probability of one annotator picking a particular category $q_a$: $\frac{n_{qa}}{N}$
Probability of both annotators picking a particular category $q_a$: $\left(\frac{n_{qa}}{N}\right)^2$

Probability of both annotators picking the same category:

$$A_e^\pi = \sum_q \left(\frac{n_q}{N}\right)^2 = \frac{1}{N^2} \sum_q n_q^2$$
Comparing $S$ and $\pi$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

$A_o = 0.8$
$S = 0.6$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>C</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

$A_o = 0.8$
$S = 0.73$
Comparing $S$ and $\pi$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>C</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

$A_o = 0.8$

$S = 0.6$

$A_e^\pi = \frac{((\frac{25+25}{2})^2+(\frac{25+25}{2})^2)}{50^2} = 0.5$

$\pi = \frac{0.8-0.5}{1-0.5} = 0.6$

$A_o = 0.8$

$S = 0.73$
## Comparing $S$ and $\pi$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>C</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

\[ A_o = 0.8 \]
\[ S = 0.6 \]
\[ A_e^{\pi} = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5 \]
\[ \pi = \frac{0.8-0.5}{1-0.5} = 0.6 \]

\[ A_o = 0.8 \]
\[ S = 0.73 \]
\[ A_e^{\pi} = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5 \]
\[ \pi = \frac{0.8-0.5}{1-0.5} = 0.6 \]
Different annotators have different interpretations of the instructions (bias/prejudice). $\kappa$ takes individual bias into account.

Total number of items: $i$

Probability of one annotator $A_x$ picking a particular category $q_a$: $\frac{n_{A_x q_a}}{i}$

Probability of both annotators picking a particular category $q_a$: $\frac{n_{A_1 q_a}}{i} \cdot \frac{n_{A_2 q_a}}{i}$

Probability of both annotators picking the same category:

$$A_e^\kappa = \sum_q \frac{n_{A_1 q}}{i} \cdot \frac{n_{A_2 q}}{i} = \frac{1}{i^2} \sum_q n_{A_1 q} n_{A_2 q}$$
Comparing $\pi$ and $\kappa$


<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

$$A_o = 0.8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$$

$$\pi = \frac{0.8-0.5}{1-0.5} = 0.6$$


<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>C</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

$$A_o = 0.8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$$

$$\pi = \frac{0.8-0.5}{1-0.5} = 0.6$$
Comparing $\pi$ and $\kappa$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

$A_o = 0.8$

$A^\pi_e = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$

$\pi = \frac{0.8 - 0.5}{1 - 0.5} = 0.6$

$A^\kappa_e = \frac{(\frac{25\times25}{50}) + (\frac{25\times25}{50})}{50} = 0.5$

$\kappa = \frac{0.8 - 0.5}{1 - 0.5} = 0.6$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>C</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

$A_o = 0.8$

$A^\pi_e = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$

$\pi = \frac{0.8 - 0.5}{1 - 0.5} = 0.6$
Comparing $\pi$ and $\kappa$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
A_o = 0.8 \\
A^\pi_e = \frac{\left(\frac{25+25}{2}\right)^2 + \left(\frac{25+25}{2}\right)^2}{50^2} = 0.5 \\
\pi = \frac{0.8-0.5}{1-0.5} = 0.6 \\
A^\kappa_e = \frac{\left(\frac{25\times25}{50}\right) + \left(\frac{25\times25}{50}\right)}{50} = 0.5 \\
\kappa = \frac{0.8-0.5}{1-0.5} = 0.6
\]

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>C</th>
<th>D</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
</tbody>
</table>

\[
A_o = 0.8 \\
A^\pi_e = \frac{\left(\frac{25+25}{2}\right)^2 + \left(\frac{25+25}{2}\right)^2}{50^2} = 0.5 \\
\pi = \frac{0.8-0.5}{1-0.5} = 0.6 \\
A^\kappa_e = \frac{\left(\frac{25\times25}{50}\right) + \left(\frac{25\times25}{50}\right)}{50} = 0.5 \\
\kappa = \frac{0.8-0.5}{1-0.5} = 0.6
\]
Comparing $\pi$ and $\kappa$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

$$A_o = 0.8$$
$$A_e^{\pi} = \frac{\left(\frac{25+25}{2}\right)^2 + \left(\frac{25+25}{2}\right)^2}{50^2} = 0.5$$
$$\pi = \frac{0.8-0.5}{1-0.5} = 0.6$$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>24</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>32</td>
<td>70</td>
</tr>
</tbody>
</table>

$$A_o = 0.68$$
$$A_e^{\pi} = \frac{\left(\frac{38+32}{2}\right)^2 + \left(\frac{32+38}{2}\right)^2}{70^2} = 0.5$$
$$\pi = \frac{0.68-0.5}{1-0.5} = 0.36$$
### Comparing $\pi$ and $\kappa$

#### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

#### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>24</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>32</td>
<td>70</td>
</tr>
</tbody>
</table>

#### Formulas

- $A_o = 0.8$
- $A_{e\pi} = \frac{\left(\frac{25+25}{2}\right)^2 + \left(\frac{25+25}{2}\right)^2}{50^2} = 0.5$
- $\pi = \frac{0.8-0.5}{1-0.5} = 0.6$
- $A_{e\kappa} = \frac{\left(\frac{25\times25}{50}\right) + \left(\frac{25\times25}{50}\right)}{50} = 0.5$
- $\kappa = \frac{0.8-0.5}{1-0.5} = 0.6$
- $A_o = 0.68$
- $A_{e\pi} = \frac{\left(\frac{38+32}{2}\right)^2 + \left(\frac{32+38}{2}\right)^2}{70^2} = 0.5$
- $\pi = \frac{0.68-0.5}{1-0.5} = 0.36$
Comparing $\pi$ and $\kappa$

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>20</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>50</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24</td>
<td>8</td>
<td>32</td>
</tr>
<tr>
<td>No</td>
<td>14</td>
<td>24</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>32</td>
<td>70</td>
</tr>
</tbody>
</table>

\[ A_o = 0.8 \]
\[ A_{\pi}^e = \frac{\left(\frac{25+25}{2}\right)^2 + \left(\frac{25+25}{2}\right)^2}{50^2} = 0.5 \]
\[ \pi = \frac{0.8 - 0.5}{1 - 0.5} = 0.6 \]
\[ A_{\kappa}^e = \frac{\left(\frac{25 \times 25}{50}\right) + \left(\frac{25 \times 25}{50}\right)}{50} = 0.5 \]
\[ \kappa = \frac{0.8 - 0.5}{1 - 0.5} = 0.6 \]

\[ A_o = 0.68 \]
\[ A_{\pi}^e = \frac{\left(\frac{38+32}{2}\right)^2 + \left(\frac{32+38}{2}\right)^2}{70^2} = 0.5 \]
\[ \pi = \frac{0.68 - 0.5}{1 - 0.5} = 0.36 \]
\[ A_{\kappa}^e = \frac{\left(\frac{38 \times 32}{70}\right) + \left(\frac{32 \times 38}{70}\right)}{70} = 0.49 \]
\[ \kappa = \frac{0.68 - 0.49}{1 - 0.49} = 0.37 \]
$S, \pi$ and $\kappa$

For any sample:

$$\begin{align*}
A_e^{\pi} & \geq A_e^S & \pi & \leq S \\
A_e^{\pi} & \geq A_e^{\kappa} & \pi & \leq \kappa
\end{align*}$$

What is a "good" $\kappa$ (or $\pi$ or $S$)?
Scales for the interpretation of Kappa

- **Landis and Koch, 1977**

  0.0 slight  |  0.2 fair    |  0.4 moderate  |  0.6 substantial  |  0.8 perfect

- **Krippendorff, 1980**

  0.67 discard  |  0.8 tentative  |  1.0 good

- **Green, 1997**

  0.0 low  |  0.4 fair / good  |  0.75 high

  “if a threshold needs to be set, 0.8 is a good value”

[Artstein and Poesio, 2008]
Scales for the interpretation of Kappa

- **Landis and Koch, 1977**
  - 0.0: slight
  - 0.2: fair
  - 0.4: moderate
  - 0.6: substantial
  - 0.8: perfect

- **Krippendorff, 1980**
  - 0.67: discard
  - 0.8: tentative
  - 1.0: good

- **Green, 1997**
  - 0.0: low
  - 0.4: fair / good
  - 0.75: high

- “if a threshold needs to be set, 0.8 is a good value”
  [Artstein and Poesio, 2008]
Scales for the interpretation of Kappa

- **Landis and Koch, 1977**
  - 0.0: slight
  - 0.2: fair
  - 0.4: moderate
  - 0.6: substantial
  - 0.8: perfect

- **Krippendorff, 1980**
  - 0.67: discard
  - 0.8: tentative
  - 1.0: good

- **Green, 1997**
  - 0.0: low
  - 0.4: fair / good
  - 0.75: high

- "if a threshold needs to be set, 0.8 is a good value" [Artstein and Poesio, 2008]
Scales for the interpretation of Kappa

- Landis and Koch, 1977
  - 0.0: slight
  - 0.2: fair
  - 0.4: moderate
  - 0.6: substantial
  - 0.8: perfect

- Krippendorff, 1980
  - 0.67: discard
  - 0.8: tentative
  - 1.0: good

- Green, 1997
  - 0.0: low
  - 0.4: fair / good
  - 0.75: high

- “if a threshold needs to be set, 0.8 is a good value” [Artstein and Poesio, 2008]
More Annotators?

Differences among coders are diluted when more coders are used.

- With many coders, difference between \( \pi \) and \( \kappa \) is small
- Another argument for using many coders
More than two annotators

Agreement is the proportion of agreeing pairs

<table>
<thead>
<tr>
<th>Item</th>
<th>Annot1</th>
<th>Annot2</th>
<th>Annot3</th>
<th>Annot4</th>
<th>Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Boxcar</td>
<td>Tanker</td>
<td>Boxcar</td>
<td>Tanker</td>
<td>2/6</td>
</tr>
<tr>
<td>b</td>
<td>Tanker</td>
<td>Boxcar</td>
<td>Boxcar</td>
<td>Boxcar</td>
<td>3/6</td>
</tr>
<tr>
<td>c</td>
<td>Boxcar</td>
<td>Boxcar</td>
<td>Boxcar</td>
<td>Boxcar</td>
<td>6/6</td>
</tr>
<tr>
<td>d</td>
<td>Tanker</td>
<td>Engine2</td>
<td>Boxcar</td>
<td>Tanker</td>
<td>1/6</td>
</tr>
<tr>
<td>e</td>
<td>Engine2</td>
<td>Tanker</td>
<td>Boxcar</td>
<td>Engine1</td>
<td>0/6</td>
</tr>
<tr>
<td>f</td>
<td>Tanker</td>
<td>Tanker</td>
<td>Tanker</td>
<td>Tanker</td>
<td>6/6</td>
</tr>
<tr>
<td>g</td>
<td>Engine1</td>
<td>Engine1</td>
<td>Engine1</td>
<td>Engine1</td>
<td>6/6</td>
</tr>
</tbody>
</table>

When 3 of 4 coders agree, only 3 of 6 pairs agree...
Beware!

\( K \) is a generalization of \( \pi \) (not \( \kappa \)!)\)

Expected agreement

The probability of agreement for an arbitrary pair of coders.

Total number of judgments: \( N \)
Probability of arbitrary annotator picking a particular category \( q_a \): \( \frac{n_{q_a}}{N} \)
Probability of two annotators picking a particular category \( q_a \): \( \left( \frac{n_{q_a}}{N} \right)^2 \)

Probability of two arbitrary annotators picking the same category:

\[
A_e^\pi = \sum_q \left( \frac{n_q}{N} \right)^2 = \frac{1}{N^2} \sum_q n_q^2
\]
Missing Points and Reflexions

I did not introduced the weighted coefficients, in particular $\alpha$ [Krippendorff, 2004]. If you are interested, have a look at [Artstein and Poesio, 2008].

There are ongoing reflexions on some issues, like:

- prevalence
- finding the “right“ negative case (we’ll see that in practical course)
- Precision, recall, F-measure
- Accuracy
- Observed agreement
- $S$, $\kappa$, $\pi$
- More than 2 annotators
Read carefully: [Hripcsak and Rothschild, 2005]
(http://ukPMC.ac.uk/articles/PMC1090460)

Apply the grid we saw in the second course to this article.


In *Proc. of the Fourth ACL Linguistic Annotation Workshop, Uppsala, Suède.*


Sense tagging: does it make sense?

In *Corpus Linguistics Conference*, Lancaster, Angleterre.