We used two metrics to analyze inter-agreement, textual agreement and overlap agreement. Textual agreement uses Cohen's kappa score (ref + expl.) to measure agreement over units that either one or both of the annotators have annotated as metaphors. Overlap agreement measures the average percentage of overlap over non-identical tags (e.g. "line" vs "line in the sand"). In our first round, we

Average textual agreement: 0.6280268867300005

Average overlap agreement: 0.4887820512820513

?? the ones Jodie calculated (20 comments)

Jan 9 (20 comments):

Average textual agreement:  0.4915213787276523  
Average overlap agreement:  0.6436586679127944

Third anno

Jan 18:

Average textual agreement between annotators R and A: 0.5517895186453347

Average overlap agreement: annotators R and A: 0.6814123376623376

Average textual agreement between annotators R and V: 0.5705820440823098

Average overlap agreement: annotators R and V: 0.5753968253968254

Average textual agreement between annotators A and V: 0.6541670606822675

Average overlap agreement: annotators A and V: 0.375

Removed ‘establishment’

Average textual agreement between annotators R and A: 0.6388179731151542  
Average overlap agreement between annotators R and A: 0.6814123376623376   
  
Average textual agreement between annotators R and V: 0.6474342747106314  
Average overlap agreement between annotators R and V: 0.5753968253968255   
  
Average textual agreement between annotators A and V: 0.6541670606822675  
Average overlap agreement between annotators A and V: 0.375

March 4 (30 comments)

Random selected

A screenshot of a computer

AI-generated content may be incorrect.

Mean of pairs: