

Statistical Models of the Annotation Process

Bob Carpenter¹ Massimo Poesio²

¹Alias-I

²Università di Trento

LREC 2010 Tutorial
17th May 2010

Many slides due to Ron Artstein

Annotated corpora

Annotated corpora are needed for:

- Supervised learning – training and evaluation
- Unsupervised learning – evaluation
- Hand-crafted systems – evaluation
- Analysis of text

Analysis of annotation quality

Typical approach to the analysis of the annotated data in CL

- Perform quality control requiring annotations to be **reliable**

Analysis of annotation quality

Typical approach to the analysis of the annotated data in CL

- Perform quality control requiring annotations to be **reliable**

If coding scheme determined to be reliable, then corpus annotated according to scheme to produce **gold standard**

- Sometimes each item annotated by a single coder, with random checks
- Better: each item annotated by two or more coders, followed by reconciliation (e.g., OntoNotes, Hovy et al 2006)

Reliability and agreement

Reliability = **consistency**. If independent annotators mark a text the same way,

- they have internalized the same scheme (instructions)
- will apply it consistently to new data
- annotations might be correct

Reliability and agreement

Reliability = **consistency**. If independent annotators mark a text the same way,

- they have internalized the same scheme (instructions)
- will apply it consistently to new data
- annotations might be correct

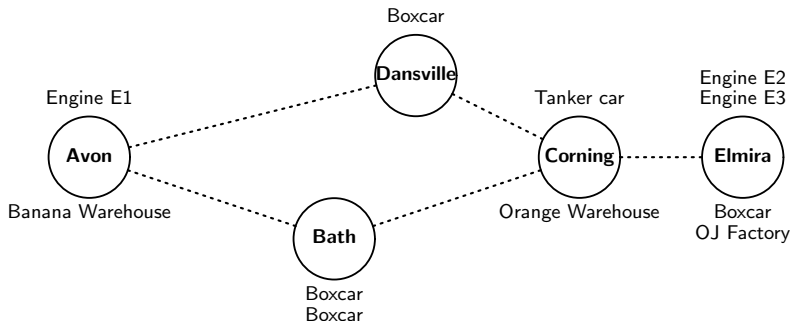
Popular measures of reliability: **coefficients of agreement**

Beyond measuring agreement

- Measuring agreement between annotators only part of the story. More systematic studies using (much) more than two annotators (Poesio and Arstein, 2005) already raised issues such as
 - Identifying poor quality annotators
 - Identifying difficult items

An example of multiple annotator study

- 18 naïve coders
- Dialogue 3.2 from the TRAINS 91 corpus
- MMAX annotation tool
- Map of the “TRAINS world”



Results

19.10: we need to get the bananas to Corning by 3

19.11: uh

19.12: maybe it's gonna be faster if we

19.13: send E1

19.14: E1's boxcar picks up at Dansville

19.15: instead of going back to Avon

19.16: have it go on to Corning

Results

19.10: we need to get the bananas to Corning by 3

19.11: uh

19.12: maybe it 's gonna be faster if we

19.13: send E1

19.14: E1 's boxcar picks up at Dansville

19.15: instead of going back to Avon

19.16: have it go on to Corning

Key: Full agreement

Results

19.10: we need to get the bananas to Corning by 3

19.11: uh

19.12: maybe it's gonna be faster if we

19.13: send E1

19.14: E1's boxcar picks up at Dansville

19.15: instead of going back to Avon

19.16: have it go on to Corning

Key: Full agreement One outlier

Results

19.10: we need to get the bananas to Corning by 3

19.11: uh

19.12: maybe it's gonna be faster if we

19.13: send E1

19.14: E1's boxcar picks up at Dansville

19.15: instead of going back to Avon

19.16: have it go on to Corning

Key: Full agreement One outlier Implicit

Results

19.10: we need to get the bananas to Corning by 3

19.11: uh

19.12: maybe it's gonna be faster if we

19.13: send E1

19.14: E1's boxcar picks up at Dansville

19.15: instead of going back to Avon

19.16: have it go on to Corning

Key: Full agreement One outlier Implicit Explicit

Beyond measuring agreement

- Measuring agreement between annotators only part of the story. More systematic studies using (much) more than two annotators (Poesio and Arstein, 2005) already raised issues such as
 - Identifying poor quality annotators
 - Identifying difficult items
- Recent trends towards using **crowdsourcing** (Mechanical Turk, Games with a Purpose) makes addressing these issues even more urgent, as well as raising others
 - Build gold standard out of multiple annotations
- More recent methods, in particular Bayesian methods, make this kind of analysis possible

This tutorial

- 1 We begin with a summary of the standard approaches to measuring agreement, following the discussion in Artstein and Poesio, CL 2008
 - Need for chance-corrected coefficients of agreement
 - Basic coefficients of agreement: Kappa, α
 - Problem and limitations of coefficients of agreement (e.g., interpreting the value)

This tutorial

- 1 We begin with a summary of the standard approaches to measuring agreement, following the discussion in Artstein and Poesio, CL 2008
 - Need for chance-corrected coefficients of agreement
 - Basic coefficients of agreement: Kappa, α
 - Problem and limitations of coefficients of agreement (e.g., interpreting the value)
- 2 Discuss problems and opportunities originated by the crowdsourcing approach to annotation

This tutorial

- 1 We begin with a summary of the standard approaches to measuring agreement, following the discussion in Artstein and Poesio, CL 2008
 - Need for chance-corrected coefficients of agreement
 - Basic coefficients of agreement: Kappa, α
 - Problem and limitations of coefficients of agreement (e.g., interpreting the value)
- 2 Discuss problems and opportunities originated by the crowdsourcing approach to annotation
- 3 How these issues can be tackled by the Bayesian approach to annotation analysis

Reliability studies

Measure the agreement between coders on a sample of the data in terms of **coefficients of agreement**

Reliability data

- Sample of the corpus
- Multiple annotators

Annotators must work **independently**

- Otherwise we can't compare them

Coefficients of agreement

Agreement measures are not hypothesis tests

- Evaluating magnitude, not existence/lack of effect
- Not comparing two hypotheses

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

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

Chance agreement

Some agreement is expected by chance alone.

- Two coders randomly assigning “Boxcar” and “Tanker” labels will 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**.

- Similar to the concept of “baseline” for system evaluation.

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

Expected agreement

Big question: how to calculate the amount of agreement expected by chance (A_e)?

S: same chance for all coders and categories

Number of category labels: q

Probability of one coder picking a particular category q_a : $\frac{1}{q}$

Probability of both coders picking a particular category q_a : $\left(\frac{1}{q}\right)^2$

Probability of both coders picking the same category:

$$A_e^S = q \cdot \left(\frac{1}{q}\right)^2 = \frac{1}{q}$$

Are all categories equally likely?

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

$$A_o = 0.88$$

$$A_e = \frac{1}{2} = 0.5$$

$$S = \frac{0.88 - 0.5}{1 - 0.5} = 0.76$$

Are all categories equally likely?

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

	A	B	C	D	Total
A	44	6	0	0	50
B	6	44	0	0	50
C	0	0	0	0	0
D	0	0	0	0	0
Total	50	50	0	0	100

$$A_o = 0.88$$

$$A_e = \frac{1}{2} = 0.5$$

$$S = \frac{0.88 - 0.5}{1 - 0.5} = 0.76$$

$$A_o = 0.88$$

$$A_e = \frac{1}{4} = 0.25$$

$$S = \frac{0.88 - 0.25}{1 - 0.25} = 0.84$$

π : different chance for different categories

Total number of judgments: **N**

Probability of one coder picking a particular category q_a : $\frac{n_{q_a}}{N}$

Probability of both coders picking a particular category q_a : $\left(\frac{n_{q_a}}{N}\right)^2$

Probability of both coders picking the same category:

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

Comparison of S and π

	A	B	C	Total
A	44	6	0	50
B	6	44	0	50
C	0	0	0	0
Total	50	50	0	100

$$A_o = 0.88$$

$$S = \frac{0.88 - 1/3}{1 - 1/3} = 0.82$$

$$\pi = \frac{0.88 - 0.5}{1 - 0.5} = 0.76$$

Comparison of S and π

	A	B	C	Total
A	44	6	0	50
B	6	44	0	50
C	0	0	0	0
Total	50	50	0	100

	A	B	C	Total
A	77	1	2	80
B	1	6	3	10
C	2	3	5	10
Total	80	10	10	100

$$A_o = 0.88$$

$$S = \frac{0.88 - 1/3}{1 - 1/3} = 0.82$$

$$\pi = \frac{0.88 - 0.5}{1 - 0.5} = 0.76$$

$$A_o = 0.88$$

$$S = \frac{0.88 - 1/3}{1 - 1/3} = 0.82$$

$$\pi = \frac{0.88 - 0.66}{1 - 0.66} \approx 0.65$$

Comparison of S and π

	A	B	C	Total
A	44	6	0	50
B	6	44	0	50
C	0	0	0	0
Total	50	50	0	100

	A	B	C	Total
A	77	1	2	80
B	1	6	3	10
C	2	3	5	10
Total	80	10	10	100

$$A_o = 0.88$$

$$S = \frac{0.88 - 1/3}{1 - 1/3} = 0.82$$

$$\pi = \frac{0.88 - 0.5}{1 - 0.5} = 0.76$$

$$A_o = 0.88$$

$$S = \frac{0.88 - 1/3}{1 - 1/3} = 0.82$$

$$\pi = \frac{0.88 - 0.66}{1 - 0.66} \approx 0.65$$

We can prove that for any sample:

$$A_e^\pi \geq A_e^S \quad \pi \leq S$$

Prevalence

Is the following annotation reliable?

Two annotators disambiguate 1000 instances of the word **love**:

- emotion
- zero (as in tennis)

Each annotator found:

- 995 instances of 'emotion'
- 5 instances of 'zero'

The annotators marked **different** instances of 'zero'. **Agr: 99%!**

Prevalence

Is the following annotation reliable?

Two annotators disambiguate 1000 instances of the word **love**:

- emotion
- zero (as in tennis)

Each annotator found:

- 995 instances of 'emotion'
- 5 instances of 'zero'

The annotators marked **different** instances of 'zero'. **Agr: 99%**

	<i>emotion</i>	<i>zero</i>	<i>Total</i>	
<i>emotion</i>	990	5	995	$A_o = 0.99$
<i>zero</i>	5	0	5	$S = \frac{0.99 - .5}{1 - .5} = 0.98$
<i>Total</i>	995	5	1000	$\pi = \frac{0.99 - 0.99005}{1 - 0.99005} \approx -0.005$

Prevalence

When one category is dominant:

- High agreement **does not indicate** high reliability
- π measures agreement on the rare category

Therefore, π is a good indicator of reliability.

κ : taking individual bias into account

Different annotators have different interpretations of the instructions (bias/prejudice). Does this affect expected agreement?

Total number of items: i

Probability of coder c_x picking a particular category q_a : $\frac{n_{c_x q_a}}{i}$

Probability of both coders picking category q_a : $\frac{n_{c_1 q_a}}{i} \cdot \frac{n_{c_2 q_a}}{i}$

Probability of both coders picking the same category:

$$A_e^\kappa = \sum_q \frac{n_{c_1 q}}{i} \cdot \frac{n_{c_2 q}}{i} = \frac{1}{i^2} \sum_q n_{c_1 q} n_{c_2 q}$$

π VS. κ

We can prove that for any sample:

$$A_e^\pi \geq A_e^\kappa \quad \pi \leq \kappa$$

π VS. κ

We can prove that for any sample:

$$A_e^\pi \geq A_e^\kappa \quad \pi \leq \kappa$$

Different interpretations of the instructions = lack of reliability.

- By this argument, π preferable to κ

π VS. κ

We can prove that for any sample:

$$A_e^\pi \geq A_e^\kappa \quad \pi \leq \kappa$$

Different interpretations of the instructions = lack of reliability.

- By this argument, π preferable to κ

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

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
g	Engine 1	Engine 1	Engine 1	Engine 1	6/6
	⋮	⋮	⋮	⋮	

Multiple coders

Numerical interpretation

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

Graphical representation

- Contingency table requires multiple dimensions. . .

Expected agreement

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

K: multiple coders

Confusing terminology: K is a generalization of π .

Total number of judgments: N

Probability of arbitrary coder picking a particular category q_a : $\frac{n_{q_a}}{N}$

Probability of two coders picking a particular category q_a : $\left(\frac{n_{q_a}}{N}\right)^2$

Probability of two arbitrary coders picking the same category:

$$A_e^K = \sum_q \left(\frac{n_q}{N}\right)^2 = \frac{1}{N^2} \sum_q n_q^2$$

Multiple coders – example

Item	Cod-1	Cod-2	Cod-3	Cod-4	Pairs
(a)	Box	Box	Box	Box	6/6
(b)	Box	Box	Box	Box	6/6
(c)	E-2	E-2	E-2	E-2	6/6
(d)	Tank	Tank	Tank	Tank	6/6
(e)	E-1	E-1	E-1	E-1	6/6
(f)	E-1	Box	E-1	E-1	3/6
(g)	Tank	Tank	Tank	Tank	6/6
(h)	Box	Box	Box	Box	6/6
(i)	Box	Box	Box	Box	6/6
(j)	Box	Box	E-1	Box	3/6
(k)	E-2	E-2	E-2	E-2	6/6
(l)	Box	Tank	Box	Box	3/6
(m)	E-1	E-1	E-1	E-1	6/6
(n)	Tank	Tank	Tank	Tank	6/6
(o)	E-1	E-1	E-1	E-1	6/6
(p)	E-2	E-2	E-2	Tank	3/6
(q)	Box	Box	Box	Box	6/6
(r)	Box	Box	Box	Box	6/6
(s)	E-1	E-1	Tank	E-1	3/6
(t)	Box	Box	Box	Box	6/6
(u)	Box	Box	Box	Box	6/6
(v)	E-1	E-1	E-1	E-1	6/6
(w)	Tank	Tank	Tank	Tank	6/6
(x)	Box	Box	Box	Box	6/6
(y)	Box	Box	Box	Tank	3/6

25 items, 100 judgments:

Box **46**, Tank **20**, E-1 **23**, E-2 **11**.

Observed agreement:

$$A_o = 132/150 = 0.88$$

Expected agreement:

$$A_e = .46^2 + .2^2 + .23^2 + .11^2 = 0.3166$$

$$K = \frac{0.88 - 0.3166}{1 - 0.3166} \approx 0.8244$$

Are all disagreements the same?

Some disagreements are more important than others

- **Boxcar/engine** more serious than **engine 1/engine 2**
- Depends on application

Need to count and weigh the disagreements

- Not only agreeing pairs
- Principled method of assigning weights

Agreement and disagreement

Observed disagreement: $D_o = 1 - A_o$

Expected disagreement: $D_e = 1 - A_e$

*Chance-corrected **disagreement:***

$$1 - \frac{D_o}{D_e}$$

Agreement and disagreement

Observed disagreement: $D_o = 1 - A_o$

Expected disagreement: $D_e = 1 - A_e$

*Chance-corrected **disagreement:***

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

Weights

Three labels: Boxcar, Engine 1, Engine 2.

Three weights:

Identical judgments: disagreement = 0 (agreement = 1)

Engine 1 / engine 2: disagreement = 0.5 (agreement = 0.5)

Boxcar / engine: disagreement = 1 (agreement = 0)

Weight table:

	<i>Box</i>	<i>E-1</i>	<i>E-2</i>
<i>Box</i>	0	1	1
<i>E-1</i>	1	0	0.5
<i>E-2</i>	1	0.5	0

Krippendorff's α : a generalized weighted coefficient

Krippendorff's α :

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

α with different distance metrics

General formula for α

$$\alpha = 1 - \frac{\text{error variance}}{\text{total variance}} = 1 - \frac{\text{mean item distance}}{\text{mean overall distance}} = 1 - \frac{D_o}{D_e}$$

Observed and expected disagreements computed with various **distance metrics**

Distance metrics for α

Interval α (numeric values)

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

Nominal α (all disagreements equal)

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

Nominal $\alpha \approx K$

Computing α : observed disagreement

Number of coders: c

Number of items: i

Distance of a single pair of labels q_a, q_b : $d_{q_a q_b}$

Observed disagreement

Number of judgment pairs per item: $c(c - 1)$

Mean distance within item i : $\frac{1}{c(c - 1)} \sum_{q_a} \sum_{q_b} n_{i q_a} n_{i q_b} d_{q_a q_b}$

Mean distance within items: $D_o = \frac{1}{i c(c - 1)} \sum_i \sum_{q_a} \sum_{q_b} n_{i q_a} n_{i q_b} d_{q_a q_b}$

Computing α : expected disagreement

Number of coders: c

Number of items: i

Distance of a single pair of labels q_a, q_b : $d_{q_a q_b}$

Expected disagreement:

Total number of judgment pairs:

$$ic(ic - 1)$$

Overall mean distance:

$$D_e = \frac{1}{ic(ic - 1)} \sum_{q_a} \sum_{q_b} n_{q_a} n_{q_b} d_{q_a q_b}$$

Example of use of α : Anaphoric chains

3.1 M: and while it's there it should pick up **the tanker₁**

...

15.1 M: we're picking up **the tanker₂**

15.2 uh **it₃** needs to then go back to Elmira

15.3 err excuse me

15.4 yes

15.5 ok

15.6 **it₄** needs to go back to Elmira

Example of use of α : Anaphoric chains

3.1 M: and while it's there it should pick up **the tanker₁**

...

15.1 M: we're picking up **the tanker₂**

15.2 uh **it₃** needs to then go back to Elmira

15.3 err excuse me

15.4 yes

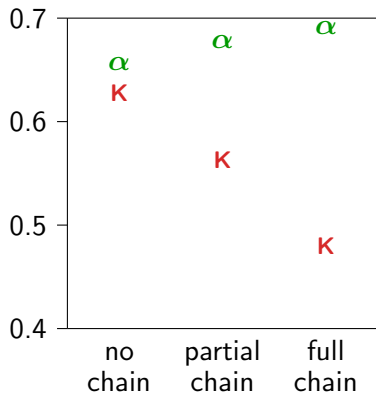
15.5 ok

15.6 **it₄** needs to go back to Elmira

No chain	Immediate antecedents	it₃ \mapsto {2}
Partial chain	Preceding items only	it₃ \mapsto {1, 2}
Full chain	Whole dialogue	it₃ \mapsto {1, 2, 3, 4}

K decreases with chain size, α increases

Chain	K	α
None	0.628	0.656
Partial	0.563	0.677
Full	0.480	0.691



Distance metrics

Jaccard: $d_{AB} = 1 - \frac{|A \cap B|}{|A \cup B|}$

Dice: $d_{AB} = 1 - \frac{2|A \cap B|}{|A| + |B|}$

Passonneau: $d_{AB} = \begin{cases} 0 & \text{if } A = B \\ 1/3 & \text{if } A \subset B \text{ or } B \subset A \\ 2/3 & \text{if } A \cap B \neq \emptyset, \text{ but } A \not\subset B \text{ and } B \not\subset A \\ 1 & \text{if } A \cap B = \emptyset \end{cases}$

Distance metrics

Jaccard: $d_{AB} = 1 - \frac{|A \cap B|}{|A \cup B|}$

Dice: $d_{AB} = 1 - \frac{2|A \cap B|}{|A| + |B|}$

Passonneau: $d_{AB} = \begin{cases} 0 & \text{if } A = B \\ 1/3 & \text{if } A \subset B \text{ or } B \subset A \\ 2/3 & \text{if } A \cap B \neq \emptyset, \text{ but } A \not\subset B \text{ and } B \not\subset A \\ 1 & \text{if } A \cap B = \emptyset \end{cases}$

	Jaccard	Dice	Passonneau
No chain	0.649	0.656	0.656
Partial	0.651	0.677	0.672
Full	0.642	0.691	0.654

α vs K : A summary

For nominal agree/disagree distinctions, $K \approx \alpha$

- Use either coefficient

For grades of agreement, use α

- Take care with choosing the distance metric

Interpreting agreement

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.

Interpreting the values is hard

- The CL literature is full of discussions on whether the standard proposed by Krippendorff and adopted by Carletta is appropriate for all types of annotation
- In annotations at the discourse level and beyond, it is often difficult to reach the .8 threshold
 - Poesio and Vieira 1998 for coreference
 - Craggs and McGee Wood 2005, Cavicchio and Poesio 2008 for emotions
- Many researchers have proposed to adopt the more lenient interpretation framework of Landis and Koch
- The problem becomes much more serious with weighted coefficients

Discourse antecedents (Artstein & Poesio 2006)

- 3.3–3.5 : ... again get the boxcar and engine to Corning
 3.6 : so the fastest way to do that is from Elmira
 3.7 : so we'll do that
 ⋮
 7.3 : so we ship one
 7.4 : boxcar
 7.5 : of oranges to Elmira
 7.6 : and **that** takes another 2 hours

that → 1 3 2 3 1

Discourse antecedents as sets of words

Item's "label" = set of words of the antecedent.

	Jaccard	Dice	Passonneau
D_o	0.53	0.43	0.43
D_e	0.95	0.94	0.94
α	0.45	0.55	0.55

Discourse antecedents as sets of words

Item's "label" = set of words of the antecedent.

	Jaccard	Dice	Passonneau
D_o	0.53	0.43	0.43
D_e	0.95	0.94	0.94
α	0.45	0.55	0.55

Antecedents of different expressions rarely coincide.

$$D_e \approx 1 \quad \textit{therefore} \quad \alpha \approx 1 - D_o$$

Discourse antecedents as word positions

Item's "label" = position of the beginning/end of the antecedent.

	beginnings	ends
D_o	54331	21237
D_e	2.82×10^7	2.88×10^7
α	0.998	0.999

Discourse antecedents as word positions

Item's "label" = position of the beginning/end of the antecedent.

	beginnings	ends
D_o	54331	21237
D_e	2.82×10^7	2.88×10^7
α	0.998	0.999

Antecedents of the same expression are close to each other, compared to the size of the whole dialogue.

$$D_o \ll D_e \quad \textit{therefore} \quad \alpha \approx 1$$

Discourse antecedents as relative positions

“label” = **relative** position of the beginning/end of the antecedent.

	beginnings	ends
D_o	54331	21237
D_e	65257	24200
α	0.167	0.122

Discourse antecedents as relative positions

“label” = **relative** position of the beginning/end of the antecedent.

	beginnings	ends
D_o	54331	21237
D_e	65257	24200
α	0.167	0.122

Variance among antecedents of each referent not much smaller than overall variance.

$$D_o \approx D_e \quad \textit{therefore} \quad \alpha \approx 0$$

Agreement and error

Agreement metrics are difficult to understand.

One way to get an interpretation is by relating the amount of agreement to an **error rate**. This can be done by making some assumptions (see Artstein and Poesio 2008)

- E.g., items are either **easy** or **hard**
- Coders always agree on easy items
- Coders classify hard items at random

Software to support calculation of coefficients of agreement

- A useful site covering what's available:
 - <http://astro.temple.edu/~lombard/reliability/>
- Various scripts for computing K exist, including our own implementation based on Mark Core's code
- The Alpha Resources page contains our own scripts for computing Alpha as well as Thomas Lippincott's and Becky Passonneau's scripts
 - <http://cswww.essex.ac.uk/Research/nle/arrau/alpha.html>

Summary of the first part of the tutorial

- Coefficients of agreement have proven rather useful tools in developing coding schemes
- But can be difficult to interpret, particularly when working with schemes that need weighting
- And do not help in analyzing the behavior of coders when many are used