Corpus Linguistics: Inter-Annotator Agreements

Karën Fort

December 15, 2011

Sources

Most of this course is largely inspired by:

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

Introduction

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

- ML learns to make same mistakes as human annotator (noise $\neq$ patterns in errors [Reidsma and Carletta, 2008])
- Misleading evaluation
- Inconclusive and misleading results from linguistic analysis and hand-crafted systems

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Validity vs. Reliability [Artstein and Poesio, 2008]

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

Achieving Reliability (consistency)

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

In all cases, measure of reliability: coefficients of agreement

Particular Case: Gold-standard

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

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

Which one?

Particular Case: Gold-standard

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

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

F-measure

Precision/Recall: back to basics

- Recall:
- Silence:
- Precision:
- Noise:

Precision/Recall: back to basics

- Recall: measures the quantity of found annotations

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

- 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 ($\beta = 0.5$)

A little more from biology and medicine

True and false, positive and negative:

	Disease is present	Disease is absent
Positive test		
Negative test		

A little more from biology and medicine

True and false, positive and negative:

	Disease is present	Disease is absent
Positive test	TP	
Negative test		TN

A little more from biology and medicine

True and false, positive and negative:

	Disease is present	Disease is absent
Positive test	TP	FP
Negative test	FN	TN

A little more from biology and medicine

- sensitivity: corresponds to recall

$$SE = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ negatives}$$

- specificity: rate of true negatives

$$SP = \frac{true\ \ negatives}{true\ \ negatives\ +\ false\ \ positives}$$

- selectivity: corresponds to precision

$$SEL = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ positives}$$

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

$$ACC = \frac{true\ \ positives\ +\ true\ \ negatives}{TP + FP + FN + TN}$$

A little more from biology and medicine

- sensitivity: corresponds to recall

$$SE = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ negatives}$$

- specificity: rate of true negatives

$$SP = \frac{true\ \ negatives}{true\ \ negatives\ +\ false\ \ positives}$$

- selectivity: corresponds to precision

$$SEL = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ positives}$$

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

$$ACC = \frac{true\ \ positives\ +\ true\ \ negatives}{TP + FP + FN + TN}$$

A little more from biology and medicine

- sensitivity: corresponds to recall

$$SE = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ negatives}$$

- specificity: rate of true negatives

$$SP = \frac{true\ \ negatives}{true\ \ negatives\ +\ false\ \ positives}$$

- selectivity: corresponds to precision

$$SEL = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ positives}$$

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

$$ACC = \frac{true\ \ positives\ +\ true\ \ negatives}{TP + FP + FN + TN}$$

A little more from biology and medicine

- sensitivity: corresponds to recall

$$SE = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ negatives}$$

- specificity: rate of true negatives

$$SP = \frac{true\ \ negatives}{true\ \ negatives\ +\ false\ \ positives}$$

- selectivity: corresponds to precision

$$SEL = \frac{true\ \ positives}{true\ \ positives\ +\ false\ \ positives}$$

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

$$ACC = \frac{true\ \ positives\ +\ true\ \ negatives}{TP + FP + FN + TN}$$

Does a "Gold-standard" exist?

- reference rarely pre-exists
- can it be "perfect"? [Fort and Sagot, 2010]

→ can we use F-measure in other cases? Reading for next class!

⇒ Back to coefficients of agreement.

Easy and Hard Tasks

[Brants, 2000] for POS and Syntax, [Véronis, 2001] for WSD.

Objective tasks

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

Subjective tasks

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

Easy and Hard Tasks

[Brants, 2000] for POS and Syntax, [Véronis, 2001] for WSD.

Objective tasks

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

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

Subjective tasks

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

Easy and Hard Tasks

[Brants, 2000] for POS and Syntax, [Véronis, 2001] for WSD.

Objective tasks

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

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

Subjective tasks

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

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

Example

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.	✓	✗	✗

→ **Agreement**?

Contingency Table and Observed Agreement

		A		
		Yes	No	Total
B	Yes	**4**	2	6
	No	2	**2**	4
	Total	6	4	**10**

Observed Agreement (A_o)

proportion of items on which 2 annotators agree.

Here:

Contingency Table and Observed Agreement

		A		
		Yes	No	Total
B	Yes	**4**	2	6
	No	2	**2**	4
	Total	6	4	**10**

Observed Agreement (A_o)

proportion of items on which 2 annotators agree.

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

Chance Agreement

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

Chance Agreement

Some agreement is expected by **chance alone**:

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

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

Taking Chance into Account

Expected Agreement (A_e)

expected value of observed agreement.

Amount of agreement above chance: $A_o - A_e$
Maximum possible agreement above chance: $1 - A_e$

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

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

Expected Agreement

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

S [Bennett et al., 1954]

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

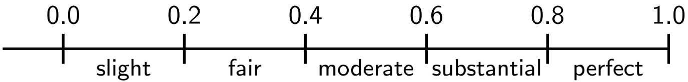
For any sample:

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

What is a "good" κ (or π or S)?

Scales for the interpretation of Kappa

- Landis and Koch, 1977



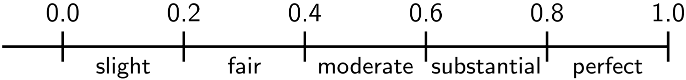
- Krippendorff, 1980

- Green, 1997

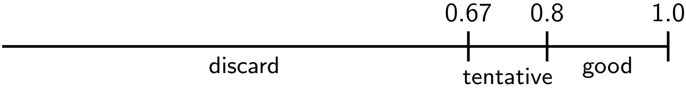
- "if a threshold needs to be set, 0.8 is a good value" [Artstein and Poesio, 2008]

Scales for the interpretation of Kappa

- Landis and Koch, 1977



- Krippendorff, 1980

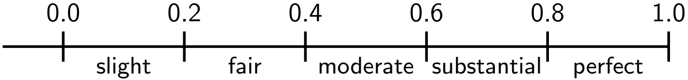


- Green, 1997

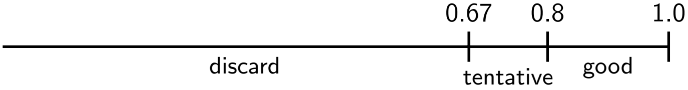
- "if a threshold needs to be set, 0.8 is a good value" [Artstein and Poesio, 2008]

Scales for the interpretation of Kappa

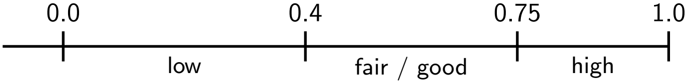
- Landis and Koch, 1977



- Krippendorff, 1980



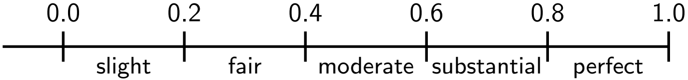
- Green, 1997



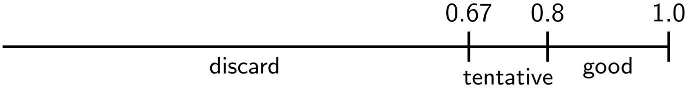
- "if a threshold needs to be set, 0.8 is a good value" [Artstein and Poesio, 2008]

Scales for the interpretation of Kappa

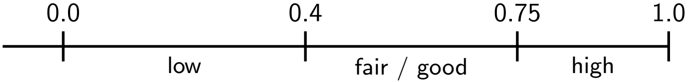
- Landis and Koch, 1977



- Krippendorff, 1980



- Green, 1997



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

More Annotators?

Differences among coders are diluted when more coders are used.

- With many coders, difference between π and κ is small
- Another argument for using many coders

More than two annotators

Multiple annotators

Agreement is the proportion of agreeing pairs

Item	Annot1	Annot2	Annot3	Annot4	**Pairs**
a	Boxcar	Tanker	Boxcar	Tanker	**2/6**
b	Tanker	Boxcar	Boxcar	Boxcar	**3/6**
c	Boxcar	Boxcar	Boxcar	Boxcar	**6/6**
d	Tanker	Engine2	Boxcar	Tanker	**1/6**
e	Engine2	Tanker	Boxcar	Engine1	**0/6**
f	Tanker	Tanker	Tanker	Tanker	**6/6**
g	Engine1	Engine1	Engine1	Engine1	**6/6**

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

K

Beware!

K is a generalization of π (not κ!)

Expected agreement

The probability of agreement for an arbitrary pair of coders.

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

Probability of two arbitrary annotators picking the same category:

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

Missing Points and Reflexions

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

There are ongoing reflexions on some issues, like:

- prevalence
- finding the "right" negative case (we'll see that in practical course)

- Precision, recall, F-measure
- Accuracy
- Observed agreement
- S, κ, π
- More than 2 annotators

- Read carefully: [Hripcsak and Rothschild, 2005] (http://ukpmc.ac.uk/articles/PMC1090460)
- Apply the grid we saw in the second course to this article.

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.

Brants, T. (2000).
Inter-annotator agreement for a german newspaper corpus.
In *In Proceedings of Second International Conference on Language Resources and Evaluation LREC-2000*.

Cohen, J. (1960).
A Coefficient of Agreement for Nominal Scales.
Educational and Psychological Measurement, 20(1):37–46.

Fort, K. and Sagot, B. (2010).
Influence of Pre-annotation on POS-tagged Corpus Development.

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

Hripcsak, G. and Rothschild, A. S. (2005).
Agreement, the f measure, and reliability in information retrieval.
J Am Med Inform Assoc., 12(3):296â8.

Krippendorff, K. (2004).
Content Analysis: An Introduction to Its Methodology, second edition, chapter 11.
Sage, Thousand Oaks, CA.

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.

Véronis, J. (2001).

Sense tagging: does it make sense?
In *Corpus Linguistics Conference*, Lancaster, Angleterre.