Evaluating Manual Annotation Quality

Karën Fort

karen.fort@loria.fr / https://members.loria.fr/KFort/

École thématique d'été "Annotations" - June 2nd, 2022

Qual Program

http://annotations-2022.llf-paris.fr/?fichier=programme

- Qual1: done
- Qual2: now
- Qual3: crowdsourcing

Some sources of inspiration

- Reference articles:
  - *Inter-Coder Agreement for Computational Linguistics* [Artstein and Poesio, 2008]
  - *The Unified and Holistic Method Gamma for Inter-Annotator Agreement Measure and Alignment* [Mathet et al., 2015]
- Presentation from Massimo Poesio at LREC on the subject (with his approval)
- Gemma Boleda and Stefan Evert's course on the subject (with their approval) at ESSLLI 2009
- Yann Mathet

Introduction

Fundamental question: **are the annotations correct?**

- systems learn errors from the human annotators (noise $\neq$ bias [Reidsma and Carletta, 2008])
- evaluation can be erroneous
- results from linguistic analyses or symbolic systems may be flawn and inconclusive

Reminder: consensus is at the heart of annotation

"agree to measure" ("convenir pour mesurer") [Desrosières, 2008]

Annotation is about quantifying

Measuring *vs* quantifying [Desrosières, 2008] :

- **measuring**: implies some measurable form (e.g. the height of Mont Blanc)
- **quantifying**: implies establishing preliminary conventions of equivalence

The consensus needs to be equipped:

- annotation guidelines (12 p. for football)
- meetings with the annotators and the campaign manager
- evaluate the consensus (consistency)

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct
- But there is no "ground truth"

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct
- But there is no "ground truth"
  - linguistic categories are determined by human judgments

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct
- But there is no "ground truth"
  - linguistic categories are determined by human judgments
  - consequences: it is impossible to measure directly if a category is correct or not

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct
- But there is no "ground truth"
  - linguistic categories are determined by human judgments
  - consequences: it is impossible to measure directly if a category is correct or not
- we can only measure the **reliability** of the annotation

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct
- But there is no "ground truth"
  - linguistic categories are determined by human judgments
  - consequences: it is impossible to measure directly if a category is correct or not
- we can only measure the **reliability** of the annotation
  - *i.e.* if the human annotators make the same decisions in a **consistent** way $\Rightarrow$ they have internalized the annotation schema

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct
- But there is no "ground truth"
  - linguistic categories are determined by human judgments
  - consequences: it is impossible to measure directly if a category is correct or not
- we can only measure the **reliability** of the annotation
  - *i.e.* if the human annotators make the same decisions in a **consistent** way $\Rightarrow$ they have internalized the annotation schema
  - underlying hypothesis: high reliability implies validity of the annotation

Validity *vs* reliability [Artstein and Poesio, 2008]

- we are interested in the **validity** of manual annotation
  - *i.e.* if the annotated categories are correct
- But there is no "ground truth"
  - linguistic categories are determined by human judgments
  - consequences: it is impossible to measure directly if a category is correct or not
- we can only measure the **reliability** of the annotation
  - *i.e.* if the human annotators make the same decisions in a **consistent** way $\Rightarrow$ they have internalized the annotation schema
  - underlying hypothesis: high reliability implies validity of the annotation
- How to evaluate this reliability?

Measuring the reliability (consistency) of the annotation

- each item is annotated by one annotator, with random checks ($\approx$ second annotation)
- some items are annotated by two or more annotators
- each item is annotated by two or more annotators - followed by a conciliation phase
- each item is annotated by two or more annotators - followed by a final decision finale made by a superannotator (expert)

In all cases, the metric used to measure reliability is an (inter-annotator) agreement coefficient

Specific Case: existing *gold-standard*

In some cases (rare and often artificial), there is a "reference":
le corpus a été annoté, au moins partiellement, et cette annotation est considérée comme “parfaite”, une référence [Fort and Sagot, 2010].

In these cases, another, additionnal metric can be used:

which one?

Specific Case: existing *gold-standard*

In some cases (rare and often artificial), there is a "reference":
le corpus a été annoté, au moins partiellement, et cette annotation est considérée comme “parfaite”, une référence [Fort and Sagot, 2010].

In these cases, another, additionnal 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

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

- Silence:
- Precision:
- Noise:

Precision / Recall: back to basics

- Recall: measures the quantity of found annotations

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

- Silence: *complement* of recall (unfound correct annotations)
- Precision:
- Noise:

Precision / Recall: back to basics

- Recall: measures the quantity of found annotations

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

- Silence: *complement* of recall (unfound correct annotations)
- 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

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

- Silence: *complement* of recall (unfound correct annotations)
- 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 (found incorrect annotations)

F-measure: back to basics (Wikipedia)

Harmonic mean of the precision and recall or balanced F-score

$$F = 2x\frac{\text{precision } x\text{recall}}{\text{precision}+\text{recall}}$$

... or the F1 measure, recall and precision having similar weights.

A specific cas of Fβ measure:

$$F\beta = (1+\beta^2)x\frac{\text{precision } x\text{recall}}{\beta^2 x\text{precision} + \text{rappel}}$$

The value of β allows to favor:

- recall ($\beta = 2$)
- precision ($\beta = 0.5$)

"Gold-standard"?

- rare that a reference already exists
- can it be "perfect"? [Fort and Sagot, 2010]

→ can we use the F-measure in other cases? See [Hripcsak and Rothschild, 2005]

⇒ Back to inter-annotator coefficients

Example

Validation of semantic annotations (content/container):

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

→ **Inter-annotator agreement**?

Synthetic representation

		A		
		✓	✗	Total
B	✓	**4**	2	6
	✗	2	**2**	4
	Total	6	4	**10**

Observed agreement (A_o)

proportion of answers on which the annotators agree.

Here:

Synthetic representation

		A		
		✓	✗	Total
B	✓	**4**	2	6
	✗	2	**2**	4
	Total	6	4	**10**

Observed agreement (A_o)

proportion of answers on which the annotators agree.

Here: $A_o = \frac{4+2}{10} = \mathbf{0.6}$

What if...

... part of the agreement was due to **chance**:
in our example, which agreement proportion can be due to chance?

What if...

... part of the agreement was due to **chance**:

- Two annotators annotating randomly will agree **half of the time** (0.5).
- Chance agreement varies according to the annotation schema and the annotated data.

The significant agreement is what is above chance.
$\rightarrow$ similar to the concept of *baseline*.

What if?

Practice

- each unit must be annotated
- 2 categories  and 
- 3 annotators: A_1, A_2 and A_3

What are the different possibilities of annotating one unit (by the 3 annotators)?

Correction and follow up

In this case, it is impossible to get a null agreement (per pair of annotators):

A_1	A_2	A_3	Nb of agreeing pairs
			?
			?
			?
			?
			?
			?
			?
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			?
			?
			?
			?
			?
			?
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			?
			?
			?
			?
			?
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			1
			?
			?
			?
			?
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			1
			3
			?
			?
			?
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			1
			3
			1
			?
			?
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			1
			3
			1
			1
			?
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			1
			3
			1
			1
			1
			?

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			1
			3
			1
			1
			1
			1

Correction and follow up

A_1	A_2	A_3	Nb of agreeing pairs
			3
			1
			1
			3
			1
			1
			1
			1

In the worse case scenario, we would get $8x1/8x3 = 0.333$

What if?

Practice (follow up)

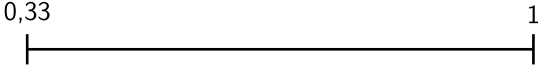
- each unit must be annotated
- 2 categories
- ~~3~~ 2 annotators

What are the different possibilities of annotating one unit?

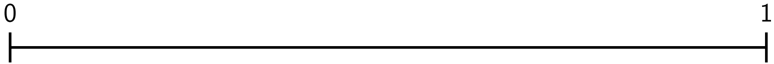
Scales of agreement coefficients

The inter-annotator agreement is not computed on the same scale depending on cases:

- Case 1: 3 annotators and 2 categories



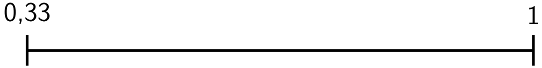
- Case 2: 2 annotators and 2 categories



Scales of agreement coefficients

The inter-annotator agreement is not computed on the same scale depending on cases:

- Case 1: 3 annotators and 2 categories



- Case 2: 2 annotators and 2 categories

→ need for a certain correction of the observed results to be able to interpret the results

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

How to compute the amount of agreement expected by chance (A_e)?

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

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

All the categories are equally likely: consequences

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

$$A_o = \frac{20+20}{50} = 0.8$$
$$A_e^S = \frac{1}{2} = 0.5$$
$$S = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

All the categories are equally likely: consequences

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

$$A_o = \frac{20+20}{50} = 0.8$$
$$A_e^S = \frac{1}{2} = 0.5$$
$$S = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

All the categories are equally likely: consequences

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

$$A_o = \frac{20+20}{50} = 0.8$$
$$A_e^S = \frac{1}{2} = 0.5$$
$$S = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

$$A_o = \frac{20+20}{50} = 0.8$$
$$A_e^S = \frac{1}{4} = 0.25$$
$$S = \frac{0.8-0.25}{1-0.25} = \mathbf{0.73}$$

π [Scott, 1955]

π

Different chance for different categories.

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

Probability of both annotators picking the same category:

$$A_e^{\pi} = \sum_q (\frac{n_q}{N})^2 = \frac{1}{N^2} \sum_q n_q^2$$

Comparing S and π

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

$A_o = 0.8$
$S = \mathbf{0.6}$

$A_o = 0.8$
$S = \mathbf{0.73}$

Comparing S and π

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

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

$A_o = 0.8$
$S = \mathbf{0.73}$

Comparing S and π

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

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

$A_o = 0.8$
$S = \mathbf{0.73}$
$A_e^\pi = \frac{((\frac{25+25}{2})^2+(\frac{25+25}{2})^2)}{50^2} = 0.5$
$\pi = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$

κ [Cohen, 1960]

κ

Different annotators have different interpretations of the instructions (bias/prejudice). κ 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: $\frac{n_{A_1 q_a}}{i} . \frac{n_{A_2 q_a}}{i}$

Probability of both annotators picking the same category:

$$A_e^{\kappa} = \sum_q \frac{n_{A_1 q}}{i} . \frac{n_{A_2 q}}{i} = \frac{1}{i^2} \sum_q n_{A_1 q} n_{A_2 q}$$

Comparing π and κ

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

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

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

Comparing π and κ

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

$$A_o = 0.8$$
$$A_e^{\pi} = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$$
$$\pi = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$
$$A_e^{\kappa} = \frac{(\frac{25x25}{50}) + (\frac{25x25}{50})}{50} = 0.5$$
$$\kappa = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

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

Comparing π and κ

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	C	D	Total
Yes	**20**	5	0	0	25
No	5	**20**	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	**50**

$$A_o = 0.8$$
$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$$
$$\pi = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$
$$A_e^\kappa = \frac{(\frac{25x25}{50}) + (\frac{25x25}{50})}{50} = 0.5$$
$$\kappa = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

$$A_o = 0.8$$
$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0.5$$
$$\pi = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$
$$A_e^\kappa = \frac{(\frac{25x25}{50}) + (\frac{25x25}{50})}{50} = 0.5$$
$$\kappa = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

Comparing π and κ

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	Total
Yes	**24**	8	32
No	14	**24**	38
Total	38	32	**70**

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

$$A_o = 0.68$$
$$A_e^\pi = \frac{((\frac{38+32}{2})^2 + (\frac{32+38}{2})^2)}{70^2} = 0.5$$
$$\pi = \frac{0.68-0.5}{1-0.5} = \mathbf{0.36}$$

Comparing π and κ

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	Total
Yes	**24**	8	32
No	14	**24**	38
Total	38	32	**70**

$$A_o = 0.8$$
$$A_e^{\pi} = \frac{((\frac{25+25}{2})^2+(\frac{25+25}{2})^2)}{50^2} = 0.5$$
$$\pi = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$
$$A_e^{\kappa} = \frac{(\frac{25x25}{50})+(\frac{25x25}{50})}{50} = 0.5$$
$$\kappa = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

$$A_o = 0.68$$
$$A_e^{\pi} = \frac{((\frac{38+32}{2})^2+(\frac{32+38}{2})^2)}{70^2} = 0.5$$
$$\pi = \frac{0.68-0.5}{1-0.5} = \mathbf{0.36}$$

Comparing π and κ

	Yes	No	Total
Yes	**20**	5	25
No	5	**20**	25
Total	25	25	**50**

	Yes	No	Total
Yes	**24**	8	32
No	14	**24**	38
Total	38	32	**70**

$$A_o = 0.8$$
$$A_e^{\pi} = \frac{((\frac{25+25}{2})^2+(\frac{25+25}{2})^2)}{50^2} = 0.5$$
$$\pi = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$
$$A_e^{\kappa} = \frac{(\frac{25x25}{50})+(\frac{25x25}{50})}{50} = 0.5$$
$$\kappa = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$$

$$A_o = 0.68$$
$$A_e^{\pi} = \frac{((\frac{38+32}{2})^2+(\frac{32+38}{2})^2)}{70^2} = 0.5$$
$$\pi = \frac{0.68-0.5}{1-0.5} = \mathbf{0.36}$$
$$A_e^{\kappa} = \frac{(\frac{38x32}{70})+(\frac{32x38}{70})}{70} = 0.49$$
$$\kappa = \frac{0.68-0.49}{1-0.49} = \mathbf{0.37}$$

S, π and κ

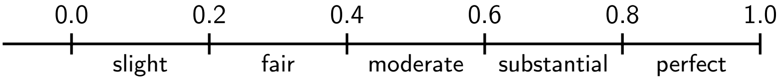
For any sample:

$$A_e^\pi \geqslant A_e^S \qquad \pi \leqslant S$$
$$A_e^\pi \geqslant A_e^\kappa \qquad \pi \leqslant \kappa$$

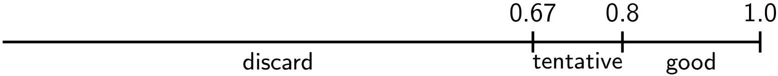
What is a "good" κ (or π or S)?

Scales of interpretation of Kappa

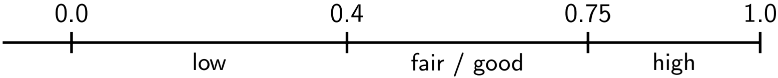
[Landis and Koch, 1977]



[Krippendorff, 1980]



[Green, 1997]



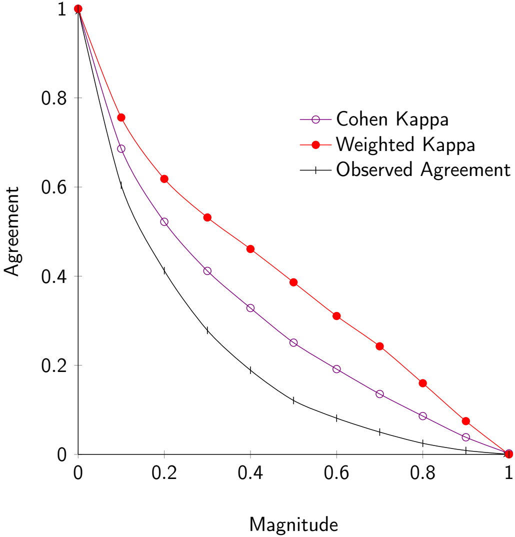
Giving meaning to the obtained results [COLING 2012a]

Creation of a "Richter" tool which:

- takes as input a reference annotation (real or automatically generated)
- generates degradations of a certain magnitude (from 0 to 1)
- applies one or several inter-annotator agreement metrics on each set of annotations (corresponding to a magnitude of degradation)

Richter on the TCOF-POS corpus

No prevalence, but proximity between categories (is taken into account):



Biases

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 "by chance"

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

Expert ?

Experts:

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

... some contradictions and shortfalls:

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

- Precision, recall, F-measure
- Accuracy (exactitude)
- Observed agreement
- S, κ, π
- Meaning

Artstein, R. and Poesio, M. (2008).
Inter-coder agreement for computational linguistics.
Computational Linguistics, 34(4):555–596.

Bennett, E. M., Alpert, R., and C.Goldstein, A. (1954).
Communications through limited questioning.
Public Opinion Quarterly, 18(3):303–308.

Cohen, J. (1960).
A coefficient of agreement for nominal scales.
Educational and Psychological Measurement, 20(1):37–46.

Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).
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.

Desrosières, A. (2008).

Pour une sociologie historique de la quantification : L'Argument statistique.
Presses de l'école des Mines de Paris.

Fort, K. and Sagot, B. (2010).
Influence of pre-annotation on POS-tagged corpus development.
In Proceedings of the Fourth ACL Linguistic Annotation Workshop, pages 56–63, Uppsala, Suède.

Green, A. M. (1997).
Kappa statistics for multiple raters using categorical classifications.
In Proceedings of the Twenty-Second Annual Conference of SAS Users Group, San Diego, USA.

Hripcsak, G. and Rothschild, A. S. (2005).
Agreement, the f measure, and reliability in information retrieval.
Journal of the American Medical Informatics Association (JAMIA), 12(3):296–298.

Krippendorff, K. (1980).
Content Analysis: An Introduction to Its Methodology.
Sage, Beverly Hills, CA., USA.

Landis, J. R. and Koch, G. G. (1977).
The measurement of observer agreement for categorical data.
Biometrics, 33(1):159–174.

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012).
Manual corpus annotation: Evaluating the evaluation metrics.
In Proceedings of the International Conference on Computational Linguistics (COLING), pages 809–818, Mumbaï, Inde.
Poster.

Mathet, Y., Widlöcher, A., and Métivier, J.-P. (2015).
The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment.
Computational Linguistics, 41(3):437–479.

Nédellec, C., Bessières, P., Bossy, R., Kotoujansky, A., and Manine, A.-P. (2006).
Annotation guidelines for machine learning-based named entity recognition in microbiology.
In et C. Nédellec, M. H., editor, Proceedings of the Data and text mining in integrative biology workshop, pages 40–54, Berlin, Allemagne.

Reidsma, D. and Carletta, J. (2008).
Reliability measurement without limits.
Computational Linguistics, 34(3):319–326.

Scott, W. A. (1955).
Reliability of content analysis: The case of nominal scale coding.
Public Opinion Quaterly, 19(3):321–325.