# An experimental attempt to give a meaning to inter-annotator agreement measures


Jean-Yves Antoine
Université François Rabelais Tours, LI

Jeanne Villaneau
Université de Bretagne Sud, IRISA

Anaïs Lefeuvre-Halftermeyer
Université d'Orléans, LIFO

Dany Bregeon
Université d'Orléans, LIFO



## MACHINE LEARNING & ANNOTATIONS

Training data are frequently provided by a manual annotation

**Example (NLP) :** attributing a category to some previously identified words

*ADJ* *N* *N*

*La maison était triste parce qu'elle avait des remords ; elle avait des remords parce qu'elle cachait un crime. "Oh! qui dit que c'est un crime? reprit Villefort*

*N* *N*

Categorization

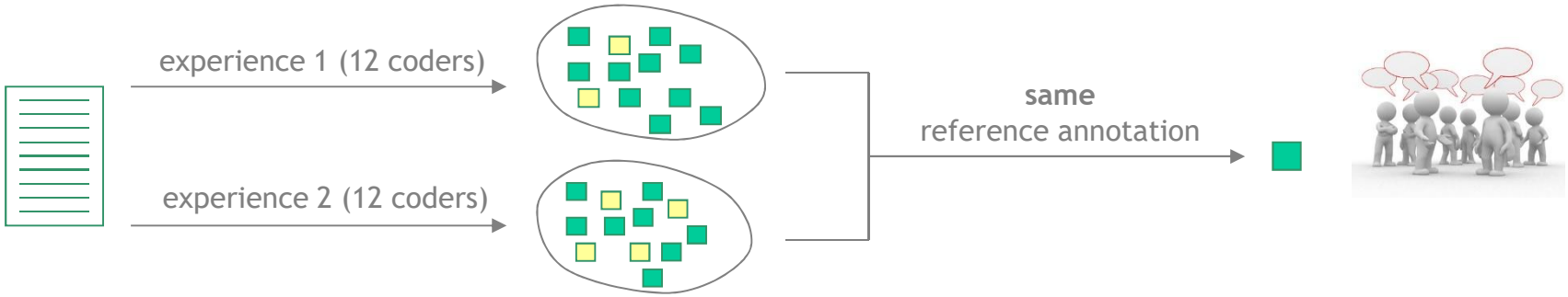
## QUALITY OF THE TRAINING DATA

Need for an objective estimation of the data reliability

# Data reliability

## ANNOTATION DATA RELIABILITY

Quality criterion : **reproducibility**



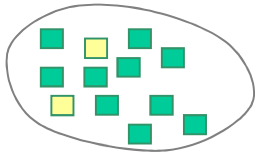

## INTER-CODER AGREEMENT

Estimation of the data reproducibility

*The more coders agree on the data they have produced, the more their annotations are likely to be reproduced by **any other set of coders***

# Data reliability

## RAW INTER-CODER AGREEMENT



annotation
(12 coders)

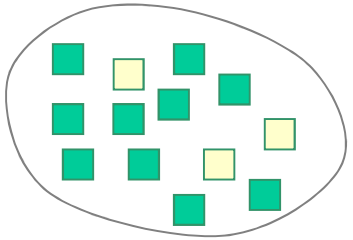
majority
vote

observed agreement

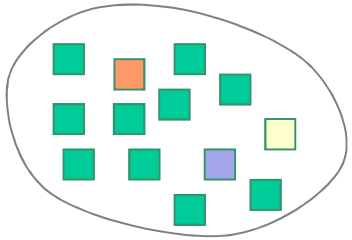
$A_o$ = 10/12 = 83.3 %

### Limitation : estimation biases

- **Example : number of categories and task difficulty** – the lesser the number of categories, the easier the chance to observe an agreement



2 categories
$A_o$ = 10/13 = 76. %

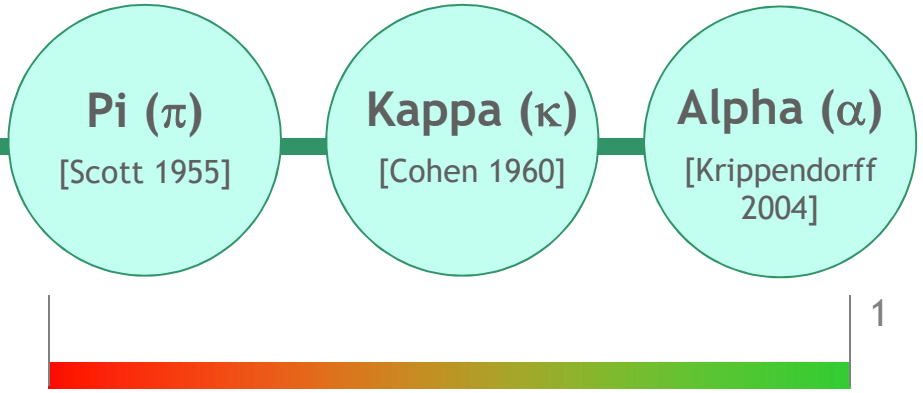


4 categories
$A_o$ = 10/13 = 76. %

- **Solution : chance corrected agreement metrics**

# Inter-coders agreement

## SEVERAL METRICS




### Agreement measure formula

Let $A_o$ =*observed agreement*, $A_e$ =*(an estimation of) chance agreement*,
$D_o$ =*observed disagreement*, $D_e$ =*(an estim. of) chance disagreement*,

$$Measure = \frac{A_o - A_e}{1 - A_e} = \frac{D_e - D_o}{D_e} = 1 - \frac{D_o}{D_e}$$

**Every metric differs from others the way they estimate chance agreement $A_e$**

# Inter-coders agreement

## SEVERAL METRICS

- **Pi ($\pi$)** [Scott 1955]
- **Kappa ($\kappa$)** [Cohen 1960]
- **Alpha ($\alpha$)** [Krippendorff 2004]

## SEVERAL QUESTIONS

**Metric reliability** – Is one metric more reliable than the other ones ?

Antoine J.-Y., Villaneau J., Lefeuvre A. (2014) Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. Proc. *14th EACL'2014*, Gothenburg, Sweden

**Metric interpretability** – How can we interpret the agreement value given by a specific metric ?

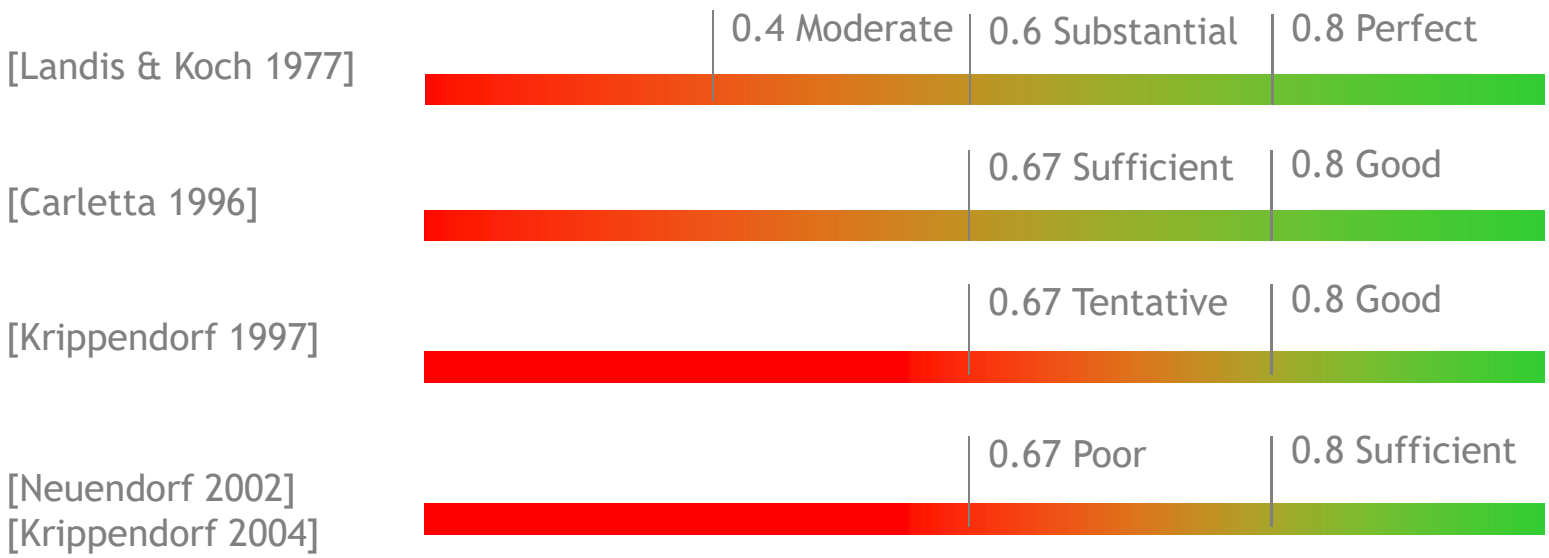


## METRICS INTERPRETABILITY

[internship Dany BREGEON, 2018]

# Metrics interpretability

## TRUSTABILITY OF DATA RELIABILITY THRESHOLD




- No scientific consensus on a satisfactory level of agreement
- No straightforward interpretation of agreement values

# Metrics interpretability

## TRUSTABILITY OF DATA RELIABILITY THRESHOLD

**Idea 1** relating agreement measures to the magnitude of the alteration of a gold standard [Mathet and al. 2012]

⇨ Indication of a correlation between agreement measure and data quality

⇨ **Limitation** – the interpretation of the magnitude remains unclear

**Idea 2** relating explicitly agreement measure and level of reproducibility

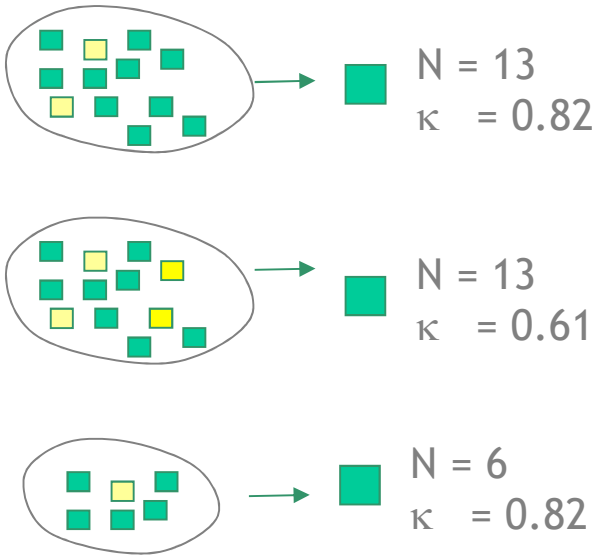

⇨ Agreement measure = probability to **reproduce** the annotation without any change (or conversely, to observe a change)

# Metrics interpretability

## TRUSTABILITY OF DATA RELIABILITY THRESHOLD

**Annotation** Majority vote with a number N of coders

**Hypothesis** Considering N, a given agreement measure A should give the probability that the annotation can be different with another set of N coders



# Metrics interpretability

## TRUSTABILITY OF DATA RELIABILITY THRESHOLD

**Aim** Table of reproducibility level

Statistical estimation of the probability that an annotation change can be observed with another coders set

| N \ agreement | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|
| 2 | 57% | 52% | 41% | 32% | 17% | 8% |
| 3 | 53% | 46% | 36% | 26% | 12% | 5% |
| 4 | 50% | 42% | 32% | 23% | 9% | 3% |
| 5 | 46% | 36% | 25% | 17% | 6% | 1% |

**Probability estimation** – % of changes observed on a representative data set

**Question** Experimental validation of the hypothesis : `P(N,Agreement)`

# Metrics reliability

## EXPERIMENTAL SETUP

4 tasks and 13 corpora to assess separately the influence of potential bias factors

### Emotion annotation

- 5 classes, from -2 (very negative) to 2 (very positive)
- Reduction to 3 classes : negative, neutral, positive
- 25 coders

### Opinion annotation

- 5 classes, from -2 (very negative) to 2 (very positive)
- Reduction to 3 classes : negative, neutral, positive
- 25 coders

### Semantic similarity annotation

- Similarity score between 0 (no similarity) and 4 (identical)
- Reduction to 5 or 3 classes
- 12 coders – 2 themes : *space conquest* and *pandemic*

### Coreference annotation

- 5 types of co-reference relations (direct, indirect,…)
- **Nominal** annotation (**no order** between classes)
- 9 coders

Emotion 5 | Emotion HC 5

Emotion 3 | Emotion HC 3

Opinion 5 | Opinion HC 5

Opinion 3 | Opinion HC 3

Space 5 | Pandemic 5

Space 3 | Pandemic 3

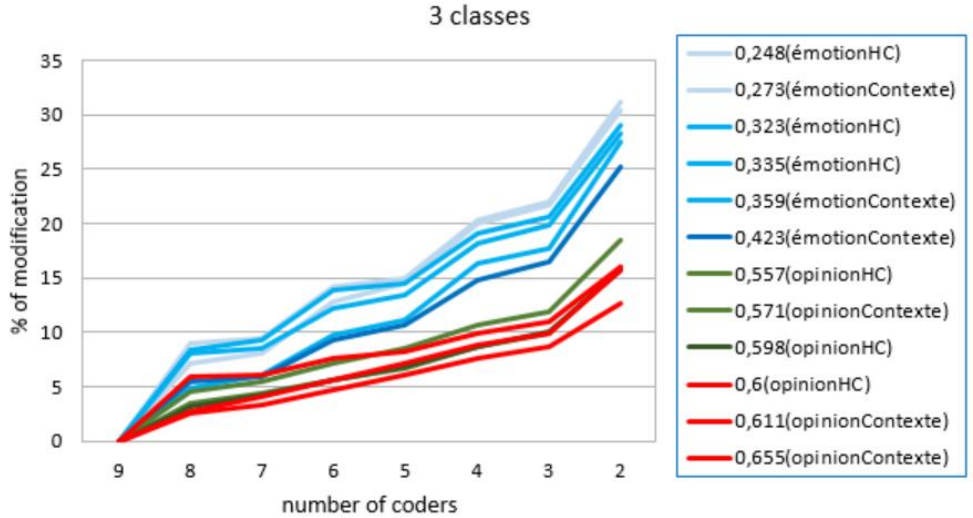
Coreference 5 classes

# Metrics interpretability

## VALIDATION ON REAL DATA

- **Real data :** opinion and emotion corpora restricted to a subset of 9 coders
- Computation of all possible annotations with N coders (N = 2 to 9) among 9
- Kappa (M-κ) metric

**Kappa (κ)**
[Cohen 1960]




## Conclusion

- Clear correlation between Kappa values and % of modification
- Clear influence of the number of coders
- Insufficient amount of data to reach an estimation for every Kappa value

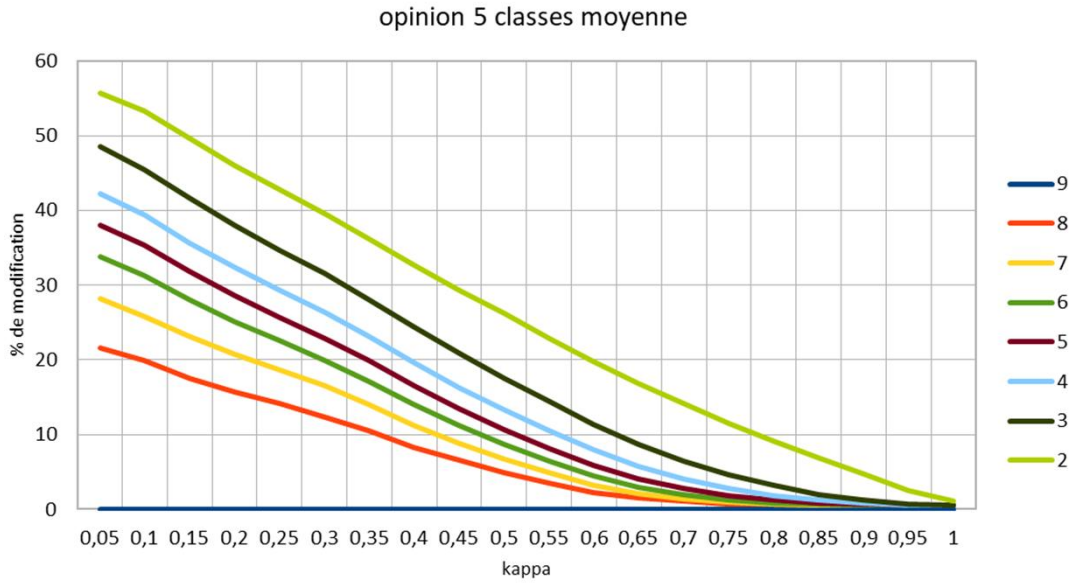
# Metrics interpretability

## VALIDATION ON ARTIFICIAL DATA

- **Artificial data** – random generation of annotations (shuffling) from real data
- **Data set of 2000 annotations** : scatter plot kappa value / % of modification


*If you conduct an annotation with 5 coding categories and with 8 coders, if you observe a Kappa value of 0.45, your annotation is reliable at a level of confidence of 92% (i.e. there is a 8% chance of observing a different annotation with 8 other coders*

# Metrics interpretability

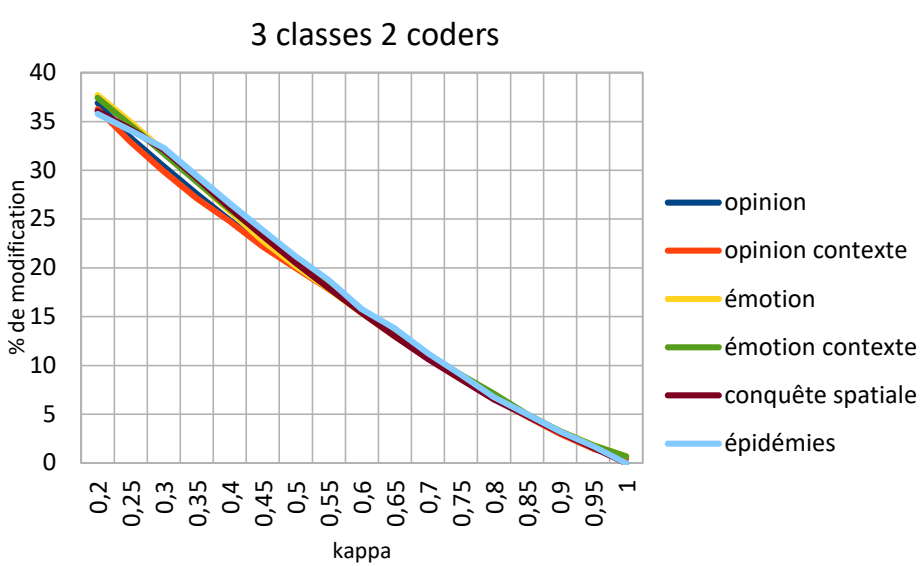



## Questions : potential biases

- Number of coding categories (classes) ?
- Influence of the task ?

# Metrics interpretability

## INFLUENCE OF THE NUMBER OF CLASSES





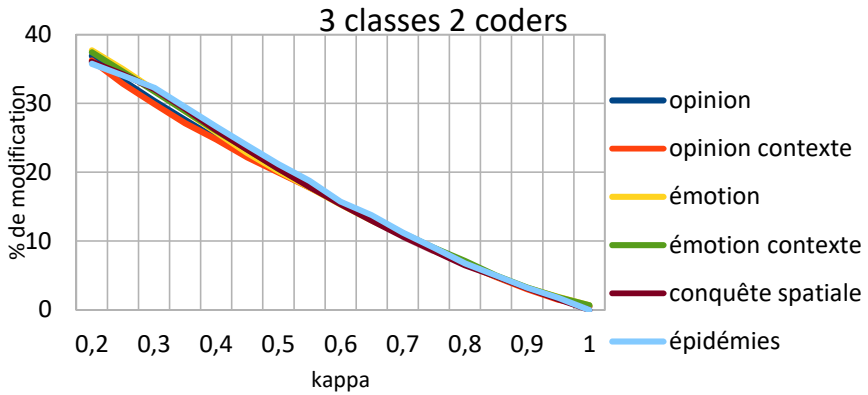
## Conclusion

- Similar behaviour of the curves
- Significant difference of magnitude

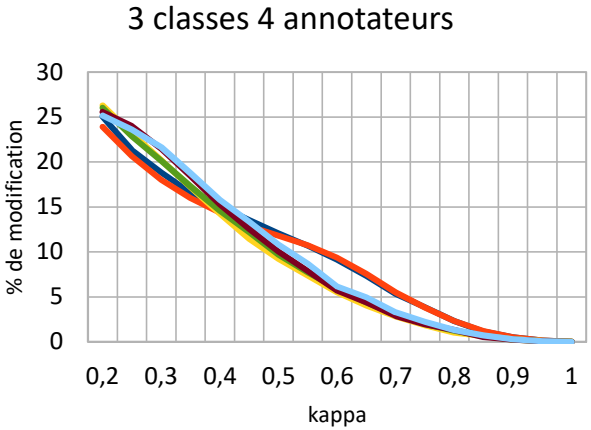
- **Recommendation** – Probabilities must be estimated for a given agreement level, a given number of coders **and a given number of coding categories**

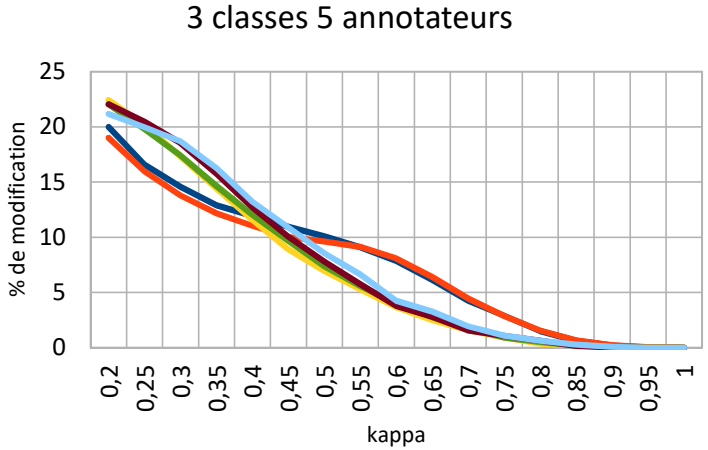
# Metrics interpretability

## INFLUENCE OF THE TASK







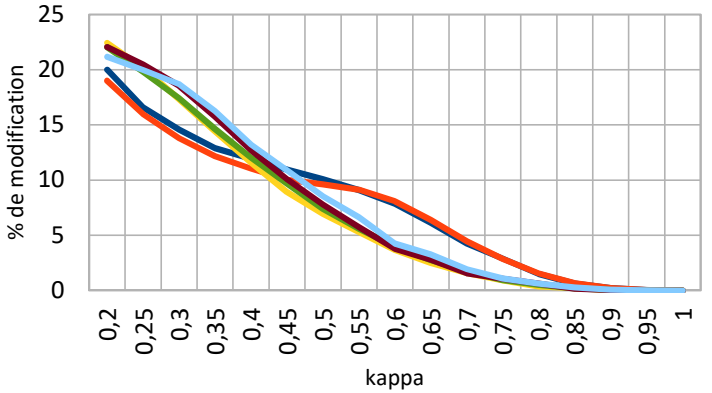

## Conclusion 1

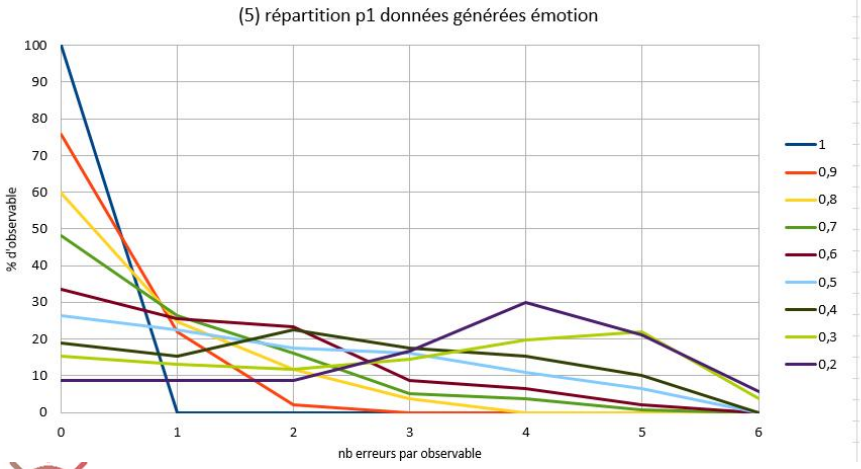
- Restricted but significative task influence

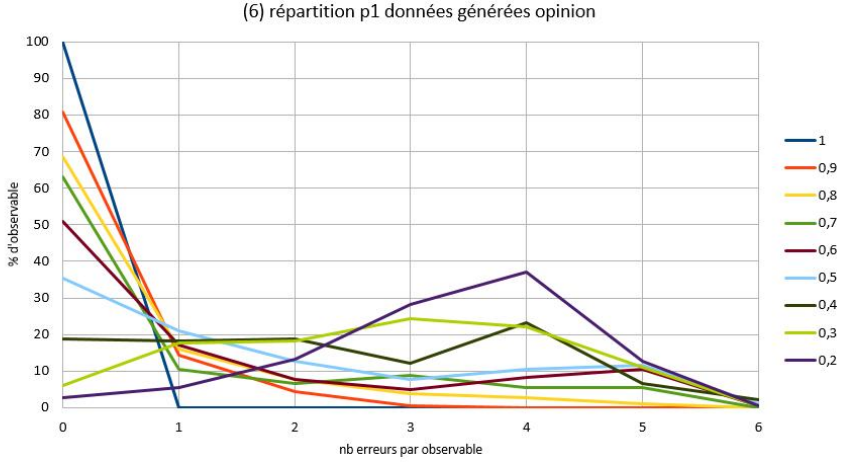
# Metrics interpretability

## INFLUENCE OF THE TASK

3 classes 5 annotateurs




## Conclusion 2

- Task influence : distribution of the errors among the data

# Conclusion

**It seems possible to interpret inter-coders agreement in terms of reproducibility level**


**Data reliability threshold must be specific to every agreement metric**

# Références


Antoine J.-Y., Villaneau J., Lefeuvre A. (2014) Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. Proc. *14th Conf. of the Europ. Chapter of the ACL, EACL'2014*, Gothenburg, Sweden

Carletta J. (1996). Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2):249-254

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

Krippendorff K. (2004). Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research,* 30(3): 411-433, 2004

Mathet Y., Widlöcher A., Fort K., François C., Galibert O., Grouin C., Kahn J., Rosset S., Zweigenbaum P. (2012) Manual Corpus Annotation: Evaluating the Evaluation Metrics. *Proc. COLING'2012*, Mumbaï, Inde.

Neuendorf K. (2002). The Content Analysis Guidebook. Sage Publications, Thousand Oaks, CA

Scott W. (1955) Reliability of content analysis: the case of nominal scale coding. *Public Opinions Quaterly*, 19:321-325