Agreement as a window to the process of corpus annotation

Ron Artstein

29 September 2012

The work depicted here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

1. Motivation
2. Agreement coefficients (Artstein & Poesio 2008, CL)
3. Usage cases
4. Conclusions

Why measure annotator agreement

Agreement can be measured between annotations of a single text.

Reliability measures consistency of an instrument.

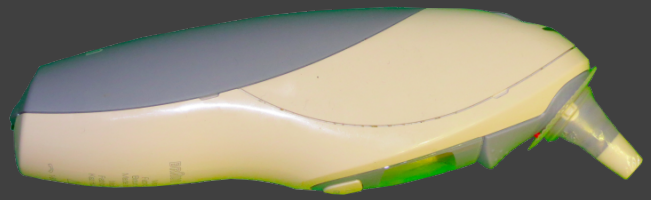
Validity is the correctness relative to a desired standard.

Reliability is a property of a process

Repeated measures with two thermometers

Mercury $\pm 0.1°C$

Infrared $\pm 0.4°C$



The mercury thermometer is more reliable.

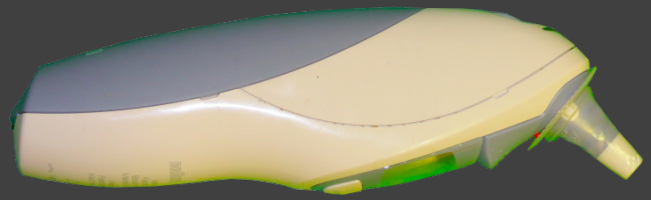
- But what if it's not calibrated properly?

Reliability is a property of a process

Repeated measures with two thermometers

Mercury $\pm 0.1^{\circ}C$

Infrared $\pm 0.4^{\circ}C$



The mercury thermometer is more reliable.

- But what if it's not calibrated properly?

Reliability is a **minimum requirement** for an annotation process.

- Qualitative evaluation also necessary.

Reliability and agreement

Reliability = **consistency** of annotation

- Needs to be measured on the same text.
- Different annotators.
- Work **independently**

If independent annotators mark a text the same way, then:

- They have internalized the same scheme (instructions).
- They will apply it consistently to new data.
- Annotations may be correct.

Results **do not generalize** from one domain to another.

1 Motivation

2 **Agreement coefficients** (Artstein & Poesio 2008, CL)

3 Usage cases

4 Conclusions

Observed agreement

Observed agreement: proportion of items on which 2 coders agree.

Detailed Listing

Item	Coder 1	Coder 2
a	Boxcar	Tanker
b	Tanker	Boxcar
c	Boxcar	Boxcar
d	Boxcar	Tanker
e	Tanker	Tanker
f	Tanker	Tanker
⋮	⋮	

Observed agreement

Observed agreement: proportion of items on which 2 coders agree.

Detailed Listing

Item	Coder 1	Coder 2
a	Boxcar	Tanker
b	Tanker	Boxcar
c	Boxcar	Boxcar
d	Boxcar	Tanker
e	Tanker	Tanker
f	Tanker	Tanker
⋮	⋮	⋮

Contingency Table

	Boxcar	Tanker	Total
Boxcar	41	3	44
Tanker	9	47	56
Total	50	50	100

Agreement: $\frac{41 + 47}{100} = 0.88$

High agreement, low reliability

Two psychiatrists evaluating 1000 patients.

	Normal	Paranoid	Total
Normal	990	5	995
Paranoid	5	0	5
Total	995	5	1000

- Observed agreement = 990/1000 = 0.99
- Most of these patients probably aren't paranoid
- No evidence that the psychiatrists identify the paranoid ones
- High agreement **does not indicate** high reliability

Chance agreement

Some agreement is expected by chance alone.

- Randomly assign two labels $\rightarrow$ agree half of the time.
- The amount expected by chance varies depending on the annotation scheme and on the annotated data.

Meaningful agreement is the agreement **above chance**.

Correction for chance

How much of the observed agreement is above chance?

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

Correction for chance

How much of the observed agreement is above chance?

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

Total			*Chance*			*Above*	
44	6		**6**	6		**38**	0
6	**44**	=	6	**6**	+	0	**38**
88			**12**			**76**	

Agreement: *88/100*
Due to chance: *12/100*
Above chance: *76/100*

Expected agreement

Observed agreement (A_o): proportion of actual agreement
Expected agreement (A_e): expected value of A_o

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

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

Scott's π, Fleiss's κ, Siegel and Castellan's K

Total number of judgments: $N = \sum_q n_q$

Probability of one coder picking category q: $\frac{n_q}{N}$

Prob. of two coders picking category q: $\left(\frac{n_q}{N}\right)^2$ *[biased estimator]*

Prob. of two coders picking same category: $A_e = \sum_q \left(\frac{n_q}{N}\right)^2$

Scott's π, Fleiss's κ, Siegel and Castellan's K

Total number of judgments: $N = \sum_q n_q$

Probability of one coder picking category q: $\frac{n_q}{N}$

Prob. of two coders picking category q: $\left(\frac{n_q}{N}\right)^2$ *[biased estimator]*

Prob. of two coders picking same category: $A_e = \sum_q \left(\frac{n_q}{N}\right)^2$

	Normal	*Paran*	*Total*
Normal	*990*	*5*	*995*
Paranoid	*5*	*0*	*5*
Total	*995*	*5*	*1000*

$$A_o = 0.99$$

$$A_e = .995^2 + .005^2 = 0.99005$$

$$K = \frac{0.99 - 0.99005}{1 - 0.99005} \approx -0.005$$

Multiple coders

Multiple coders: Agreement is the proportion of agreeing **pairs**

Item	Coder 1	Coder 2	Coder 3	Coder 4	Pairs
a	Boxcar	Tanker	Boxcar	Tanker	2/6
b	Tanker	Boxcar	Boxcar	Boxcar	3/6
c	Boxcar	Boxcar	Boxcar	Boxcar	6/6
d	Tanker	Engine 2	Boxcar	Tanker	1/6
e	Engine 2	Tanker	Boxcar	Engine 1	0/6
f	Tanker	Tanker	Tanker	Tanker	6/6
	⋮	⋮	⋮	⋮	

Expected agreement

- The probability of agreement for an **arbitrary pair** of coders

Krippendorff's α: weighted and generalized

Krippendorff's α:

- Weighted: various distance metrics
- Allows multiple coders
- Similar to K when categories are nominal
- Allows numerical category labels
  - Related to ANOVA (analysis of variance)

General formula for α

α is calculated using observed and expected **disagreement**:

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

- Disagreement can be in units outside the range $[0, 1]$
- Disagreements computed with various **distance metrics**

Analysis of variance

Numerical judgments (e.g. magnitude estimation)

- Single-variable ANOVA, each item = separate level

Analysis of variance

Numerical judgments (e.g. magnitude estimation)

- Single-variable ANOVA, each item = separate level

$$F = \frac{\text{between-level variance}}{\text{error variance}}$$

$F = 1$: Levels non-distinct
Random

$F > 1$: Levels distinct
Effect exists

Analysis of variance

Numerical judgments (e.g. magnitude estimation)

- Single-variable ANOVA, each item = separate level

$$F = \frac{\text{between-level variance}}{\text{error variance}}$$

$$\frac{\text{error variance}}{\text{total variance}}$$

$F = 1$: Levels non-distinct
Random

$F > 1$: Levels distinct
Effect exists

0: No error; perfect agreement
1: Random; no distinction
2: Maximal value

Analysis of variance

Numerical judgments (e.g. magnitude estimation)

- Single-variable ANOVA, each item = separate level

$$F = \frac{\text{between-level variance}}{\text{error variance}} \qquad \frac{\text{error variance}}{\text{total variance}}$$

$F = 1$: Levels non-distinct
Random

$F > 1$: Levels distinct
Effect exists

0: No error; perfect agreement
1: Random; no distinction
2: Maximal value

$$\alpha = 1 - \frac{\textit{error variance}}{\textit{total variance}} = 1 - \frac{D_o}{D_e}$$

Example of α

Item	C-1	C-2	C-3	C-4	C-5	Mean	Variance
(a)	7	7	7	7	7	7.0	0.0
(b)	5	4	5	6	5	5.0	0.5
(c)	5	5	5	6	4	5.0	0.5
(d)	7	8	6	7	7	7.0	0.5
(e)	4	2	3	3	2	2.8	0.7
(f)	6	7	6	6	6	6.2	0.2
(g)	6	6	6	5	6	5.8	0.2
(h)	7	6	9	6	9	7.4	2.3
(i)	5	5	5	4	5	4.8	0.2
(j)	4	5	2	4	6	4.2	2.2
(k)	3	5	2	4	4	3.6	1.3
(l)	5	5	6	6	5	5.4	0.3
(m)	3	4	2	3	3	3.0	0.5
(n)	2	3	4	3	4	3.2	0.7
(o)	7	7	6	7	7	6.8	0.2
(p)	7	8	7	8	7	7.4	0.3
(q)	3	3	3	1	3	2.6	0.8
(r)	4	2	4	2	4	3.2	1.2
(s)	3	2	3	3	3	2.8	0.2
(t)	4	4	2	4	4	3.6	0.8
(u)	5	6	4	5	6	5.2	0.7
(v)	4	3	4	3	1	3.0	1.5
(w)	6	6	7	5	7	6.2	0.7
(x)	4	5	2	4	3	3.6	1.3
(y)	4	5	5	6	5	5.0	0.5

Mean variance per item: **0.732**

Example of α

Item	C-1	C-2	C-3	C-4	C-5	Mean	Variance
(a)	7	7	7	7	7	7.0	0.0
(b)	5	4	5	6	5	5.0	0.5
(c)	5	5	5	6	4	5.0	0.5
(d)	7	8	6	7	7	7.0	0.5
(e)	4	2	3	3	2	2.8	0.7
(f)	6	7	6	6	6	6.2	0.2
(g)	6	6	6	5	6	5.8	0.2
(h)	7	6	9	6	9	7.4	2.3
(i)	5	5	5	4	5	4.8	0.2
(j)	4	5	2	4	6	4.2	2.2
(k)	3	5	2	4	4	3.6	1.3
(l)	5	5	6	6	5	5.4	0.3
(m)	3	4	2	3	3	3.0	0.5
(n)	2	3	4	3	4	3.2	0.7
(o)	7	7	6	7	7	6.8	0.2
(p)	7	8	7	8	7	7.4	0.3
(q)	3	3	3	1	3	2.6	0.8
(r)	4	2	4	2	4	3.2	1.2
(s)	3	2	3	3	3	2.8	0.2
(t)	4	4	2	4	4	3.6	0.8
(u)	5	6	4	5	6	5.2	0.7
(v)	4	3	4	3	1	3.0	1.5
(w)	6	6	7	5	7	6.2	0.7
(x)	4	5	2	4	3	3.6	1.3
(y)	4	5	5	6	5	5.0	0.5

Mean variance per item: **0.732**

Overall variance: **3.085**

'1' **2**	'2' **11**	'3' **19**	'4' **24**	'5' **23**
'6' **22**	'7' **19**	'8' **3**	'9' **2**	Mean 4.792

Example of α

Item	C-1	C-2	C-3	C-4	C-5	Mean	Variance
(a)	7	7	7	7	7	7.0	0.0
(b)	5	4	5	6	5	5.0	0.5
(c)	5	5	5	6	4	5.0	0.5
(d)	7	8	6	7	7	7.0	0.5
(e)	4	2	3	3	2	2.8	0.7
(f)	6	7	6	6	6	6.2	0.2
(g)	6	6	6	5	6	5.8	0.2
(h)	7	6	9	6	9	7.4	2.3
(i)	5	5	5	4	5	4.8	0.2
(j)	4	5	2	4	6	4.2	2.2
(k)	3	5	2	4	4	3.6	1.3
(l)	5	5	6	6	5	5.4	0.3
(m)	3	4	2	3	3	3.0	0.5
(n)	2	3	4	3	4	3.2	0.7
(o)	7	7	6	7	7	6.8	0.2
(p)	7	8	7	8	7	7.4	0.3
(q)	3	3	3	1	3	2.6	0.8
(r)	4	2	4	2	4	3.2	1.2
(s)	3	2	3	3	3	2.8	0.2
(t)	4	4	2	4	4	3.6	0.8
(u)	5	6	4	5	6	5.2	0.7
(v)	4	3	4	3	1	3.0	1.5
(w)	6	6	7	5	7	6.2	0.7
(x)	4	5	2	4	3	3.6	1.3
(y)	4	5	5	6	5	5.0	0.5

Mean variance per item: **0.732**

Overall variance: **3.085**

'1' **2** '2' **11** '3' **19** '4' **24** '5' **23**
'6' **22** '7' **19** '8' **3** '9' **2** Mean 4.792

$$\alpha = 1 - \frac{0.732}{3.085} = 0.763$$

$$F(24, 100) = \frac{12.891}{0.732} = 17.611, p < 1^{-15}$$

Distance metrics for α

Interval α (numeric values)

$$\mathbf{d}_{ab} = (a - b)^2$$

Nominal α (all disagreements equal)

$$\mathbf{d}_{ab} = \begin{cases} 0 \ if \ a = b \\ 1 \ if \ a \neq b \end{cases}$$

Nominal $\alpha \approx K$

Agreement measures are not hypothesis tests

- Evaluating magnitude, not existence/lack of effect
- Not comparing two hypotheses
- No clear probabilistic interpretation

Agreement values (historical note)

Krippendorff 1980, page 147:

> *In a study by Brouwer et al. (1969) we adopted the policy of reporting on variables only if their reliability was above .8 and admitted variables with reliability between .67 and .8 only for drawing highly tentative and cautious conclusions. These standards have been continued in work on cultural indicators (Gerbner et al., 1979) and might serve as a guideline elsewhere.*

Agreement values (historical note)

Krippendorff 1980, page 147:

> *In a study by Brouwer et al. (1969) we adopted the policy of reporting on variables only if their reliability was above .8 and admitted variables with reliability between .67 and .8 only for drawing highly tentative and cautious conclusions. These standards have been continued in work on cultural indicators (Gerbner et al., 1979) and might serve as a guideline elsewhere.*

Carletta 1996, page 252:

> *[Krippendorff] says that content analysis researchers generally think of $K > .8$ as good reliability, with $.67 < K < .8$ allowing tentative conclusions to be drawn.*

1 Motivation

2 Agreement coefficients (Artstein & Poesio 2008, CL)

3 Usage cases

4 Conclusions

Textbook usage paradigm

Conduct a reliability study with:

- Written annotation guidelines
- Generally available coders
- Representative sample of annotation materials

In order to validate annotation scheme and procedure.

Not all coders are equal

Scott, Barone and Koeling, LREC 2012

- Annotate hedges in medical text as likelihood

> **Possible** early pneumonia. . .
> . . . **could** represent pneumonia. . .

- Two annotator populations differ in **medical training**
- Systematic differences between annotators: medically trained interpret hedges as expressing greater likelihood

Each population of coders (instrument) has a certain reliability, but one is probably more correct.

Differences among coders

Coders agree to different extents (Artstein et al. 2009, LNCS)

	All Raters	Excluding Outlier	Range
Oct. 2007	0.786	0.886	0.676–0.901
June 2008	0.583	0.655	0.351–0.680
Oct. 2008	0.699	0.757	0.614–0.763

- 3 datasets, 4 coders each.
- Conf. intervals generalize over items (Hayes & Krippendorff).
- No generalization available over coders.

Learning from annotators' disagreements

Utterances ⇒ dialogue acts (Artstein et al. 2009, Semdial)

How well do the dialogue acts capture what users say?

- Virtual character.
- 16 dialogues.
- 224 unique user utterances.
- 3 annotators.

Instructions:

Match each user utterance to the most appropriate player speech act; if none is appropriate, match to "unknown".

Example annotations

Are you a school teacher?

3 **ynq** amani / work / teacher

Thank you and good night.

1 **thanks** 2 **closing**

Can you tell me about the sniper?

1 **whq** 1 **ynq** 1 **unknown**

Reliability of annotating dialogue acts

$\alpha = 1 - \frac{D_o}{D_e}$	Krippendorff's Alpha	Observed Disagreement	Expected Disagreement
Dialogue act	0.489	0.455	0.891
Dialogue act type	0.502	0.415	0.834
In domain?	0.383	0.259	0.420

- Reliability measures **straightforwardness** of the task.
- Improved with more explicit guidelines.
- Substantial disagreement on whether utterance fits scheme.

Adequacy of dialogue acts

Calculated after an individual analysis of the disagreements.

User utterances	N	%	
Fully covered	72	32	80%
Immaterial disagreement	57	25	
Covered with extensions	50	22	
Hard to deal with	45	20	
Total	224	100	

Follow-up study found coverage to be 72–76%.

Reliability of different parts of the data

Coherence of virtual character (Artstein et al. 2009, LNCS)

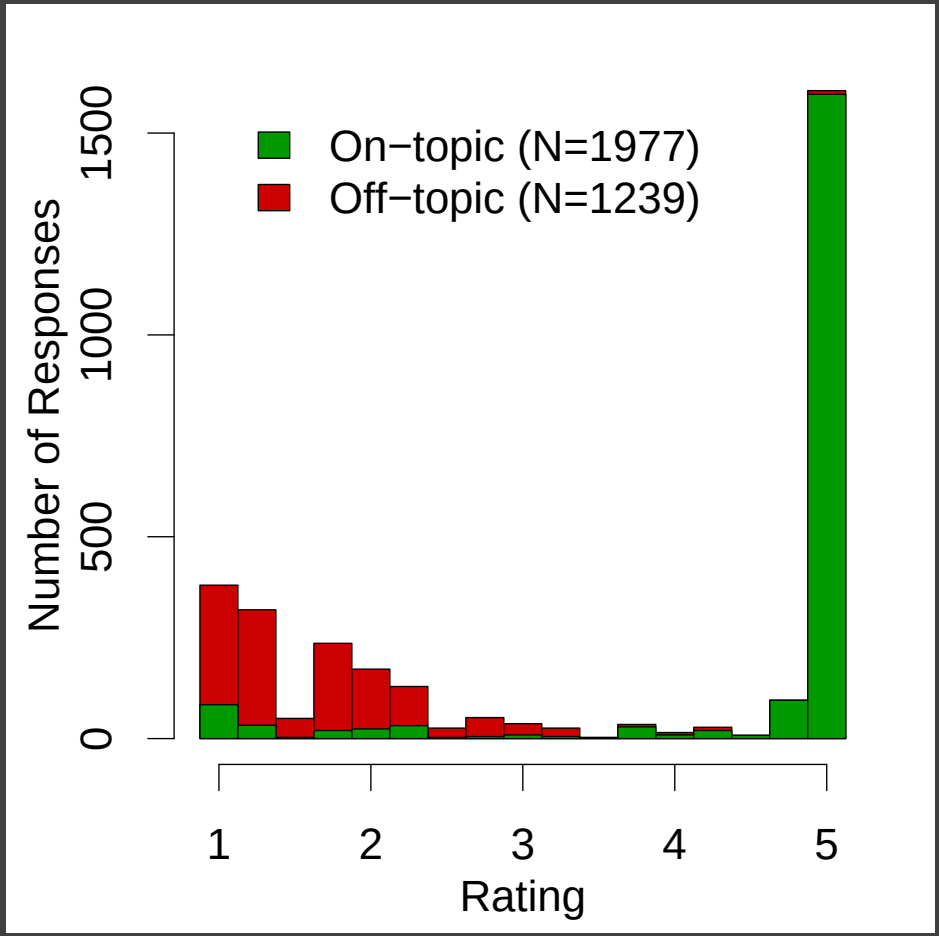
user	you wanna talk about the army	
sgt star	Hooah! I love talking about the Army!	★★★★★
user	hey sergeant not sure everyone here knows about hooah	
sgt star	It's the universal Soldier's word. everyone knows what hooah means.	★★★★☆
user	well what is hooah	
sgt star	Well, I think that's just going to have to remain a mystery.	★★★★☆
user	there could be a couple people who don't know so go ahead and cure their curiosity	
sgt star	Hooah	★☆☆☆☆
user	well explain hooah	
sgt star	Hooah can mean I copy, roger, good, alright, message received, yes, you got it, amen.	★★★★★

3216 responses: 703 exact match to training data
2513 rated by 4 judges on a scale of 1–5

Reliability of coherence ratings

- Distribution of ratings:

- Krippendorff's α
  Overall: 0.786
  On-topic: 0.794
  Off-topic: 0.097



Differences in the annotated material

Kang et al. 2012, AAMAS: identify smiles in videos

- Smiles are easier to detect on some people than others

Differences in the annotated material

Park et al. 2012, CrowdMM: identify nonverbal behavior in videos

- In-house experts
- Amazon Mechanical Turkers less reliable
- Majority vote among Turkers: only one instrument available
- Majority instrument vs. in-house: same reliability

1. Motivation
2. Agreement coefficients (Artstein & Poesio 2008, CL)
3. Usage cases
4. Conclusions

Conclusions

Reasons to conduct agreement studies:

- Validate annotation schemes and guidelines.
- Learn about how annotators work.
- Identify patterns in the underlying data.
- Point out directions for qualitative studies.

Results need to be interpreted.