The good, the bad and the metrics: understanding inter-annotator agreement

Karën Fort and Yann Mathet

karen.fort@paris-sorbonne.fr / yann.mathet@unicaen.fr

June 26th, 2018
Some sources of inspiration

- The reference articles:
  - *Inter-Coder Agreement for Computational Linguistics* [Artstein and Poesio, 2008]
  - *On the reliability of unitizing textual continua: Further developments* [Krippendorff et al., 2016]
  - *The Unified and Holistic Method Gamma for Inter-Annotator Agreement Measure and Alignment, and Gama-Cat* [Mathet et al., 2015], [Mathet, 2017]
  - Massimo Poesio’s presentation at LREC on the subject (with his permission)
  - Gemma Boleda and Stefan Evert’s course at ESSLLI 2009 (with their permission)
  - *An article about some mistakes and misconceptions regarding agreement measures* [Mathet and Widlöcher, 2016]
Evaluating the quality of manual annotation
Back to basics: what is annotating?
Agreement vs validity
Metrics of|with reference

Agreement coefficients/metrics

Beyond categorization: unitizing

Giving meaning to the metrics

Annotating: back on chance
Evaluating the quality of manual annotation
Back to basics: what is annotating?
Agreement vs validity
Metrics of with reference

Agreement coefficients/metrics

Beyond categorization: unitizing

Giving meaning to the metrics

Annotating: back on chance
Annotating

CATEGORIZATION

SEGMENTATION

Source signal
Annotating

Adding interpretative information [Leech, 1997, Habert, 2005]
The science of categorization
Measuring vs quantifying

Some realities are immediately measurable:

- the height of Mount Everest (8,848 m)

others are not:

- the number of unemployed persons → what is an unemployed person?

"Mais précisément la définition et la mesure de la population active et du chômage relèvent d’une autre épistémologie que celle de l’étoile polaire. Elles impliquent des conventions (analogues aux principes généraux des lois et des codes votés par les Parlements) et des décisions (analogues à celles d’un juge) d’affecter tel cas à telle classe.” [Desrosières, 2001]
"Un codage est une décision conventionnelle de construire une classe d’équivalence entre divers objets, la ’classe’ étant jugée plus ’générale’ que tout objet singulier. La 1ère condition pour cela est de supposer que tous ces objets peuvent être comparés, ce qui ne va pas de soi”
[Desrosières, 1989]

Quantifying, is to agree, then to measure [Desrosières, 2008]
How to ”agree”?

annotation guidelines

and how to check that the consensus is understood and applied?

inter-annotator agreement
Evaluating the quality of manual annotation

Back to basics: what is annotating?

Agreement vs validity

Metrics of with reference

Agreement coefficients/metrics

Beyond categorization: unitizing

Giving meaning to the metrics

Annotating: back on chance
Four main concepts interacting

- Multi-annotations
- Agreement (and agreement measures)
- Reference (Gold-Standard)
- Validity (and validity measures)
From agreement to a reference

High agreement

Low Agreement

no possible reference

1

2

3

getting a reference

STRATEGY

12 / 99
Agreement vs Validity

The center of the target is a metaphor of the perfect annotation (validity)
Agreement vs Validity

Benefiting from reproducibility

The whole corpus

a part of the corpus

Three annotators named 1, 2 and 3: 1 2 3
Benifiting from reproducibility
Agreement vs Validity

Benefiting from reproducibility
Agreement vs Validity

Benefiting from reproducibility

high agreement reproducibility reproducibility reproducibility validity validity validity validity
Evaluating the quality of manual annotation

Back to basics: what is annotating?
Agreement vs validity
Metrics of with reference

Agreement coefficients/metrics

Beyond categorization: unitizing

Giving meaning to the metrics

Annotating: back on chance
Particular case: with a *gold-standard*

In some cases (rare and often artificial), a “gold-standard” pre-exists:
the corpus has been annotated, at least partly, and this annotation is considered as being “perfect”, a reference [Fort and Sagot, 2010].

In those cases, another *complementary* metric 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 \((\beta = 0.5)\)
“Gold-standard”?

- it is very rare that a reference already exists
- can it be “perfect”? [Fort and Sagot, 2010]
- is it advisable to use the F-measure in other cases? See [Hripcsak and Rothschild, 2005]
- back to inter-annotator agreement coefficients
Evaluating the quality of manual annotation

**Agreement coefficients/metrics**
- Observed agreement
- Correction needed
- (Some) Coefficients

**Beyond categorization: unitizing**

**Giving meaning to the metrics**

**Annotating: back on chance**
Evaluating the quality of manual annotation

**Agreement coefficients/metrics**

- Observed agreement
- Correction needed
- (Some) Coefficients

**Beyond categorization: unitizing**

**Giving meaning to the metrics**

**Annotating: back on chance**
Example

Validation of semantic annotations (container/contained):

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

→ **Inter-annotator agreement?**
Synthetic representation: confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>✗</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 annotations on which the annotators agree

Here:

\[
\frac{30}{99}
\]
Synthetic representation : confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>✓</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>✗</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

Observed agreement \((A_o)\)

proportion of annotations on which the annotators agree

Here : \(A_o = \frac{4 + 2}{10} = 0.6\)
Evaluating the quality of manual annotation

Agreement coefficients/metrics
  Observed agreement
  Correction needed
  (Some) CoefficientS

Beyond categorization: unitizing

Giving meaning to the metrics

Annotating: back on chance
What if…

Exercise

- each unit has to be annotated
- 2 categories 📚 and 🌟
- 3 annotators: $A_1$, $A_2$ et $A_3$

Enumerate the different annotation possibilities (on one unit)?
Correction and follow up

In this particular case, it is impossible to obtain no agreement at all:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>?</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>( A_1 )</th>
<th>( A_2 )</th>
<th>( A_3 )</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>3</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>?</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>?</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>?</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>?</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>?</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>?</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>?</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>✅</td>
<td>✅</td>
<td>✅</td>
<td>3</td>
</tr>
<tr>
<td>✅</td>
<td>✅</td>
<td>☹</td>
<td>1</td>
</tr>
<tr>
<td>☹</td>
<td>✅</td>
<td>☹</td>
<td>?</td>
</tr>
<tr>
<td>☹</td>
<td>☹</td>
<td>☹</td>
<td>?</td>
</tr>
<tr>
<td>☹</td>
<td>☹</td>
<td>☹</td>
<td>?</td>
</tr>
<tr>
<td>☹</td>
<td>☹</td>
<td>☹</td>
<td>?</td>
</tr>
<tr>
<td>☹</td>
<td>☹</td>
<td>☹</td>
<td>?</td>
</tr>
<tr>
<td>☹</td>
<td>☹</td>
<td>☹</td>
<td>?</td>
</tr>
</tbody>
</table>


Correction and follow up

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>3</td>
</tr>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>1</td>
</tr>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>1</td>
</tr>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>?</td>
</tr>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>?</td>
</tr>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>?</td>
</tr>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>?</td>
</tr>
<tr>
<td>✧</td>
<td>✧</td>
<td>✧</td>
<td>?</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>A₁</th>
<th>A₂</th>
<th>A₃</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>✠</td>
<td>✠</td>
<td>✠</td>
<td>3</td>
</tr>
<tr>
<td>✠</td>
<td>✠</td>
<td>☢</td>
<td>1</td>
</tr>
<tr>
<td>☢</td>
<td>✠</td>
<td>✠</td>
<td>1</td>
</tr>
<tr>
<td>✠</td>
<td>☢</td>
<td>☢</td>
<td>3</td>
</tr>
<tr>
<td>☢</td>
<td>☢</td>
<td>✠</td>
<td>?</td>
</tr>
<tr>
<td>☢</td>
<td>✠</td>
<td>☢</td>
<td>?</td>
</tr>
<tr>
<td>☢</td>
<td>☢</td>
<td>✠</td>
<td>?</td>
</tr>
<tr>
<td>☢</td>
<td>✠</td>
<td>☢</td>
<td>?</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="#" alt="Image 1" /></td>
<td><img src="#" alt="Image 2" /></td>
<td><img src="#" alt="Image 3" /></td>
<td>3</td>
</tr>
<tr>
<td><img src="#" alt="Image 4" /></td>
<td><img src="#" alt="Image 5" /></td>
<td><img src="#" alt="Image 6" /></td>
<td>1</td>
</tr>
<tr>
<td><img src="#" alt="Image 7" /></td>
<td><img src="#" alt="Image 8" /></td>
<td><img src="#" alt="Image 9" /></td>
<td>1</td>
</tr>
<tr>
<td><img src="#" alt="Image 10" /></td>
<td><img src="#" alt="Image 11" /></td>
<td><img src="#" alt="Image 12" /></td>
<td>3</td>
</tr>
<tr>
<td><img src="#" alt="Image 13" /></td>
<td><img src="#" alt="Image 14" /></td>
<td><img src="#" alt="Image 15" /></td>
<td>1</td>
</tr>
<tr>
<td><img src="#" alt="Image 16" /></td>
<td><img src="#" alt="Image 17" /></td>
<td><img src="#" alt="Image 18" /></td>
<td>?</td>
</tr>
<tr>
<td><img src="#" alt="Image 19" /></td>
<td><img src="#" alt="Image 20" /></td>
<td><img src="#" alt="Image 21" /></td>
<td>?</td>
</tr>
<tr>
<td><img src="#" alt="Image 22" /></td>
<td><img src="#" alt="Image 23" /></td>
<td><img src="#" alt="Image 24" /></td>
<td>?</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>3</td>
</tr>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>1</td>
</tr>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>1</td>
</tr>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>3</td>
</tr>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>1</td>
</tr>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>1</td>
</tr>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>?</td>
</tr>
<tr>
<td>✈✈</td>
<td>✈✈</td>
<td>✈✈</td>
<td>?</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>🟡🟡</td>
<td>🟡🟡</td>
<td>🟡🟡</td>
<td>3</td>
</tr>
<tr>
<td>🟡🟡</td>
<td>🟡🔵</td>
<td>🟡🟡</td>
<td>1</td>
</tr>
<tr>
<td>🟡🔵</td>
<td>🟡🟡</td>
<td>🟡🔵</td>
<td>1</td>
</tr>
<tr>
<td>🟡🟡</td>
<td>🟡🔵</td>
<td>🟡🟡</td>
<td>3</td>
</tr>
<tr>
<td>🟡🔵</td>
<td>🟡🔵</td>
<td>🟡🔵</td>
<td>1</td>
</tr>
<tr>
<td>🟡🔵</td>
<td>🟡🔵</td>
<td>🟡🔵</td>
<td>1</td>
</tr>
<tr>
<td>🟡🔵</td>
<td>🟡🔵</td>
<td>🟡🔵</td>
<td>?</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>3</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>1</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>1</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>3</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>1</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>1</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>1</td>
</tr>
<tr>
<td>✖️</td>
<td>✖️</td>
<td>✖️</td>
<td>1</td>
</tr>
</tbody>
</table>
Correction and follow up

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>$A_2$</th>
<th>$A_3$</th>
<th>Agreeing pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>3</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>❌</td>
<td>1</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>❌</td>
<td>1</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>✖️</td>
<td>3</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>1</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>1</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>1</td>
</tr>
<tr>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
<td>1</td>
</tr>
</tbody>
</table>

In the worst case scenario, we get $8 \times 1 / (8 \times 3) = 0.333$
Exercise (follow up)

- each unit has to be annotated
- 2 categories
- 3 annotators

Enumerate the different annotation possibilities (on one unit)?
Agreement scale

The inter-annotator agreement will not be computed according to the same scale depending on the cases:

- **Case #1**: 3 annotators and 2 categories
  
  ![Scale 0.33](0.33 to 1)

- **Case #2**: 2 annotators and 2 categories
  
  ![Scale 0](0 to 1)
Agreement scale

The inter-annotator agreement will not be computed according to the same scale depending on the cases:

- Case #1: 3 annotators and 2 categories
  
  \[
  0.33 \quad 1
  \]

- Case #2: 2 annotators and 2 categories
  
  \[
  0 \quad 1
  \]

→ need for some **correction** of the observed results to be able to interpret the results
What if... part of the agreement was due to chance? In our example, which proportion can be due to chance?
What if... part of the agreement was due to chance?

- two annotators who annotate randomly would agree in half of the cases (0.5).
- the part of agreement which is due to chance varies according to the annotation scheme and the data to annotate

The meaningful part of the agreement is that above chance → similar to the concept of baseline
Taking chance into account

Expected agreement \( A_e \)
expected value of the observed agreement

Agreement above chance : \( A_o - A_e \)
Maximum possible agreement above chance : \( 1 - A_e \)

Proportion of agreement above chance reached : \( \frac{A_o - A_e}{1 - A_e} \)

Prefect agreement : \( \frac{1 - A_e}{1 - A_e} \)
Perfect disagreement : \( \frac{-A_e}{1 - A_e} \)
A change in reference

0

Ae

Ao

A

1

raw values

0

1

corrected values

50 / 99
Is this fair?

- Sometimes considered as unfair: chance correction would be considering annotators are playing dice...
- Wrong argument: the part of chance does not come from "playing dice", but from disagreement
- Disagreement $\implies$ annotation is not 100% under control $\implies$ chance may occur
- The more the observed agreement, the lesser the chance correction:

$$A_{deducted} = A_o - A$$

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

$$= \frac{A_e}{1 - A_e} \cdot (1 - A_o)$$

(1)
Expected agreement

How to compute the amount of expected agreement due to chance \( (A_e) \)?
Evaluating the quality of manual annotation

**Agreement coefficients/metrics**
- Observed agreement
- Correction needed
- *(Some) CoefficientS*

**Beyond categorization: unitizing**

**Giving meaning to the metrics**

**Annotating: back on chance**
S [Bennett et al., 1954]

S

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 $q_a : (\frac{1}{q})^2$

Probability of both annotators picking the same category:

$$A_e^S = q.(\frac{1}{q})^2 = \frac{1}{q}$$
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>

\[
A_o = \frac{20+20}{50} = 0.8
\]

\[
A_e^S = \frac{1}{2} = 0.5
\]

\[
S = \frac{0.8-0.5}{1-0.5} = 0.6
\]
All the categories are equally likely: consequences

\[
\begin{array}{c|ccc|c}
& \text{Yes} & \text{No} & \text{Total} \\
\hline
\text{Yes} & 20 & 5 & 25 \\
\text{No} & 5 & 20 & 25 \\
\text{Total} & 25 & 25 & 50 \\
\end{array}
\]

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

\[
\begin{array}{c|cccc|c}
& \text{Yes} & \text{No} & \text{C} & \text{D} & \text{Total} \\
\hline
\text{Yes} & 20 & 5 & 0 & 0 & 25 \\
\text{No} & 5 & 20 & 0 & 0 & 25 \\
\text{C} & 0 & 0 & 0 & 0 & 0 \\
\text{D} & 0 & 0 & 0 & 0 & 0 \\
\text{Total} & 25 & 25 & 0 & 0 & 50 \\
\end{array}
\]
All the categories are equally likely: consequences

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

$A_0 = 0.8$

$S = 0.6$

<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>

$A_0 = 0.8$

$S = 0.73$
### 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>

$A_o = 0.8$

$S = 0.6$

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

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

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>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>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>

$$A_0 = 0.8$$

$$S = 0.6$$

$$A_\pi^e = \frac{\left(\frac{25+25}{2}\right)^2 + \left(\frac{25+25}{2}\right)^2}{50^2} = 0.5$$

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

<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>

$$A_0 = 0.8$$

$$S = 0.73$$

$$A_\pi^e = \frac{\left(\frac{25+25}{2}\right)^2 + \left(\frac{25+25}{2}\right)^2}{50^2} = 0.5$$

$$\pi = \frac{0.8-0.5}{1-0.5} = 0.6$$
\[ \kappa \quad [\text{Cohen, 1960}] \]

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>

$A_o = 0.8$

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

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

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>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>

$A_o = 0.8$

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

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

$A_o = 0.8$

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

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

$A_e^\kappa = \frac{\left(\frac{25\times25}{50}\right) + \left(\frac{25\times25}{50}\right)}{50} = 0.5$

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

$$A_o = 0.8$$
$$A_{\pi} = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$$
$$\pi = \frac{0.8 - 0.5}{1 - 0.5} = 0.6$$
$$A_{\kappa} = \frac{\frac{25 \times 25}{50} + \frac{25 \times 25}{50}}{50} = 0.5$$
$$\kappa = \frac{0.8 - 0.5}{1 - 0.5} = 0.6$$

<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>

$$A_o = 0.8$$
$$A_{\pi} = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$$
$$\pi = \frac{0.8 - 0.5}{1 - 0.5} = 0.6$$
$$A_{\kappa} = \frac{\frac{25 \times 25}{50} + \frac{25 \times 25}{50}}{50} = 0.5$$
$$\kappa = \frac{0.8 - 0.5}{1 - 0.5} = 0.6$$
## 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>

<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>
<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>

$$A_o = 0.8$$

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

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

$$A_{e}^{\kappa} = \frac{\left(\frac{25\times25}{50}\right) + \left(\frac{25\times25}{50}\right)}{50} = 0.5$$

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

$$A_o = 0.68$$

$$A_{e}^{\pi} = \frac{((\frac{38+32}{2})^2 + (\frac{32+38}{2})^2)}{70^2} = 0.5$$

$$\pi = \frac{0.68-0.5}{1-0.5} = 0.36$$
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>

$A_o = 0.8$

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

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

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

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

$A_o = 0.68$

$A^\pi_e = \frac{((\frac{38+32}{2})^2 + (\frac{32+38}{2})^2)}{70^2} = 0.5$

$\pi = \frac{0.68-0.5}{1-0.5} = 0.36$

$A^\kappa_e = \frac{(\frac{38\times32}{70}) + (\frac{32\times38}{70})}{70} = 0.49$

$\kappa = \frac{0.68-0.49}{1-0.49} = 0.37$
For any sample:

\[ A_e^{\pi} \geq A_e^{S} \quad \pi \leq S \]

\[ A_e^{\pi} \geq A_e^{\kappa} \quad \pi \leq \kappa \]

What is a "good" \( \kappa \) (or \( \pi \) or \( S \))?
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization: unitizing

Introducing unitizing

Three different strategies to assess agreement on unitizing
Further comparisons between strategies
Introducing the gamma coefficients

Giving meaning to the metrics

Annotating: back on chance
Unitizing

- **Unitizing** = annotators freely put units of various sizes and categories on a continuum
- Until recently, only the Krippendorff’s $\alpha$ were designed for unitizing
- A challenge: **how to compare annotations from different annotators?**
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization: unitizing
  Introducing unitizing
  Three different strategies to assess agreement on unitizing
  Further comparisons between strategies
  Introducing the gamma coefficients

Giving meaning to the metrics

Annotating: back on chance
First strategy : Atomizing

A workaround based on categorizing measures : atomizing the continuum

Leads to severe biases :

- two contiguous units seen as one in (1)
- a slight shift in (3) is as important as a false positive in (2)
- agreement on blanks (what is not annotated)
- unit overlapping is not possible

Still seen when reviewing for ACL 2017 and 2018 (A* conference in NLP)
Second strategy: Relying on intersections

- Krippendorff’s $\alpha$ coefficients go further
- They rely on intersecting parts of units rather than on atoms
- Biases are more limited, but still present (as shown by the CST)
- I discovered a paradox in $\upsilon\alpha$: positional agreement may increase overall disagreement

\[ 2^2 + 6^2 + 2^2 < 10^2 \]
The third strategy (by Gamma) : Relying on an alignment

- Gamma proposal : to rely on an alignment which provides associations between units of different annotators
- Each unit of an annotator is associated with at most one unit of each of the others
- short units are as important as long ones
- Overlapping is natively accepted
Alignment

A **holistic** mechanism

annotator A

annotator B
Alignment

A holistic mechanism

annotator A

annotator B
Alignment

A holistic mechanism
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization : unitizing
  Introducing unitizing
  Three different strategies to assess agreement on unitizing
  Further comparisons between strategies
  Introducing the gamma coefficients

Giving meaning to the metrics

Annotating : back on chance
Length of units

A potential bias (depending on your needs...) inherent to the intersection strategies

- In CL and NLP, units of different sizes have the same importance

  Barack Hussein Obama II is the 44th and current (...). In 2004, Obama received national (...)

- This cannot be done with atomizing nor intersecting parts

- At word level, "Barak Hussein Obama II" is 4 times as long as "Obama"
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization: unitizing
  Introducing unitizing
  Three different strategies to assess agreement on unitizing
  Further comparisons between strategies
  Introducing the gamma coefficients

Giving meaning to the metrics

Annotating: back on chance
Alignments vs intersections

Units with no intersection may refer to the same intent

- Microsoft Corporation (MSFT) ➔ 1
- Microsoft Corporation (MSFT) ➔ 2
- Microsoft Corporation (MSFT) ➔ 3
The agreement metric $\gamma$

$\gamma$ in a nutshell (Mathet, Widlöcher, Métivier, 2015) with the support of S. Bouvry

1. A chance corrected agreement measure based on dissimilarity and alignments. It introduces:
   ▶ dissimilarity between two units depending on their positions and categories
   ▶ unitary alignment grouping at most one unit of each annotator (and using "empty units")
   ▶ alignment as a set of unitary alignments covering all units of all annotators (a.k.a. a partition)
   ▶ disorder of a unitary alignment resulting from dissimilarities between its units
   ▶ disorder associated with an alignment depending on its unitary alignments

2. $\gamma$ chooses the alignment having the minimal disorder

3. It is then chance corrected, by also computing the average disorder of random sets of annotations
The agreement metric $\gamma$

Example of 3 possible alignments:

- 3 unitary alignments found by Gamma
- 3 empty units found by Gamma
Chance correction of $\gamma$

Re-sampling annotations to get the part of "chance" of $\gamma$

1. Re-sampling at local level

```
position 15
position 24
position 38
continuum length = 44
real annotations (by one annotator)
```

```
examples of transformed annotations (number 15, 24 and 38)
by circular shifts of real annotations
```

2. Re-sampling at corpus level

```
random annotations from 2-2, 4-3, 7-1
```
Focusing on categorization with $\gamma_{cat}$

$\gamma_{cat}$ : a complement to $\gamma$ focused on categories

- $\gamma$ is an overall measure for all discrepancies (position, category, false positive, split)
- When it is not as high as wished, it is useful to get more details about the disagreements
- $\gamma_{cat}$ is focused on categories, leaving aside all other kinds of disagreement
- It somehow corresponds to what could be the agreement if annotators had not had to unitize
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization: unitizing

Giving meaning to the metrics
   Interpretations
   Semantics

Annotating: back on chance
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization: unitizing

Giving meaning to the metrics
  Interpretations
  Semantics

Annotating: back on chance
Scales for the interpretation of Kappa

- **Landis and Koch, 1977**

  - 0.0
  - 0.2
  - 0.4
  - 0.6
  - 0.8
  - 1.0

  slight | fair | moderate | substantial | perfect

- **Krippendorff, 1980**

  - 0.67
  - 0.8
  - 1.0

  discard | tentative | good

- **Green, 1997**

  - 0.0
  - 0.4
  - 0.75
  - 1.0

  low  | fair / good | 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
  
  \[
  \begin{array}{cccccc}
  0.0 & 0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
  \text{slight} & \text{fair} & \text{moderate} & \text{substantial} & \text{perfect} \\
  \end{array}
  \]

- Krippendorff, 1980
  
  \[
  \begin{array}{cccccc}
  0.0 & 0.67 & 0.8 & 1.0 \\
  \text{discard} & \text{tentative} & \text{good} \\
  \end{array}
  \]

“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
  - 1.0

"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]
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization: unitizing

Giving meaning to the metrics
  Interpretations
  Semantics

Annotating: back on chance
Giving meaning to the obtained results [COLING 2012a]

Creation of a ”Richter” tool (CorpusShufflingTool) that:

- takes as input a (real or artificial) reference annotation
- generates alterations of a certain (controlled) magnitude (between 0 and 1)
- applies one or more inter-annotator agreement metrics on each set of annotations (corresponding to a alteration magnitude)
”Richter” tool on TCOF-POS [Benzitoun et al., 2012]

No prevalence, but proximity between categories taken into account:

![Graph showing agreement and magnitude](image)

- **Cohen’s Kappa**
- **Weighted Kappa**
Evaluating the quality of manual annotation

Agreement coefficients/metrics

Beyond categorization : unitizing

Giving meaning to the metrics

Annotating : back on chance
  Annotators and biases
  Experts, but of what?
Bias

Well-trained annotators are less sensitive to biases:
  ▶ of pre-annotation [Fort and Sagot, 2010]
  ▶ of the annotation tool [Dandapat et al., 2009]
and annotate less ”randomly”

Using annotation guidelines allows to obtain better annotations
[Nédellec et al., 2006]
Experts:

- of the domain: annotation in microbiology (gene renaming), football, etc.
- of the task: annotation in structured named entities

... contradictions and shortfalls:

→ to annotate structured named entities in old press, do we need names entities specialists or historians?
Appendices

Examples

Bibliography
Annodis

http://redac.univ-tlse2.fr/corpus/annodis/

Selon nous, la lutte contre le terrorisme ne serait pas similaire aux évolutions entraînées par les autres conflits. Elle débouche en fait sur un activismisme tous-azimut, qui concerne aussi bien l’Etat fédéral que les États fédérés et les autorités locales (villes, comtés). Plutôt que de parler de centralisation, il faudrait évoquer un renforcement des fonctions légittimes de chacun des niveaux du gouvernement. L’effet de ces derniers est la protection des citoyens. Pour le niveau fédéral, les autorités locales, elles, gèrent les moyens de réponse immédiats aux agressions terroristes (police, pompier, santé). L’essentiel des problèmes suscités par la protection du territoire contre le terrorisme réside dans la coordination entre les différents organes. L’administration actuelle s’engage résolument dans cette voie, et entame une réorganisation massive des administrations nationales.
Annodis Agreement Values

\[ \gamma = 0.55 \quad (0.53 \leq \gamma \leq 0.57) \]
observed disagreement = 0.838
expected disagreement = 1.864 ± 5%
number of resulting alignments = 195

\[ \gamma_{cat} = 0.74 \quad (0.73 \leq \gamma_{cat} \leq 0.75) \]
observed disagreement = 0.090
expected disagreement = 0.347 ± 5%
number of resulting alignments = 195.0
\[ \gamma_{k(amo\text{rc})} = 1.000 \]
\[ \gamma_{k(cloture)} = 1.000 \]
\[ \gamma_{k(enumera Theme)} = 0.936 \]
\[ \gamma_{k(indice)} = 0.636 \]
\[ \gamma_{k(item)} = 1.000 \]
\[ \gamma_{k(segment)} = 1.000 \]
\[ \gamma_{k(todo)} = -0.082 \]
Football Matches Transcriptions

K. Fort & V. Claveau, TALN 2012

Fabien Lévêque : N’Diaye en retrait vers Revel qui allonge le jeu directement.

Ekobo et Sammaritano de la tête maintenant, c’est bien fait ! Quintin.

Xavier Gravelaine : Allez il va falloir y aller. Quintin tout seul. Même si Traoré avait bien fermé son couloir.

Fabien Lévêque : Sammaritano qui perd son ballon. Chamakh. Il y aura une faute siflée par Freddy Fautrel, faute de Franck Jurietti, tout va bien. Oui Daniel?

Daniel Lauclair : Oui, je suis en compagnie d’un homme qui est frustré et qui a été privé de finale pour deuxième carton jaune cet été à Auxerre, c’est il est un peu triste, il est dans les tribunes, il est à nos côtés Alou Diarra. Alou, c’est la meilleure des entames pour Bordeaux qui mène après deux minutes de jeu.

Alou Diarra : Ça a débuté fort. C’est quand même fort d’avoir marqué dans les premières minutes. Faut continuer dans ce sens là, et mettre un deuxième but, pas s’arrêter après ce but, parce qu’ils sont capables de revenir. Ils sont dangereux on le sait. Eh
Football Agreement Values

\[ \gamma = 0.72 \ (0.70 \leq \gamma \leq 0.73) \]
observed disagreement = 0.554
expected disagreement = 1.951 ± 5%
number of resulting alignments = 355

\[ \gamma_{cat} \approx 0.93 \ (0.93 \leq \gamma_{cat} \leq 0.93) \]
observed disagreement = 0.056
expected disagreement = 0.805 ± 5%
number of resulting alignments = 355

\[ Y_k(A_{ActionDuPublic}) = NA \]
\[ Y_k(A_{ArreterBut}) = 1.000 \]
\[ Y_k(A_{Centrer}) = 0.000 \]
\[ Y_k(A_{FaireFauteDeJeu}) = 0.000 \]
\[ Y_k(A_{FaireTentative2Centre}) = 0.000 \]
\[ Y_k(A_{HorsJeu}) = 0.667 \]
\[ Y_k(A_{InterceptorBallon}) = 0.958 \]
\[ Y_k(A_{MarquerBut}) = 1.000 \]
\[ Y_k(A_{PossederBallon}) = -0.068 \]
\[ Y_k(A_{PrendreCartonJaune}) = 1.000 \]
\[ Y_k(A_{PrendreRappelALOrdre}) = 0.667 \]
\[ Y_k(A_{RaterBut}) = 0.857 \]
\[ Y_k(A_{TirerCorner}) = 1.000 \]
\[ Y_k(A_{TirerCoupFrancDirect}) = 0.833 \]
\[ Y_k(\text{ActionPourActeurVide}) = 0.633 \]
Smell Recognition Dialog Transcriptions

C2O : Ongoing project in Psychology, Caen, leaded by Maryse Delaunay-El Allam
\[ \gamma = 0.39 \ (0.36 \leq \gamma \leq 0.42) \]
observed disagreement = 1.666
expected disagreement = 2.747 ± 5%
number of resulting alignments = 130

\[ \gamma_{\text{cat}} = 0.71 \ (0.70 \leq \gamma_{\text{cat}} \leq 0.73) \]
observed disagreement = 0.248
expected disagreement = 0.860 ± 5%
number of resulting alignments = 130.0

\[ \gamma_k(\text{ActionSurEmotions}) = \text{NA} \]
\[ \gamma_k(\text{ActivitésCulinaires}) = 0.397 \]
\[ \gamma_k(\text{ActivitésHygièneCorporelle}) = 0.896 \]
\[ \gamma_k(\text{ActivitésHygièneDomestique}) = 0.281 \]
\[ \gamma_k(\text{ActivitésJardinage}) = 0.721 \]
\[ \gamma_k(\text{ActivitésOlfactives}) = 0.000 \]
\[ \gamma_k(\text{ActivitésPromenades}) = 0.450 \]
\[ \gamma_k(\text{CapacitésOlfactives}) = 0.863 \]
\[ \gamma_k(\text{CoutumeFestive}) = \text{NA} \]
\[ \gamma_k(\text{EnvironnementOdorant}) = 0.385 \]
\[ \gamma_k(\text{ÉvénementSpirituel}) = \text{NA} \]
\[ \gamma_k(\text{Flairage}) = 0.750 \]
\[ \gamma_k(\text{GesteSymbolique}) = 0.494 \]
\[ \gamma_k(\text{Hédonisme}) = 0.870 \]
\[ \gamma_k(\text{Imitation}) = 0.000 \]
\[ \gamma_k(\text{Initiation-Délétérivités}) = 0.478 \]


Complex linguistic annotation - no easy way out! a case from bangla and hindi POS labeling tasks.
In Proceedings of the third ACL Linguistic Annotation Workshop, Singapour.

Comment faire des choses qui tiennent : histoire sociale et statistique.

Enter réalisme métrologique et conventions d’équivalence : les ambiguïtés de la sociologie quantitative.
Genèses, 2(43) :112–127.

Pour une sociologie historique de la quantification :
L’Argument statistique I.

Fort, K. and Sagot, B. (2010).
Influence of pre-annotation on POS-tagged corpus development.


Mathet, Y. (2017). The agreement measure gamma-cat ($\gamma_{cat}$), a complement to gamma focused on categorization of a continuum. Computational Linguistics, 43(3) :661–681.


The unified and holistic method gamma ($\gamma$) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3) :437–479.

