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

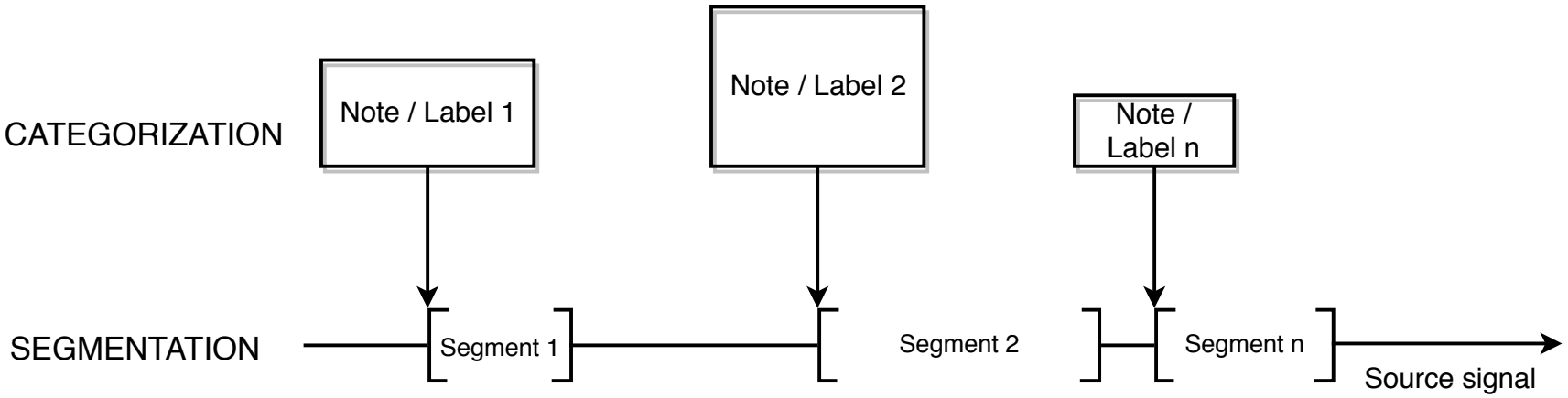
Agreement coefficients/metrics

Beyond categorization : unitizing

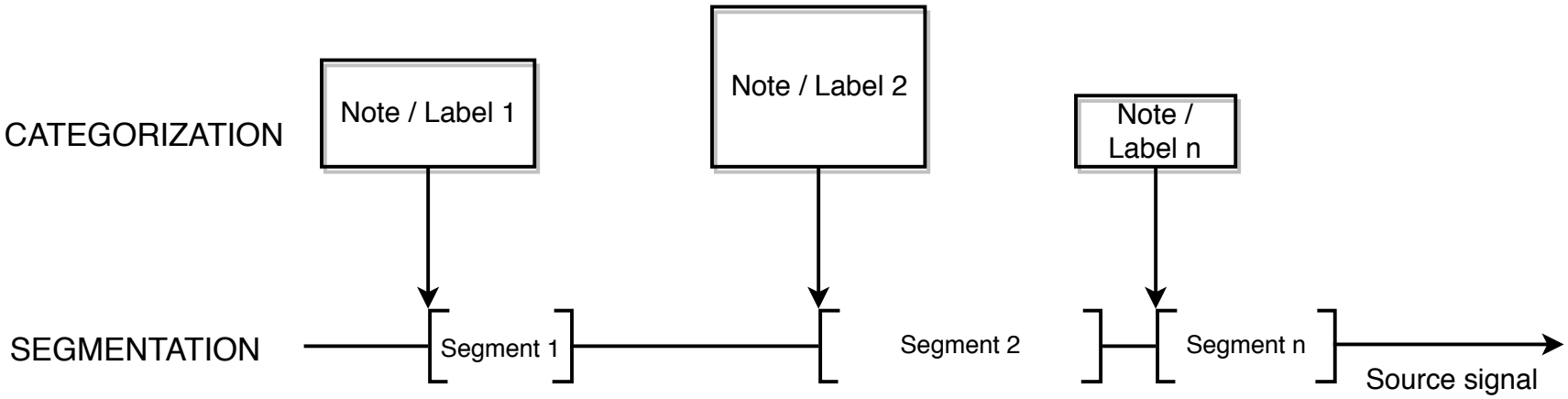
Giving meaning to the metrics

Annotating : back on chance

Annotating


CATEGORIZATION
Note / Label 1
Note / Label 2
Note / Label n
SEGMENTATION
Segment 1
Segment 2
Segment n
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 précisémment 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]

Coding

"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

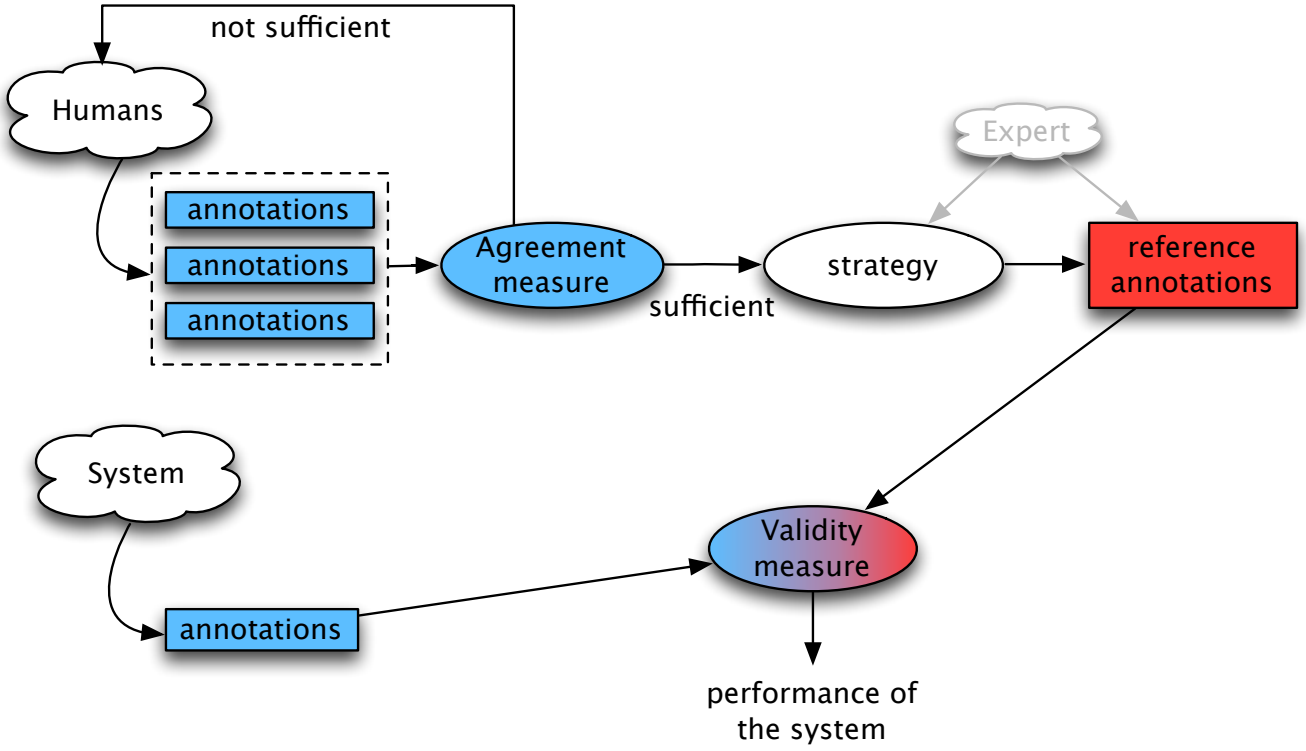
Beyond categorization : unitizing

Giving meaning to the metrics

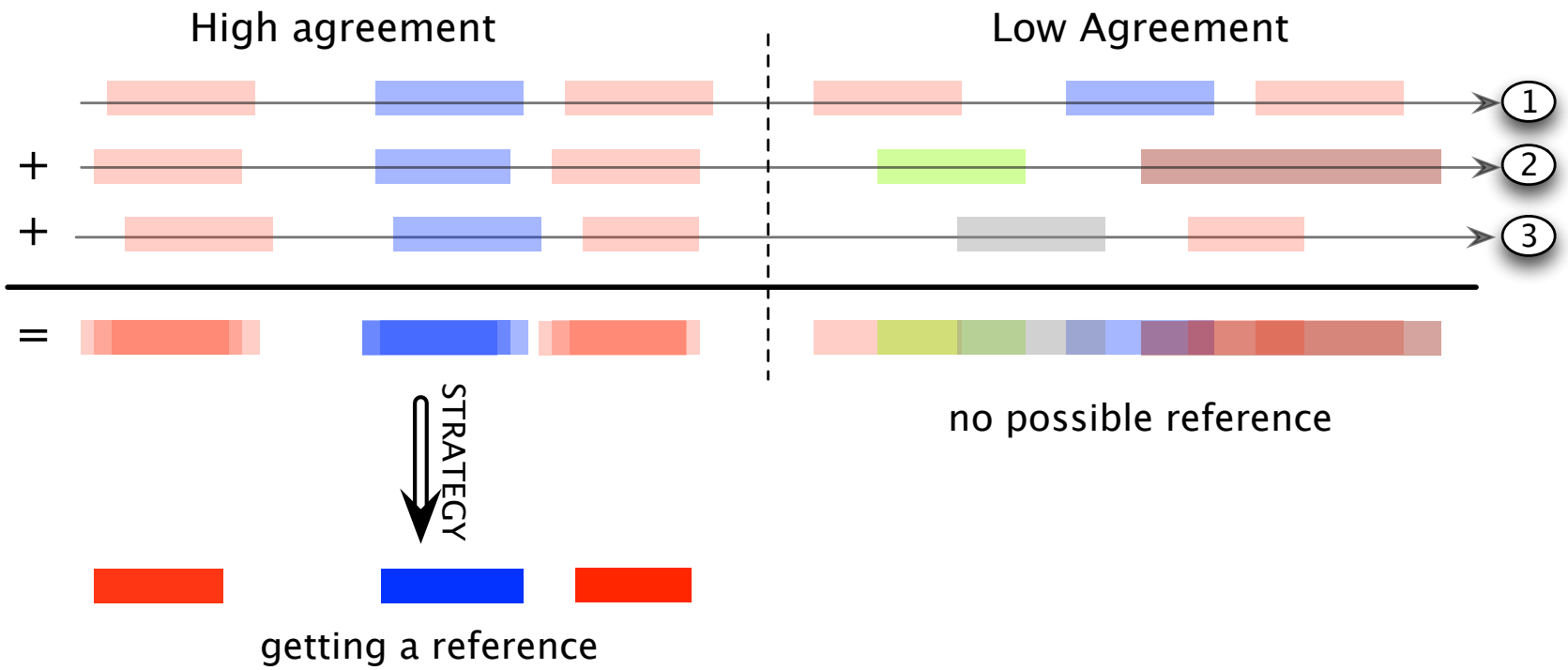
Annotating : back on chance

Four main concepts interacting

- Multi-annotations
- Agreement (and agreement measures)
- Reference (Gold-Standard)
- Validity (and validity measures)

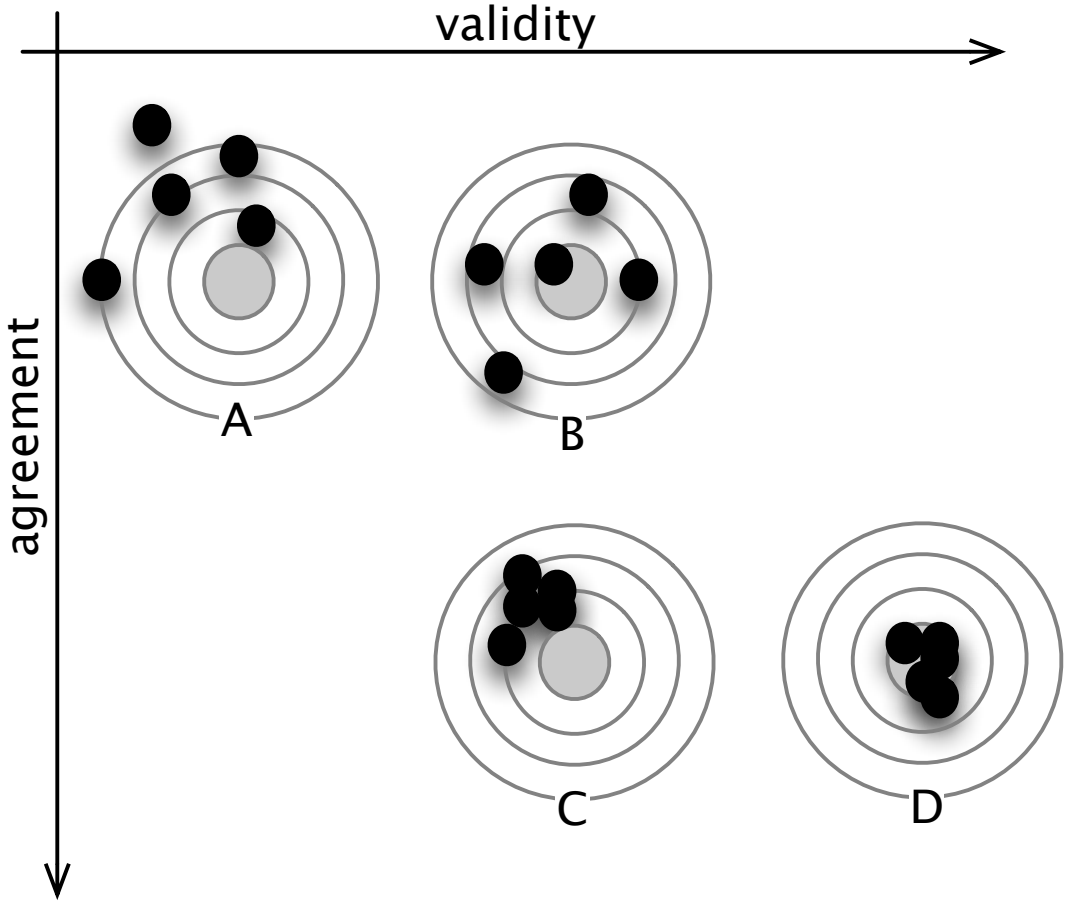


From agreement to a reference



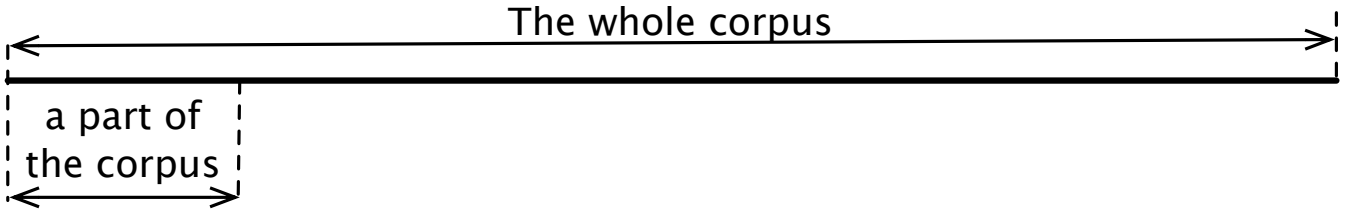
Agreement *vs* Validity

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



Agreement *vs* Validity

Benefiting from reproducibility



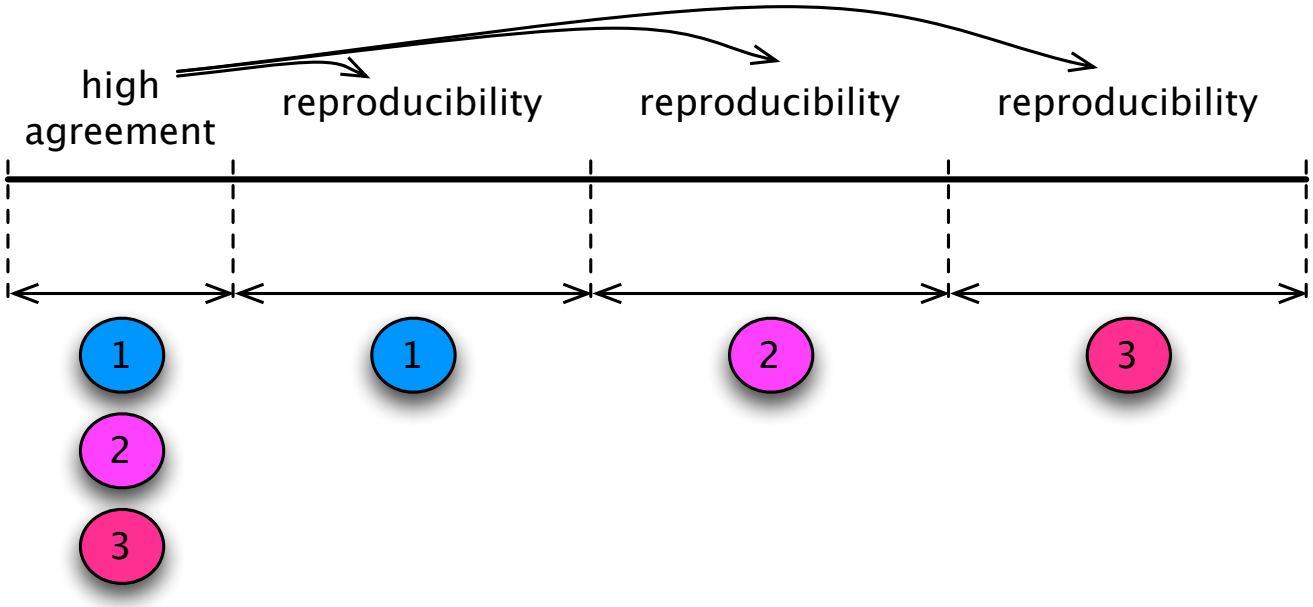
Three annotators named 1, 2 and 3 : 

Agreement *vs* Validity

Benefiting from reproducibility

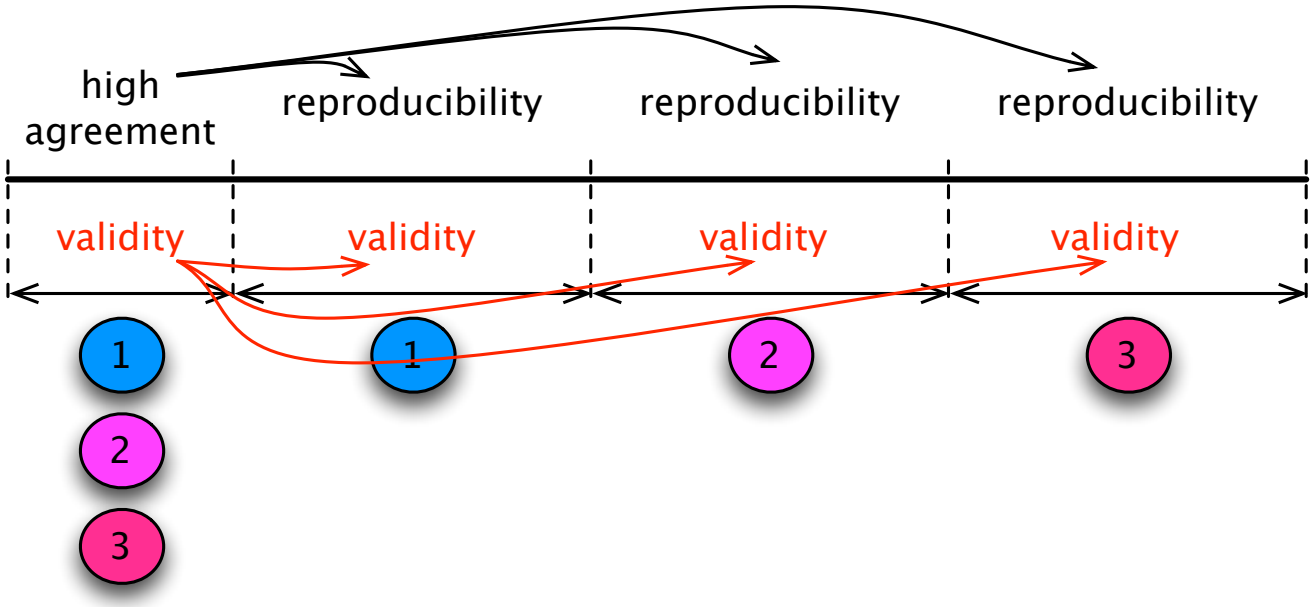
Agreement *vs* Validity

Benefiting from reproducibility



Agreement *vs* Validity

Benefiting from reproducibility



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

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

- Silence :
- Precision :

- Noise :

Precision/Recall : back to basics

- Recall : measures the quantity of found annotations

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

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

- Noise :

Precision/Recall : back to basics

- Recall : measures the quantity of found annotations

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

- Silence : *complement* of recall (correct annotations not found)
- Precision : measures the quality of found annotations

$$Precision = \frac{Nb\ of\ correct\ found\ annotations}{Total\ nb\ of\ found\ annotations}$$

- Noise :

Precision/Recall : back to basics

- Recall : measures the quantity of found annotations

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

- Silence : *complement* of recall (correct annotations not found)
- Precision : measures the quality of found annotations

$$Precision = \frac{Nb\ of\ correct\ found\ annotations}{Total\ nb\ of\ found\ annotations}$$

- 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

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

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

It is a special case of the general Fβ measure :

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

The value of β 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) :

Sentence	A	B	Agreed ?
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 : confusion matrix

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

Observed agreement (A_o)

proportion of annotations on which the annotators agree

Here :

Synthetic representation : confusion matrix

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

Observed agreement (A_o)

proportion of annotations on which the annotators agree

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

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 :

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	3
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	?
🧩	🧩	🧩	?
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	?
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	?

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	1

Correction and follow up

A_1	A_2	A_3	Agreeing pairs
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	3
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	1
🧩	🧩	🧩	1

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

What if...

Exercise (follow up)

- each unit has to be annotated
- 2 categories
- ~~3~~ 2 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 :

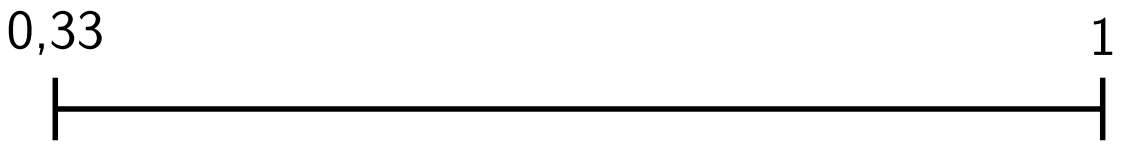
- Case #1 : 3 annotators and 2 categories

- Case #2 : 2 annotators and 2 categories

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



- Case #2 : 2 annotators and 2 categories

$\rightarrow$ 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
$\rightarrow$ similar to the concept of *baseline*

Taking chance into account

Expected agreement A_e)

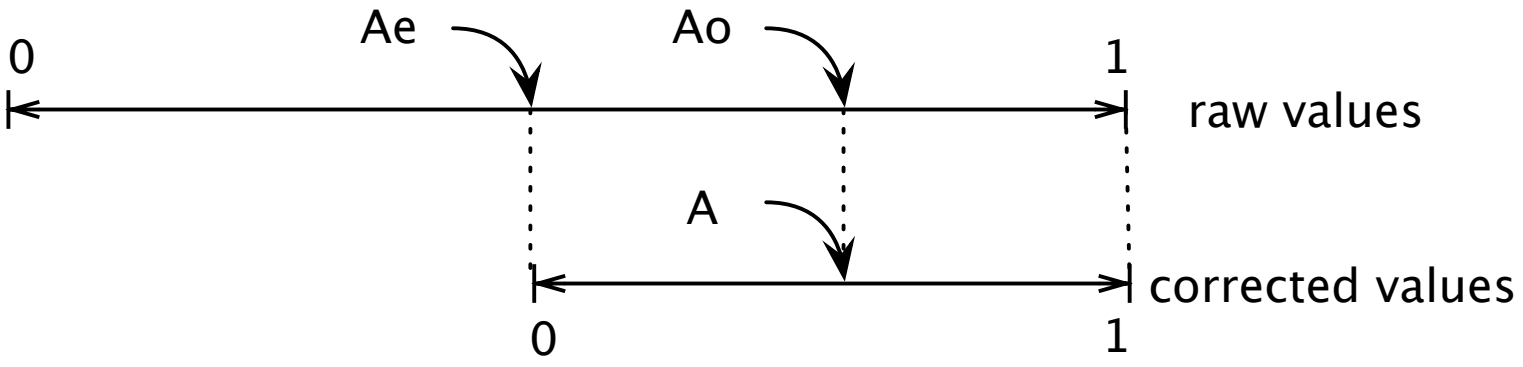
expected value of the observed agreement

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



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 :

$$\begin{aligned} A_{deducted} &= A_o - A \\ &= A_o - \frac{A_o - A_e}{1 - A_e} \\ &= \frac{A_e}{1 - A_e} \cdot (1 - A_o) \end{aligned} \tag{1}$$

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

	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**

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

$S = \frac{0.8-0.5}{1-0.5} = \mathbf{0.6}$

	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**

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

$S = \frac{0.8-0.25}{1-0.25} = \mathbf{0.73}$

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**

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

	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**

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

Comparing S and π

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

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

	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**

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

Comparing S and π

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

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

	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**

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

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 : $\frac{n_{A_1 q_a}}{i} \cdot \frac{n_{A_2 q_a}}{i}$

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 π and κ

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

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

	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**

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

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**

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

	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**

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

Comparing π and κ

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

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

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

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

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**

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

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

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

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) ?

Evaluating the quality of manual annotation

Agreement coefficients/metrics

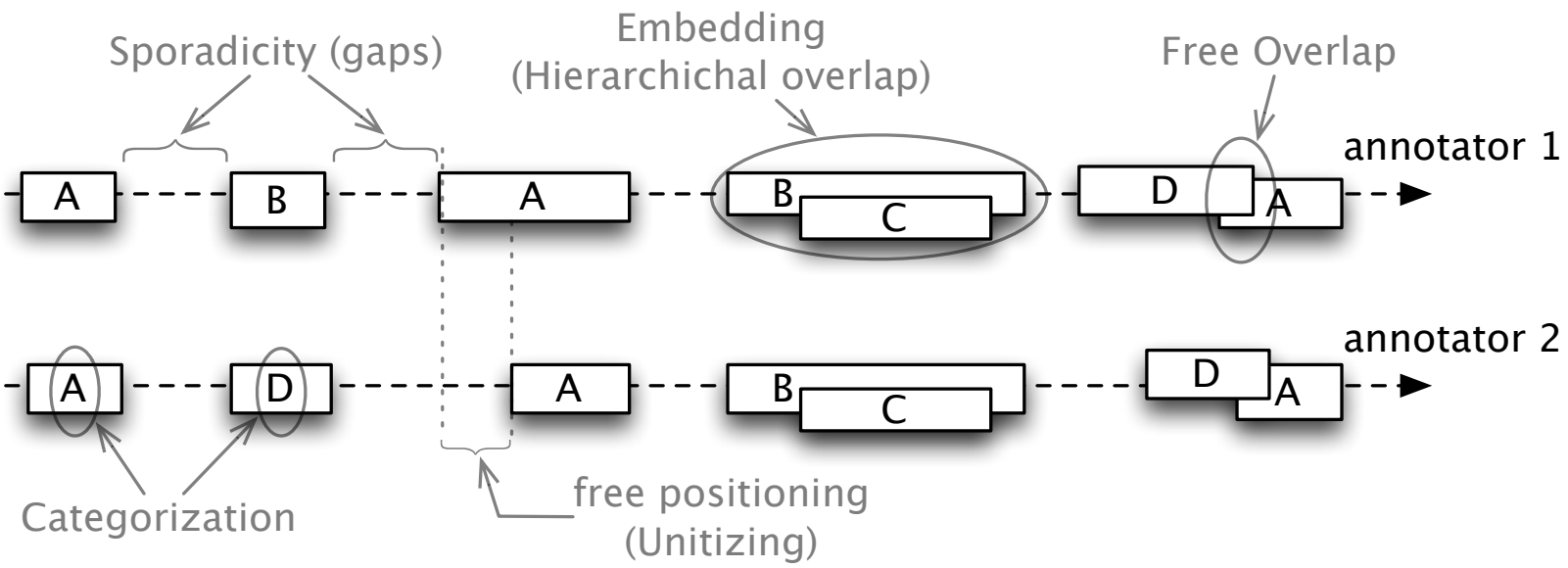
Beyond categorization : unitizing
- Introducing unitizing
- Three different strategies to assess agreement on unitizing
- Further comparisons between strategies
- Introducing the gamma coefficients

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 α were designed for unitizing
- A challenge : **how to compare annotations from different annotators ?**



Evaluating the quality of manual annotation

Agreement coefficients/metrics

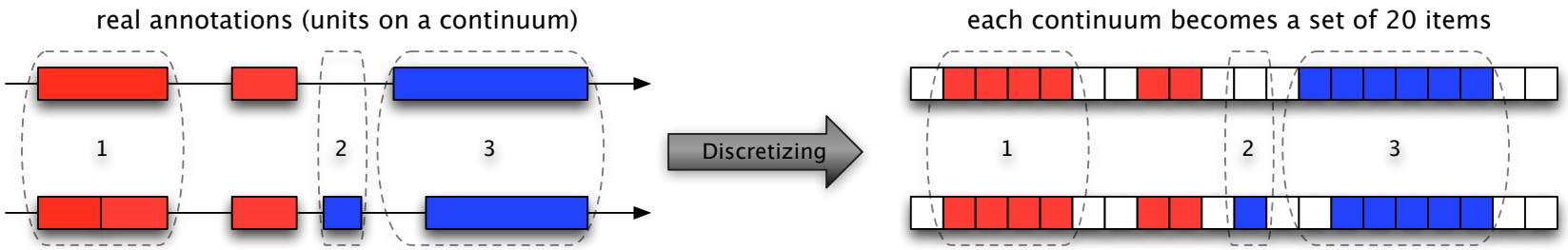
Beyond categorization : unitizing
- Introducing unitizing
- Three different strategies to assess agreement on unitizing
- Further comparisons between strategies
- Introducing the gamma coefficients

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 :

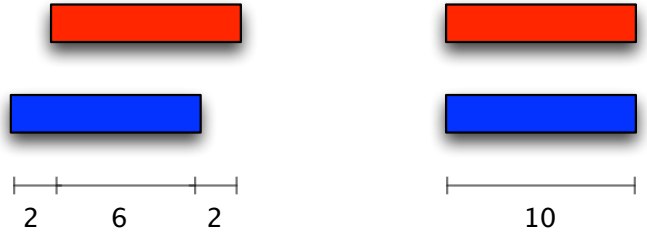
- two contiguous units seen as one in (1)
- a slight shift in (3) is as important as a false positive in (2)
- agreement on blanks (what is not annotated)
- unit overlapping is not possible

Still seen when reviewing for ACL 2017 and 2018 (A* conference in NLP)

Second strategy : Relying on intersections

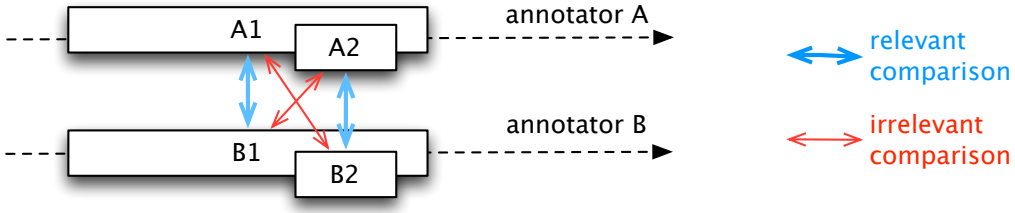
- Krippendorff's α coefficients go further
- They rely on intersecting parts of units rather than on atoms
- Biases are more limited, but still present (as shown by the CST)
- I discovered a paradox in $_U\alpha$: positional agreement may increase overall disagreement

$$2^2 + 6^2 + 2^2 < 10^2$$



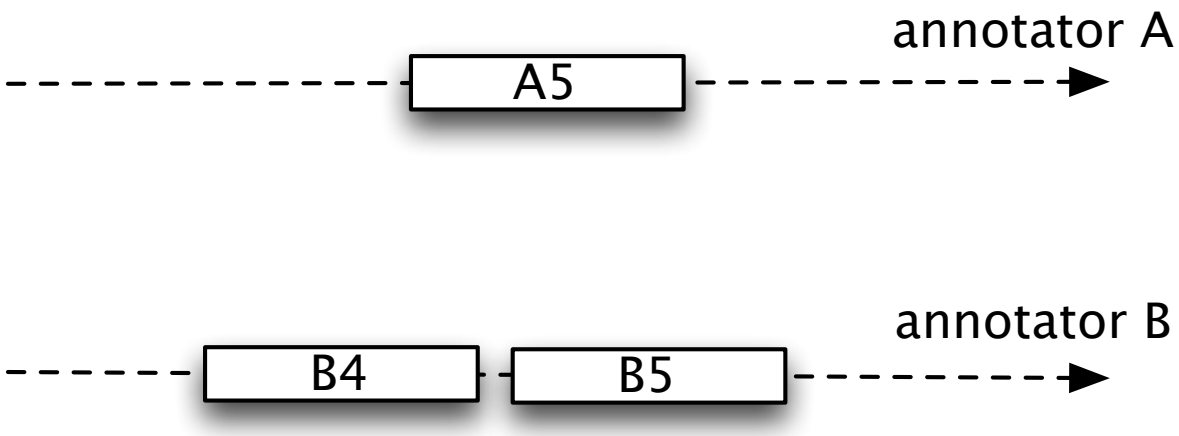
The third strategy (by Gamma) : Relying on an alignment

- 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



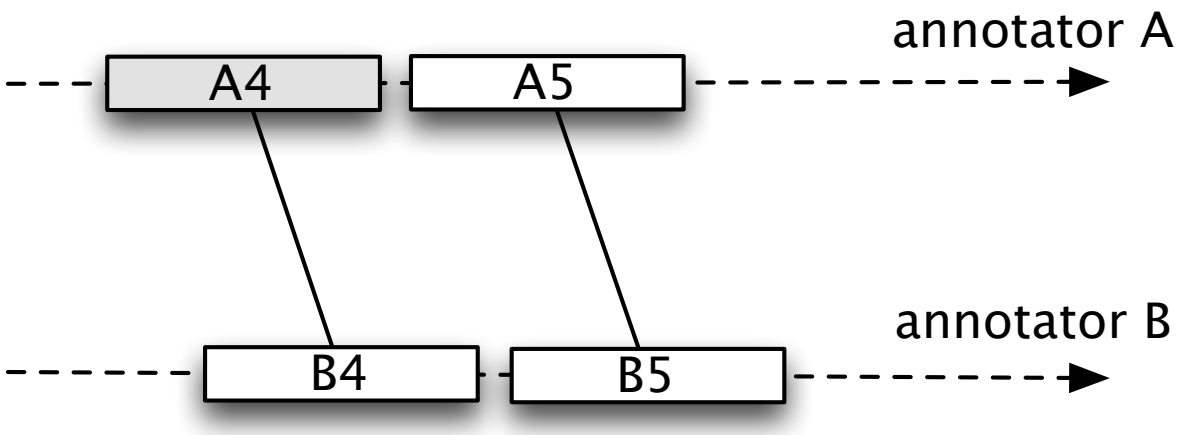
Alignment

A **holistic** mechanism



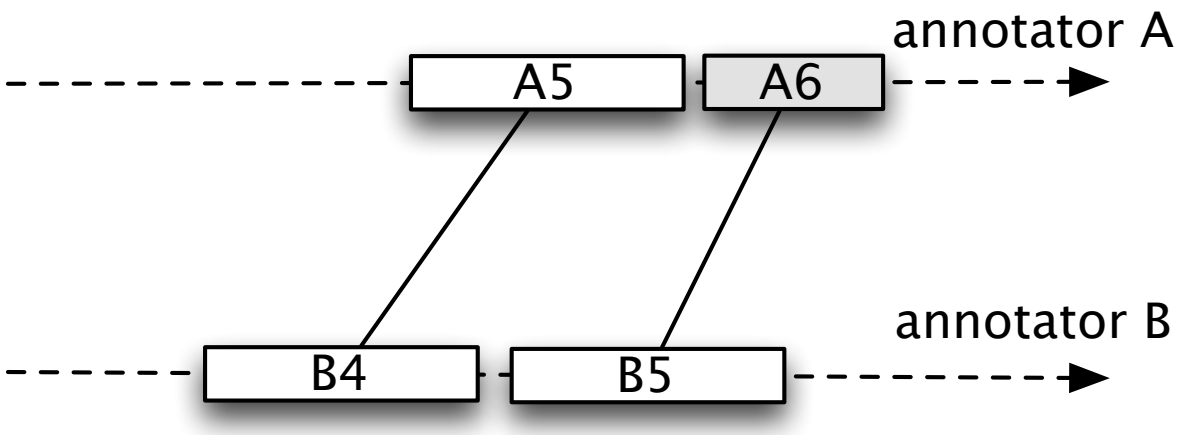
Alignment

A **holistic** mechanism



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

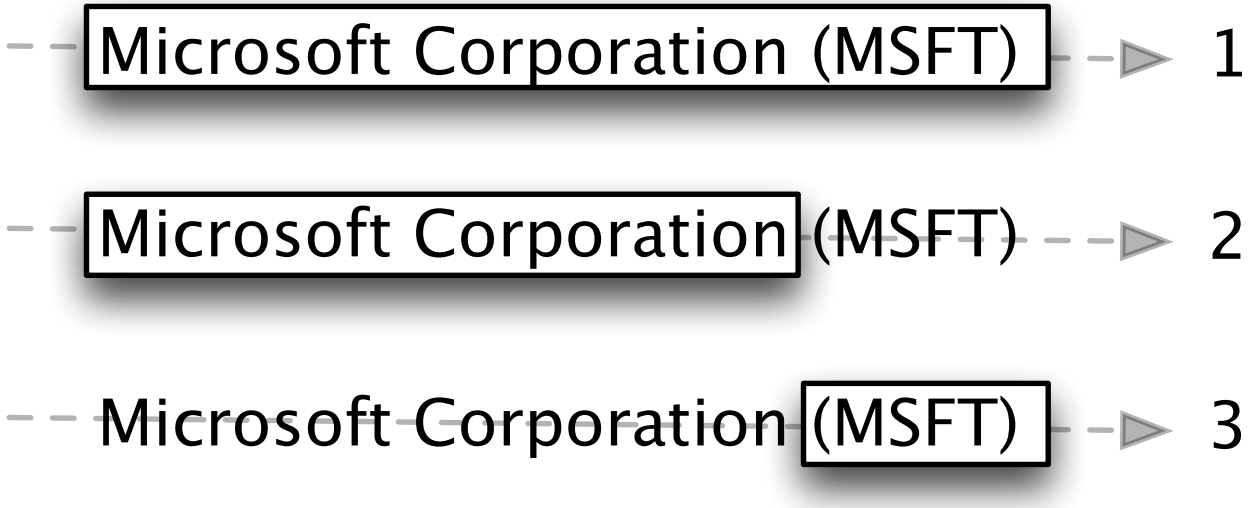
Beyond categorization : unitizing
- Introducing unitizing
- Three different strategies to assess agreement on unitizing
- Further comparisons between strategies
- Introducing the gamma coefficients

Giving meaning to the metrics

Annotating : back on chance

Alignments vs intersections

Units with no intersection may refer to the same intent



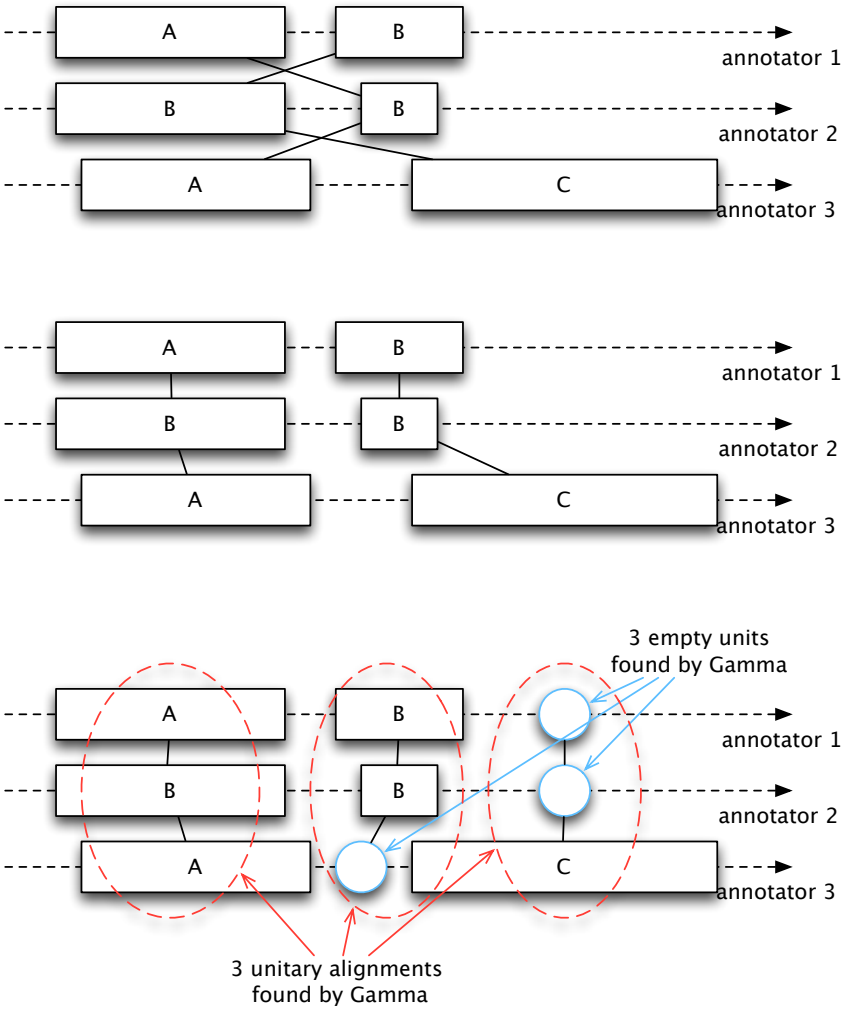
The agreement metric γ

γ 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. γ 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 γ

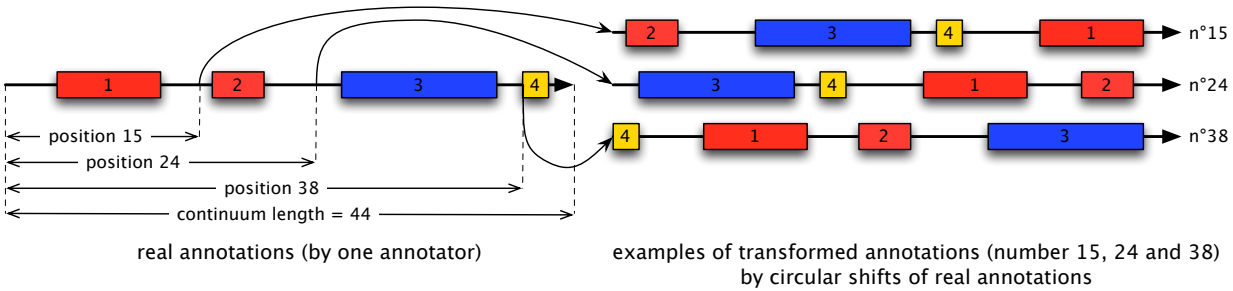
Example of 3 possible alignments :



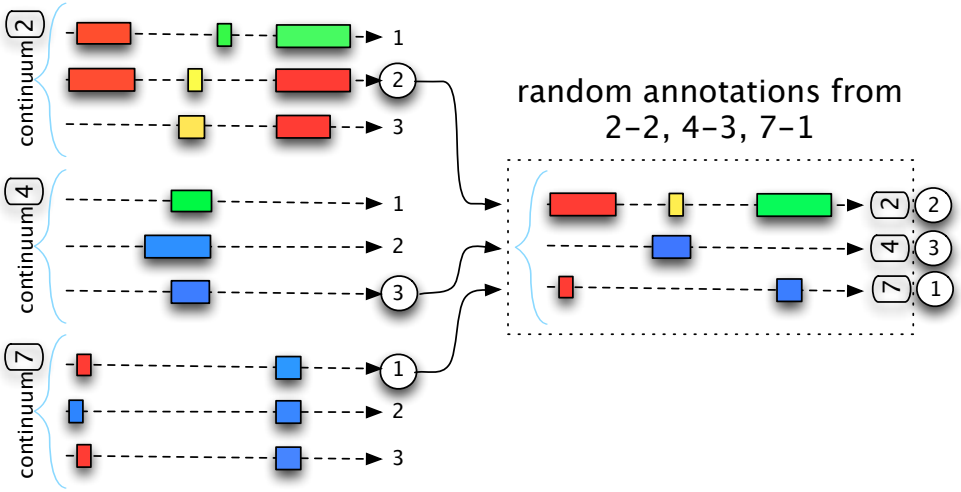
Chance correction of γ

Re-sampling annotations to get the part of "chance" of γ

1. Re-sampling at local level



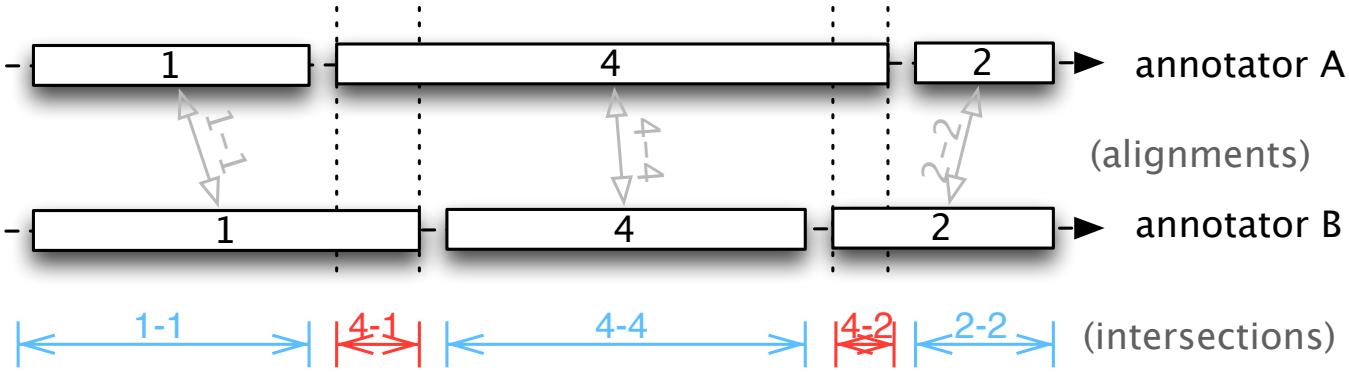
2. Re-sampling at corpus level



Focusing on categorization with γ_{cat}

γ_{cat} : a complement to γ focused on categories

- γ 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
- γ_{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

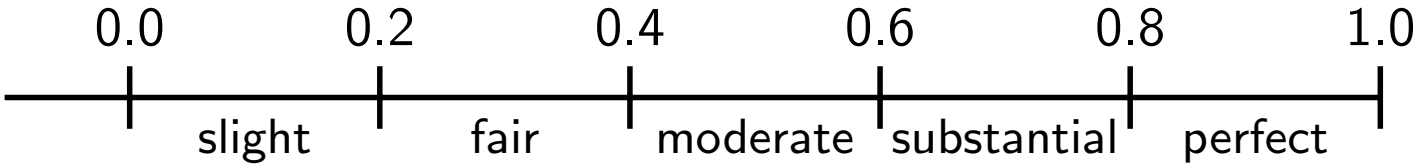
Beyond categorization : unitizing

Giving meaning to the metrics
- Interpretations
- Semantics

Annotating : back on chance

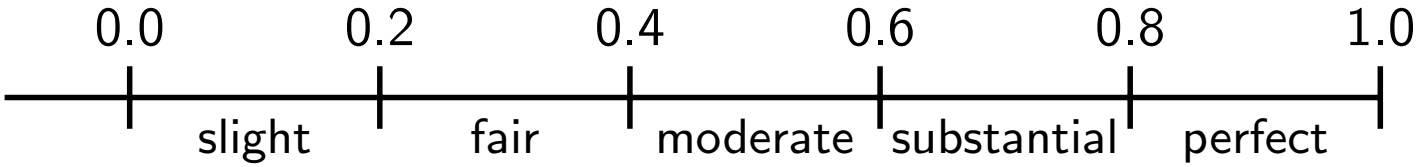
Scales for the interpretation of Kappa

- Landis and Koch, 1977

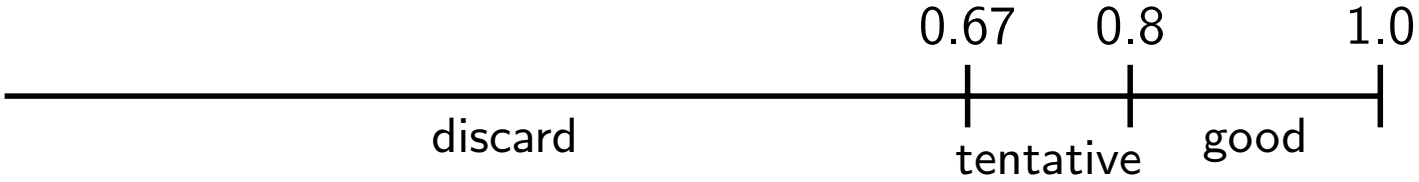


Scales for the interpretation of Kappa

- Landis and Koch, 1977

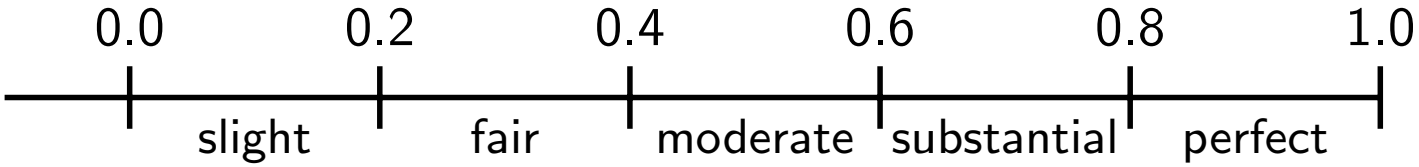


- Krippendorff, 1980

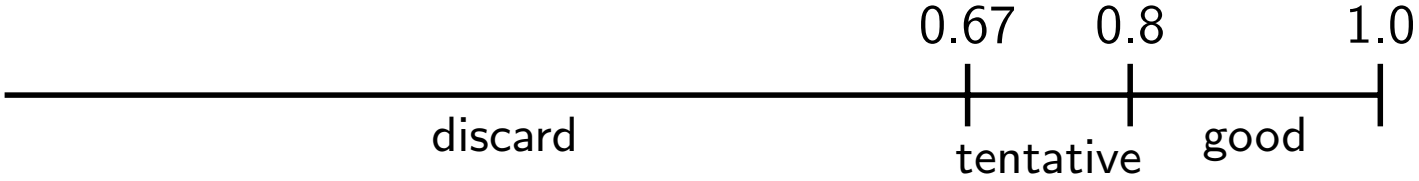


Scales for the interpretation of Kappa

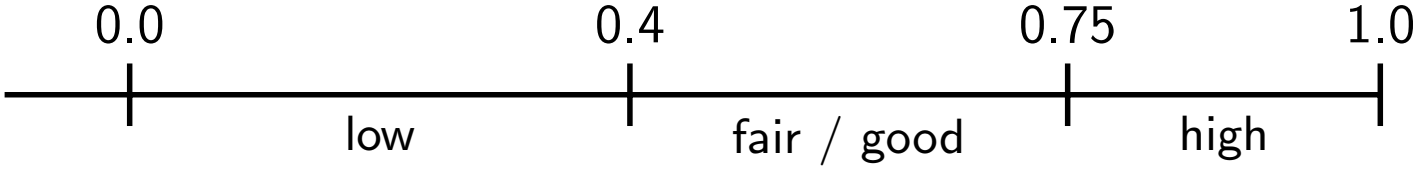
- Landis and Koch, 1977



- Krippendorff, 1980

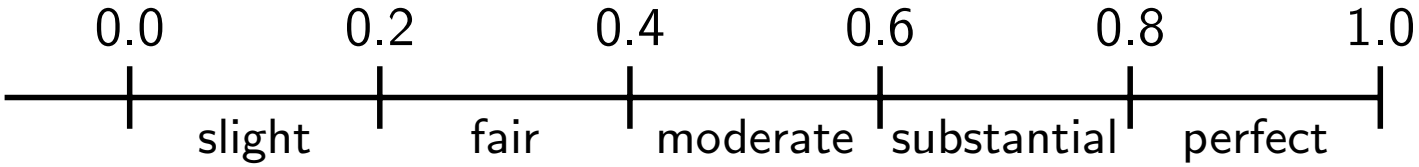


- Green, 1997

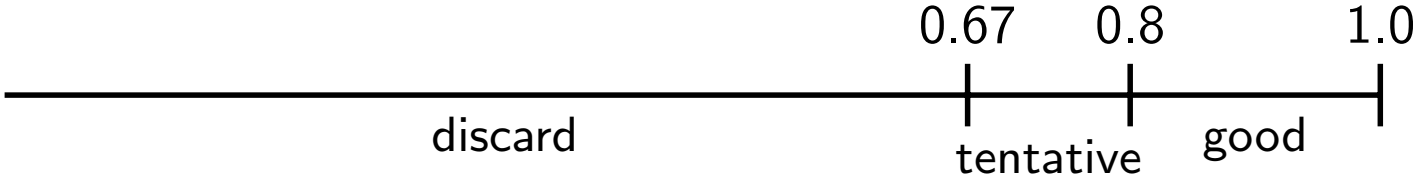


Scales for the interpretation of Kappa

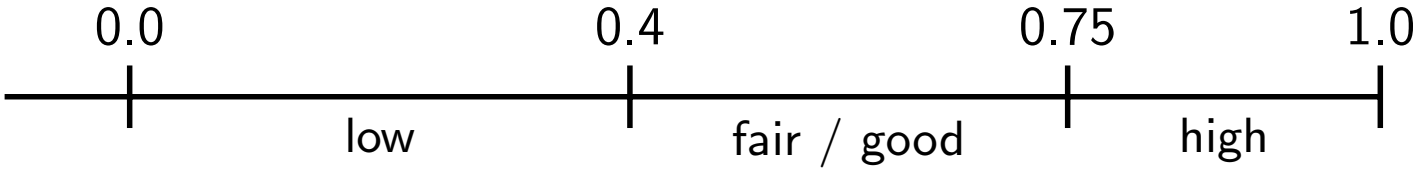
- Landis and Koch, 1977



- Krippendorff, 1980



- Green, 1997



- "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

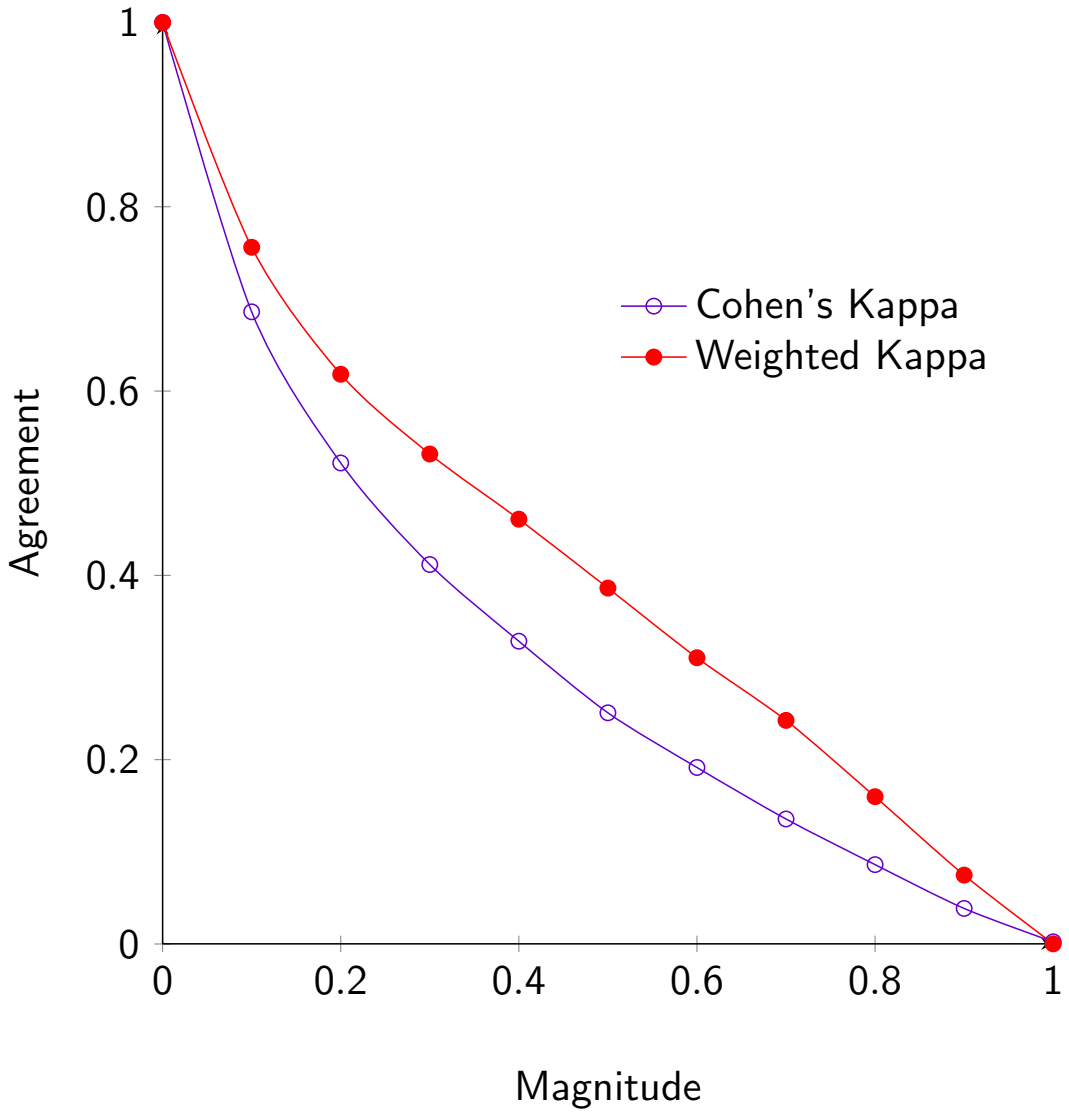
Giving meaning to the obtained results [COLING 2012a]

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 :



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]

Expert ?

Experts :

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

... contradictions and shortfalls :

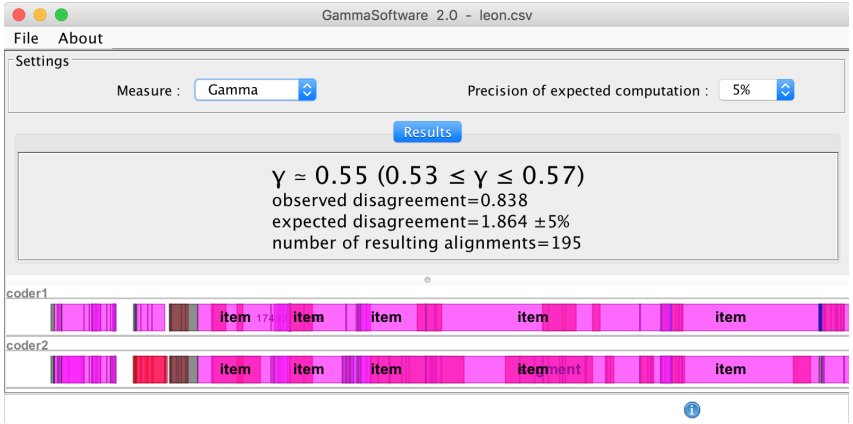
→ to annotate structured named entities in old press, do we need names entities specialists or historians ?

Appendices
Examples
Bibliography

Annodis

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

Annodis Agreement Values



$\gamma_{cat} \simeq 0.74\ (0.73 \leq \gamma_{cat} \leq 0.75)$
observed disagreement=0.090
expected disagreement=0.347 ±5%
number of resulting alignments=195.0
$\gamma_{k(amorce)} \simeq 1.000$
$\gamma_{k(cloture)} \simeq 1.000$
$\gamma_{k(enumeraTheme)} \simeq 0.936$
$\gamma_{k(indice)} \simeq 0.636$
$\gamma_{k(item)} \simeq 1.000$
$\gamma_{k(segment)} \simeq 1.000$
$\gamma_{k(todo)} \simeq -0.082$

Football Matches Transcriptions

K. Fort & V. Claveau, TALN 2012

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

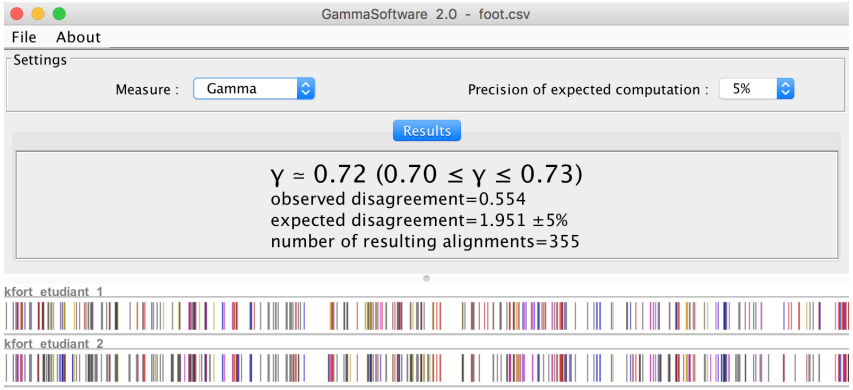
Fabien Lévêque : Sammaritano qui perd son ballon . Chamakh . Il y aura une faute sifflée par Freddy Fautrel , faute de Franck Jurietti . tout va bien . Oui Daniel ?

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 .

ID=749 ID=400

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

Football Agreement Values



$\gamma_{cat} \simeq 0.93\ (0.93 \leq \gamma_{cat} \leq 0.93)$

observed disagreement=0.056

expected disagreement=0.805 ±5%

number of resulting alignments=355.0

$\gamma_{k(A\_ActionDuPublic)} \simeq NA$

$\gamma_{k(A\_ArreterBut)} \simeq 1.000$

$\gamma_{k(A\_Centrer)} \simeq 0.000$

$\gamma_{k(A\_FaireFauteDeJeu)} \simeq 0.000$

$\gamma_{k(A\_FaireTentative2Centre)} \simeq 0.000$

$\gamma_{k(A\_HorsJeu)} \simeq 0.667$

$\gamma_{k(A\_IntercepterBallon)} \simeq 0.958$

$\gamma_{k(A\_MarquerBut)} \simeq 1.000$

$\gamma_{k(A\_PossederBallon)} \simeq -0.068$

$\gamma_{k(A\_PrendreCartonJaune)} \simeq 1.000$

$\gamma_{k(A\_PrendreRappelALOrdre)} \simeq 0.667$

$\gamma_{k(A\_RaterBut)} \simeq 0.857$

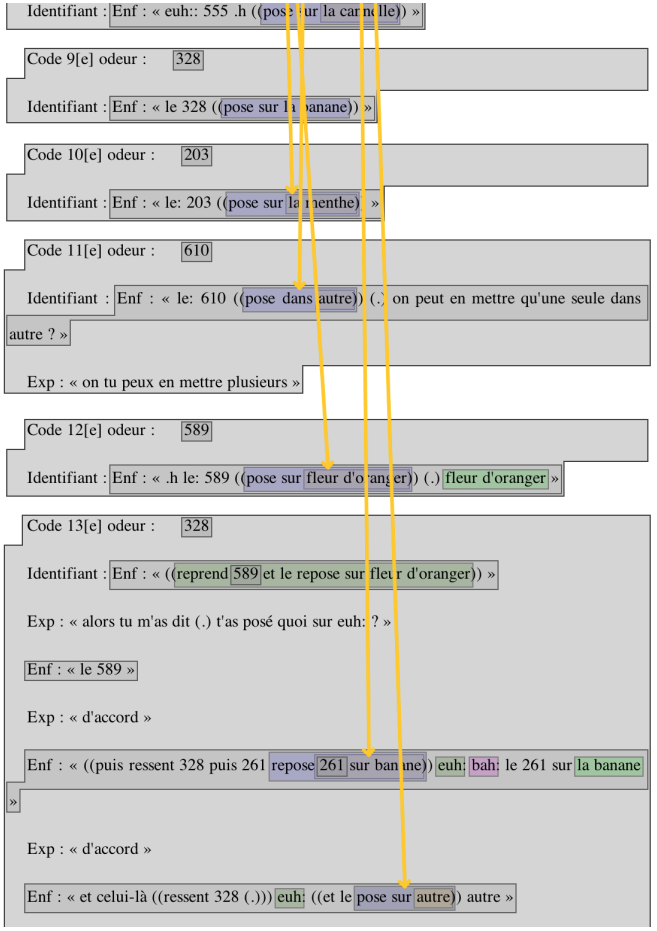
$\gamma_{k(A\_TirerCorner)} \simeq 1.000$

$\gamma_{k(A\_TirerCoupFrancDirect)} \simeq 0.833$

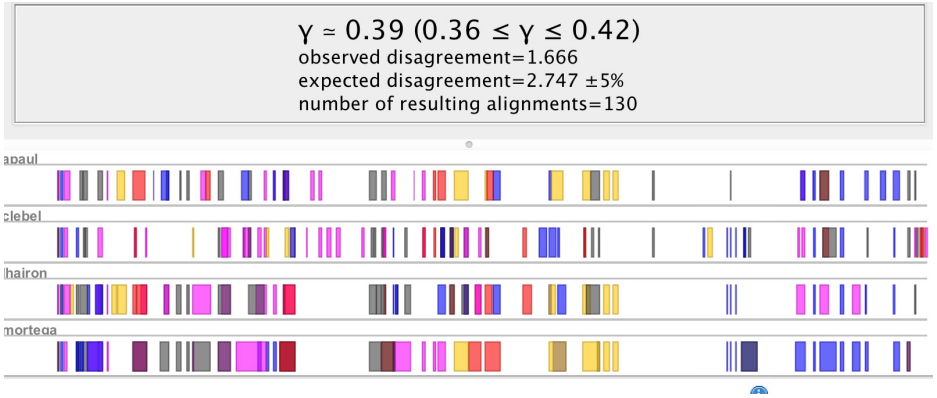
$\gamma_{k(ActionPourActeurVide)} \simeq 0.633$

Smell Recognition Dialog Transcriptions

C2O : Ongoing project in Psychology, Caen, leaded by Maryse Delaunay-El Allam



Smell Recognition Dialog Transcriptions



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

$\gamma_{k(ActionSurEmotions)} \simeq NA$
$\gamma_{k(ActivitésCulinaires)} \simeq 0.397$
$\gamma_{k(ActivitésHygièneCorporelle)} \simeq 0.896$
$\gamma_{k(ActivitésHygièneDomestique)} \simeq 0.281$
$\gamma_{k(ActivitésJardinage)} \simeq 0.721$
$\gamma_{k(ActivitésOlfactives)} \simeq 0.000$
$\gamma_{k(ActivitésPromenades)} \simeq 0.450$
$\gamma_{k(CapacitésOlfactives)} \simeq 0.863$
$\gamma_{k(CoutumeFestive)} \simeq NA$
$\gamma_{k(EnvironnementOdorant)} \simeq 0.385$
$\gamma_{k(EvènementSpirituel)} \simeq NA$
$\gamma_{k(Flairage)} \simeq 0.750$
$\gamma_{k(GesteSymbolique)} \simeq 0.494$
$\gamma_{k(Hédonisme)} \simeq 0.870$
$\gamma_{k(Imitation)} \simeq 0.000$
$\gamma_{k(Initiative\ Activité)} \simeq 0.478$

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.

Benzitoun, C., Fort, K., and Sagot, B. (2012).
TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe.
In Actes de Traitement Automatique des Langues Naturelles (TALN), pages 99–112, Grenoble, France.

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. (1989).
Comment faire des choses qui tiennent : histoire sociale et statistique.
Histoire & Mesure, 4(4) :225–242.

Desrosières, A. (2001).
Entre réalisme métrologique et conventions d'équivalence : les ambiguïtés de la sociologie quantitative.
Genèses, 2(43) :112–127.

Desrosières, A. (2008).
Pour une sociologie historique de la quantification : L'Argument statistique I.
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.

Habert, B. (2005).
Portrait de linguiste(s) à l'instrument.
Texto !, vol. X(4).

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., Mathet, Y., Bouvry, S., and Widlöcher, A. (2016).
On the reliability of unitizing textual continua : Further developments.
Quality and Quantity, 50(6) :2347–2364.

Leech, G. (1997).
Corpus annotation : Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.
Longman, Londres, Angleterre.

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

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.

Mathet, Y. and Widlöcher, A. (2016).
Évaluation des annotations : ses principes et ses pièges.
Revue T.A.L., 52(2) :73–98.

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.

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