# 7 - Kappa

#### Inter-Rater Reliability

- Dialogue Act Classification
  - o can be straightforward, i.e. question, declaration, apology
  - o can be subject to interpretation
    - yeah, right agreement or sarcasm?
    - what!? question, exclamation, or reaction?
  - o solution test how well two people agree on given dialogue acts
    - inter-rater reliability
- **inter-rater reliability** degree of agreement between raters where raters work independently of each other
  - o application validation of rating protocols
- · useful when rating protocols are ambiguous
  - applying dialogue act tags
  - codes from thematic analysis
  - o judging the quality of something

## Agreement Calculations

• agreement - probability that you and your partner selected the same tag for an item on the list

$$agreement = \frac{count(item\ rated\ the\ same)}{count(item)}$$

- observed vs. expected agreement determine what agreement was likely due to chance
  - observed agreement probability that items were rated the same

$$P(items\ rated\ the\ same)$$

• expected agreement - sum over all ratings

$$P(item\ rated\ by\ both\ as\ X)$$

$$= P(judge\ 1\ rated\ X \cap judge\ 2\ rated\ X)$$

• if judges rated independently

$$P(judge\ 1\ rated\ X) * P(judge\ 2\ rated\ X)$$

- example
  - rate 20 items good or bad

- o rater 1 rated 1 item bad rest good
- o rater 2 rated 2 itmes bad rest good
- o all the bad rates, the other rater rated that item as good
- $\circ$  observed agreement = 17 / 20 = 0.85
- expected agreement make table where entry is the count that the rater rated items that class out of all items

	Rater 1	Rater 2
Bad	0.05	0.10
Good	0.95	0.90

$$\circ$$
 bad = 0.05 x 0.10 = 0.005

$$\circ$$
 good = 0.95 x 0.90 = 0.855

$$\circ$$
 total = 0.855 + 0.005

### Cohen's Kappa

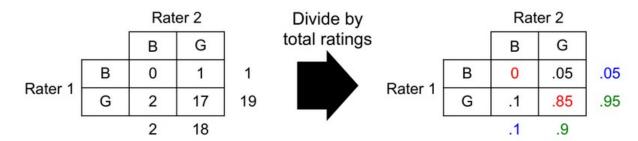
• measures the degree to which two raters' agreement exceeds chance

$$k=rac{O-E}{1-E}$$

- O is observed agreement, E expected agreement
- from previous example

#### **Raw Frequencies**

#### **Relative Frequencies**



$$\circ$$
 O = 0 + 0.85 = 0.85

$$\circ$$
 E = (0.05 x 0.1) + (0.95 x 0.9) = 0.86

$$\circ$$
 k = (0.85 - 0.86) / (1 - 0.86) = -0.071, poor agreement

- kappa ranges from -1 to 1
  - k > 0 indicates agreement better than chance
    - k = 1 perfect agreement

- $\circ$  k < 0 indicates agreement worse than chance
  - k = -1 perfect disagreement and 50% expected agreement
- $\circ\,$  applicable when dara are nominal and unordered

Score	Interpretation
< 0	poor
0 - 0.2	slight
0.2 - 0.4	fair
0.41 - 0.6	moderate
0.61 - 0.8	substantial
0.81 - 1	almost perfect

		Rater 2			
		В	G	Meh	
Rater 1	В	5	1	0	6
	G	1	9	1	11
	Meh	1	1	1	3
		7	11	2	

		Rater 2				
		В	G	Meh		
Rater 1	В	.25	.05	0	.3	
	G	.05	.45	.05	.55	
	Meh	.05	.05	.05	.15	
		.35	.55	.1		

• example