Inter Annotator Agreement

Workshop SANTA

Nils Reiter

Sept. 18, 2018

Section 1

Introduction

Motivation

- IAA expresses agreement between annotators/raters quantitatively
- Often used as an upper bound in NLP:
  Computers can't be expected to perform better than human agreement
- Annotations with high IAA are considered more reliable
- Sometimes used to steer guideline/resource development
  - '90% solution': Remove word senses for which annotators achieve less than 90% Hovy et al. (2006)
- Corpus releases should be accompanied by IAA values, to allow estimation of annotation quality

Different Metrics

- Not all annotation tasks are the same
  - PoS tagging: Assign each word to a category
    - Only categorizing
  - Sentence splitting: Mark sentence boundaries
    - Only unitizing
  - Named entities: Select a span *and* assign it to a category
    - Unitizing, categorizing
- Different metrics for different tasks!

Cohen 1960; Fleiss 1971; Fournier and Inkpen 2012; Mathet et al. 2015

Different Metrics

Common Properties

- All metrics incorporate *observed* and *expected* agreement
- Observed agreement: Extracted from the annotations
- Expected agreement: Agreement to be expected by chance annotations
  - Indicates difficulty of the annotation task
  - Allows comparing agreement values with different numbers of categories!

Expected Agreement

If two annotators assign word classes (noun, verb, adjective, other) by throwing a 4-sided die, they achieve a certain level of agreement (this is a categorization task).

Gamma

Section 2

Gamma

Metric γ has been published in this paper:
Yann Mathet et al. "The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3 (2015), pp. 437–479

Three Components

- Combination of expected and observed agreement
- Calculation of expected agreement
- Calculation of observed agreement

Combining Expected and Observed Agreement

Note: γ is defined based on disagreements!
Assuming we have calculated observed (δ_o) and expected (δ_e) disagreement

$$\gamma = 1 - \frac{\delta_o}{\delta_e} \tag{1}$$

Combining Expected and Observed Agreement

Examples

$$\gamma = 1 - \frac{\delta_o}{\delta_e}$$

Combining Expected and Observed Agreement

Examples

$$\gamma = 1 - \frac{\delta_o}{\delta_e}$$

δ_o	δ_e	γ	
0.99	0.01	0.98	(upper bound: 1)
0.01	0.99	-98	(lower bound: $-\infty$)
0.5	0.25	-1	
0.5	0.5	0	
0.5	0.75	0.33	
0.25	0.5	0.5	
0.5	0.5	0	
0.75	0.5	-0.5	

Table: γ scores for observed (δ_o) and expected (δ_e) disagreement

Calculating Expected Agreement

- Random annotations need to be *realistic* w.r.t. several criteria
  - Distribution of units per annotator
  - Distribution of categories
  - ...
- γ's expected disagreement is based on real annotations
  1. Take the annotations created by a real annotator
  2. Split the text at a random point
  3. Permute the two parts
  4. Repeat multiple times and calculate disagreement
- This doesn't work if the text only contains a single annotation that spans the entire text

Calculating Observed Agreement

Basics

- Local level: Measuring dissimilarity between two annotations
- Global level: Create unitary alignments over all annotations by all annotators

Calculating Observed Agreement

Situations

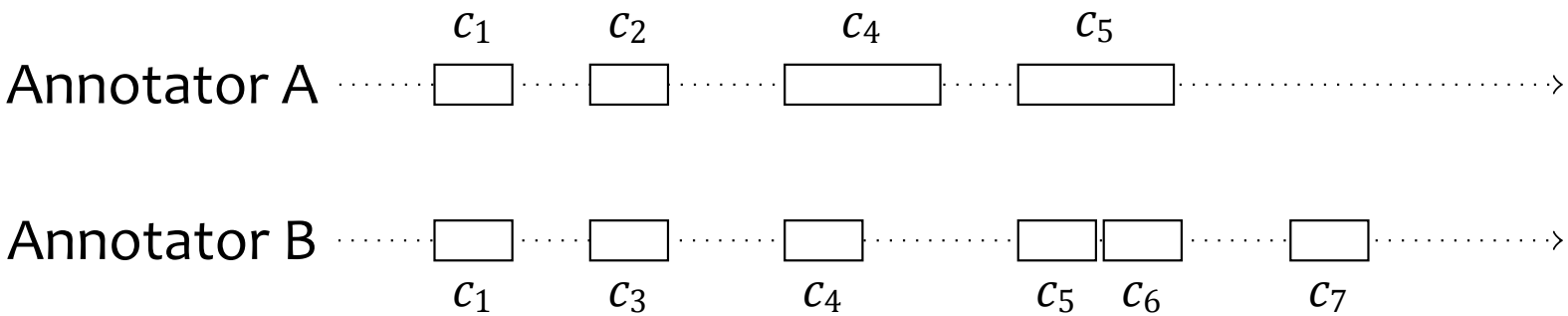


Figure: Two annotators and (some) possible situations

One Annotation is defined by

- begin/end
- feature values (including category)

If these are the same, we consider two annotations to be equal

Calculating Observed Agreement

Positional Dissimilarity

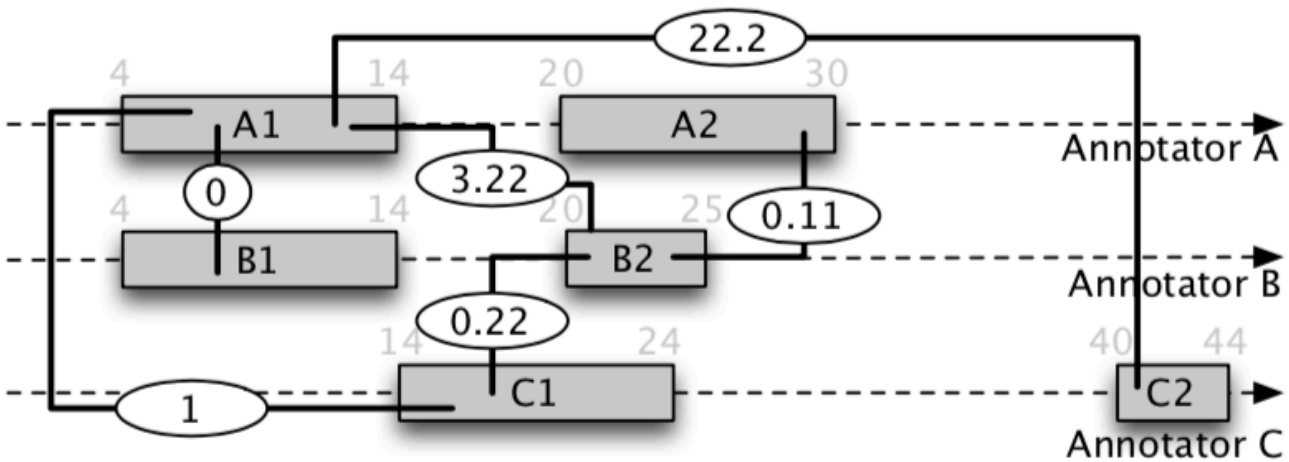
$$d_{pos}(u,v) = \left(\frac{\overbrace{|\text{start}(u) - \text{start}(v)|}^{\text{start difference}} + \overbrace{|\text{end}(u) - \text{end}(v)|}^{\text{end difference}}}{\underbrace{(\text{end}(u) - \text{start}(u))}_{\text{length of u}} + \underbrace{(\text{end}(v) - \text{start}(v))}_{\text{length of v}}} \right)^2$$

Calculating Observed Agreement

Positional Dissimilarity

$$d_{pos}(u,v) = \left(\frac{\overbrace{|\text{start}(u) - \text{start}(v)|}^{\text{start difference}} + \overbrace{|\text{end}(u) - \text{end}(v)|}^{\text{end difference}}}{\underbrace{(\text{end}(u) - \text{start}(u))}_{\text{length of u}} + \underbrace{(\text{end}(v) - \text{start}(v))}_{\text{length of v}}} \right)^2$$

Examples



Calculating Observed Agreement

Positional Dissimilarity

$$d_{pos}(u,v) = \left(\frac{\overbrace{|\text{start}(u) - \text{start}(v)|}^{\text{start difference}} + \overbrace{|\text{end}(u) - \text{end}(v)|}^{\text{end difference}}}{\underbrace{(\text{end}(u) - \text{start}(u))}_{\text{length of u}} + \underbrace{(\text{end}(v) - \text{start}(v))}_{\text{length of v}}} \right)^2$$

Here: Token numbers as positions (using a heuristic tokenizer).

Calculating Observed Agreement

Categorial Dissimilarity

Gamma

Define dissimilarity between categories in a matrix

	c_1	c_2	c_3
c_1	0	0.5	1
c_2	0.5	0	0.25
c_3	1	0.25	0

Calculating Observed Agreement

Categorial Dissimilarity

Gamma

Define dissimilarity between categories in a matrix

	c_1	c_2	c_3
c_1	0	0.5	1
c_2	0.5	0	0.25
c_3	1	0.25	0

SANTA

$$d_{cat}(u,v) = \begin{cases} 0 & \text{if } \text{cat}(u) = \text{cat}(v) \\ 1 & \text{otherwise} \end{cases}$$

I.e.: We don't use graded dissimilarity here

Calculating Observed Agreement

Categorial Dissimilarity: Features vs. Categories

Most guidelines define several individual features, instead of a single category. Feature assignments have been merged into a single string to represent a ‘category’.

Calculating Observed Agreement

Categorial Dissimilarity: Features vs. Categories

Most guidelines define several individual features, instead of a single category. Feature assignments have been merged into a single string to represent a 'category'.

Example

Addressee	:	Mouse
Speaker	:	Mouse

becomes the 'category' `Addressee=Mouse+Speaker=Mouse`

Calculating Observed Agreement

Categorial Dissimilarity: Features vs. Categories

Most guidelines define several individual features, instead of a single category. Feature assignments have been merged into a single string to represent a 'category'.

Example

Addressee	:	Mouse
Speaker	:	Mouse

becomes the 'category' `Addressee=Mouse+Speaker=Mouse`

- This is a shortcoming
- Guideline authors: Define severity of disagreement between categories

Calculating Observed Agreement

Combining Dissimilarity

$$d_{combi}(u,v) = \alpha d_{pos}(u,v) + \beta d_{cat}(u,v)$$

Calculating Observed Agreement

Combining Dissimilarity

$$d_{combi}(u, v) = \alpha d_{pos}(u, v) + \beta d_{cat}(u, v)$$

Intuitions and Remarks

- α and β can be used to express importance
  - Our setting, $\alpha = \beta = 1$
  - I.e., positional and categorial disagreement are equally important

Calculating Observed Agreement

Combining Dissimilarity

$$d_{combi}(u, v) = \alpha d_{pos}(u, v) + \beta d_{cat}(u, v)$$

Intuitions and Remarks

- α and β can be used to express importance
  - Our setting, $\alpha = \beta = 1$
  - I.e., positional and categorial disagreement are equally important
- Dissimilarity between two annotations is roughly between 0 (zero) and the squared length of the text (because of the positional dissimilarity)

Calculating Observed Agreement

Alignment

- Pairwise comparison of annotations ✓
- Which pairs do we compare?

Calculating Observed Agreement

Alignment

- Pairwise comparison of annotations ✓
- Which pairs do we compare?

Alignment

An alignment defines, which annotation of annotator 1 corresponds to which annotation of annotator 2 (if any)

Calculating Observed Agreement

Alignment

- Pairwise comparison of annotations ✓
- Which pairs do we compare?

Alignment

An alignment defines, which annotation of annotator 1 corresponds to which annotation of annotator 2 (if any)

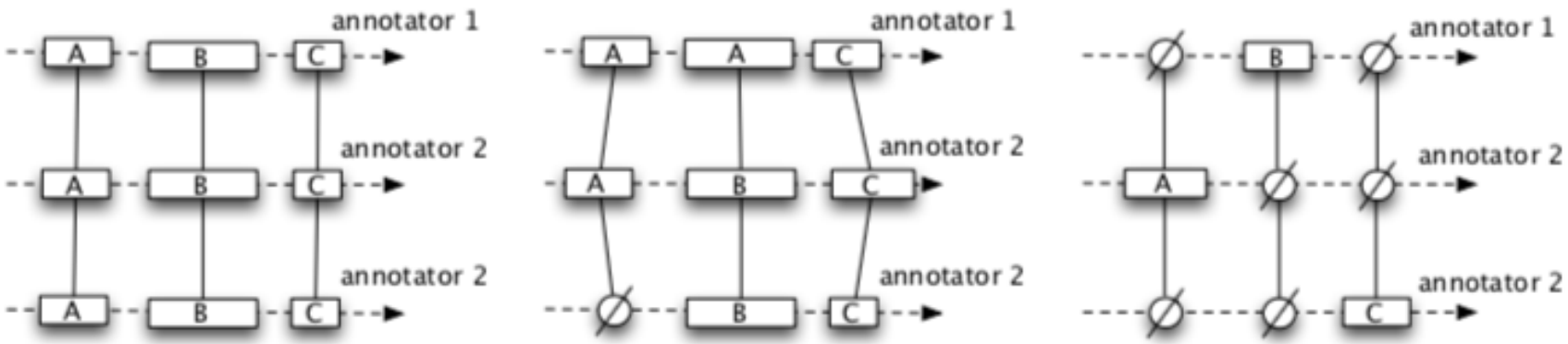


Figure: Different alignments between three annotators

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u,v)$$

with $\hat{a}$ being a set of aligned annotations

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u,v)$$

  with $\hat{a}$ being a set of aligned annotations

- Calculate disagreement over a set of annotators: Average

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u,v)$$

with $\hat{a}$ being a set of aligned annotations

- Calculate disagreement over a set of annotators: Average

$$\bar{\delta}(A) = \frac{1}{|x|} \sum_{i=1}^{|\hat{a}|} \hat{\delta}(\hat{a}_i)$$

with A being a set of annotators, and $|x|$ the mean number of annotations per annotator

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u,v)$$

  with $\hat{a}$ being a set of aligned annotations

- Calculate disagreement over a set of annotators: Average

$$\bar{\delta}(A) = \frac{1}{|x|} \sum_{i=1}^{|\hat{a}|} \hat{\delta}(\hat{a}_i)$$

  with A being a set of annotators, and $|x|$ the mean number of annotations per annotator

- Alignment is created such that $\bar{\delta}(A)$ is minimal

Calculating Observed Agreement

Summary

- Gamma combines alignment and agreement calculation
- Core: Compare annotations pairwise, w.r.t.
  - their position
  - their categories
- Settable parameters
  - Dissimilarity of categories
  - Weighting between dissimilarity types
  - Position metric (SANTA: token numbers)
- Computationally expensive
- Implementation by Mathet et al. (2015) using ILP
  `https://gamma.greyc.fr`

Results

Results

Note

The report shows Observed and Expected Disorder, i.e., the lower the better. Gamma scores represent agreement, i.e., the higher the better

All Annotations

Results

Observed Disorder (all annotations)

SANTA_	1	2	3	4	5	6	7	8
Buechner	2.07	1.87	2.47	2.57	1.96	2.18	2.36	1.39
Chekhov	1.85	1.90	2.62	2.67	2.41	3.04	0.88	2.50
Henry	1.93	2.57	2.27	2.17	2.60	2.16	1.35	1.70
Kafka	1.38	1.03	2.68	1.50	1.49	1.82	1.27	1.00
Kleist	1.99	1.40	2.79	1.63	2.11	1.53	1.36	0.75
Lagerloef	1.81	2.75	2.36	2.45	1.50	2.81	0.91	0.67
Storm	2.07	1.62	2.48	2.35	2.12	1.97	1.42	1.45
Tieck	1.78	1.99	2.18	2.79	2.32	2.23	1.64	1.85
Min	1.38	1.03	2.18	1.50	1.49	1.53	0.88	0.67
Mean	1.86	1.89	2.48	2.27	2.06	2.22	1.40	1.41
Max	2.07	2.75	2.79	2.79	2.60	3.04	2.36	2.50
Stddev.	0.22	0.57	0.21	0.48	0.40	0.49	0.46	0.61

Table: Observed Disorder (the lower the better)

Results

Expected Disorder (all annotations)

SANTA_	1	2	3	4	5	6	7	8
Buechner	2.28	2.41	2.66	2.34	2.27	2.34	2.33	
Chekhov	2.10	2.24	2.47	2.54	2.51	2.88	1.96	
Henry	2.05	2.56	2.22	2.49	2.41	2.03	2.15	
Kafka	2.14				2.40	2.79	2.03	
Kleist	2.42	2.37	2.13			2.61	1.52	1.46
Lagerloef	1.82	2.82	2.38	2.26	2.03	2.39	1.63	
Storm	2.14	2.06	2.49	2.23	2.29	2.05	2.16	2.13
Tieck	2.02	2.05	2.55	2.25	2.39	2.08	2.38	
Min	1.82	2.05	2.13	2.23	2.03	2.03	1.52	1.46
Mean	2.12	2.36	2.41	2.35	2.33	2.40	2.02	1.80
Max	2.42	2.82	2.66	2.54	2.51	2.88	2.38	2.13
Stddev.	0.18	0.28	0.19	0.13	0.15	0.34	0.31	0.47

Table: Expected Disorder (the lower the better)

Results

Inter-Annotator Agreement Gamma (all annotations)

SANTA_	1	2	3	4	5	6	7	8
Buechner	0.09	0.22	0.07	−0.10	0.14	0.07	−0.01	
Chekhov	0.12	0.15	−0.06	−0.05	0.04	−0.05	0.55	
Henry	0.06	0.00	−0.02	0.13	−0.08	−0.06	0.37	
Kafka	0.36				0.38	0.35	0.37	
Kleist	0.18	0.41	−0.31			0.41	0.11	0.49
Lagerloef	0.01	0.02	0.01	−0.08	0.26	−0.18	0.44	
Storm	0.03	0.21	0.00	−0.05	0.07	0.04	0.34	0.32
Tieck	0.12	0.03	0.15	−0.24	0.03	−0.07	0.31	
Min	0.01	0.00	−0.31	−0.24	−0.08	−0.18	−0.01	0.32
Mean	0.12	0.15	−0.02	−0.07	0.12	0.06	0.31	0.40
Max	0.36	0.41	0.15	0.13	0.38	0.41	0.55	0.49
Stddev.	0.11	0.15	0.14	0.12	0.15	0.21	0.18	0.12

Table: Inter-Annotator Agreement Gamma (the higher the better)

Own vs. Foreign Annotations

Results

Observed Disorder (own vs. foreign)

SANTA_	1	2	3	4	5	6	7	8
Buechner	1.72	1.64	2.00	1.99	1.47	1.39	1.49	1.11
Chekhov	1.54	1.33	1.90	1.96	1.68	2.00	0.56	1.84
Henry	1.61	1.64	2.00	1.42	1.00	1.26	1.20	2.12
Kafka	1.25	1.07	1.62	1.50	1.22	1.18	0.80	0.00
Kleist	1.83	0.91	1.95	1.37	2.00	1.44	0.00	0.00
Lagerloef	1.75	1.73	2.05	1.82	1.85	2.50	0.50	1.00
Storm	1.94	1.17	1.92	1.89	1.55	1.68	1.23	1.59
Tieck	1.61	1.55	1.94	2.00	1.75	1.85	1.80	0.93
Min	1.25	0.91	1.62	1.37	1.00	1.18	0.00	0.00
Mean	1.65	1.38	1.92	1.74	1.57	1.66	0.95	1.07
Max	1.94	1.73	2.05	2.00	2.00	2.50	1.80	2.12
Stddev.	0.22	0.57	0.21	0.48	0.40	0.49	0.46	0.61

Table: Observed Disorder (the lower the better)

Results

Expected Disorder (own vs. foreign)

SANTA_	1	2	3	4	5	6	7	8
Buechner	1.68	1.81	1.60	1.79	1.83	1.44	1.49	1.55
Chekhov	1.66	1.67	1.79	1.56	1.81	1.51	1.50	1.70
Henry	1.71	1.78	1.59	1.83	1.55	1.29	1.66	1.67
Kafka	1.61	1.80	1.50	1.56	1.91	1.70	1.42	
Kleist	1.66	1.65	1.39	1.56		1.81	0.00	0.00
Lagerloef	1.58	1.90	1.51	1.52	1.68	1.93	1.32	
Storm	1.59	1.64	1.65	1.69	1.62	1.63	1.64	1.63
Tieck	1.59	1.60	1.82	1.70	1.81	1.58	1.81	1.38
Min	1.58	1.60	1.39	1.52	1.55	1.29	0.00	0.00
Mean	1.64	1.73	1.61	1.65	1.75	1.61	1.35	1.32
Max	1.71	1.90	1.82	1.83	1.91	1.93	1.81	1.70
Stddev.	0.18	0.28	0.19	0.13	0.15	0.34	0.31	0.47

Table: Expected Disorder (the lower the better)

Results

Inter-Annotator Agreement Gamma (own vs. foreign)

SANTA_	1	2	3	4	5	6	7	8
Buechner	−0.02	0.09	−0.25	−0.11	0.20	0.04	0.00	0.28
Chekhov	0.07	0.21	−0.06	−0.26	0.07	−0.32	0.63	−0.08
Henry	0.06	0.08	−0.26	0.22	0.36	0.02	0.28	−0.27
Kafka	0.23	0.41	−0.08	0.04	0.36	0.31	0.44	1.00
Kleist	−0.10	0.45	−0.40	0.12		0.20	1.00	1.00
Lagerloef	−0.10	0.09	−0.36	−0.19	−0.10	−0.30	0.62	
Storm	−0.21	0.29	−0.17	−0.11	0.05	−0.03	0.25	0.02
Tieck	−0.01	0.03	−0.06	−0.18	0.03	−0.17	0.00	0.32
Min	−0.21	0.03	−0.40	−0.26	−0.10	−0.32	0.00	−0.27
Mean	−0.01	0.21	−0.21	−0.06	0.14	−0.03	0.40	0.33
Max	0.23	0.45	−0.06	0.22	0.36	0.31	1.00	1.00
Stddev.	0.11	0.15	0.14	0.12	0.15	0.21	0.18	0.12

Table: Inter-Annotator Agreement Gamma (the higher the better)

Own vs. Student Annotations

Results

Observed Disorder (own vs. student)

SANTA_	1	2	3	4	5	6	7	8
Buechner	1.48	1.07	1.77	1.00	1.53	1.99	1.87	1.33
Chekhov	1.70	1.69	1.86	1.33	1.72	2.50	0.84	1.33
Henry	1.59	2.11	1.59	1.67	1.69	1.75	0.97	1.50
Kafka	1.33	1.00	2.00	0.00	1.16	2.25	0.90	1.00
Kleist	1.62	1.31	1.95	1.00	1.73	1.25	1.36	0.75
Lagerloef	1.58	2.00	1.59	0.67	0.63	2.01	0.95	0.00
Storm	1.47	1.30	1.82	0.57	1.70	1.71	1.05	0.42
Tieck	1.52	1.49	1.54	1.32	1.90	1.70	0.92	1.00
Min	1.33	1.00	1.54	0.00	0.63	1.25	0.84	0.00
Mean	1.53	1.50	1.76	0.94	1.51	1.90	1.11	0.92
Max	1.70	2.11	2.00	1.67	1.90	2.50	1.87	1.50
Stddev.	0.22	0.57	0.21	0.48	0.40	0.49	0.46	0.61

Table: Observed Disorder (the lower the better)

Results

Expected Disorder (own vs. student)

SANTA_	1	2	3	4	5	6	7	8
Buechner	1.58	1.75	1.86		1.65	1.75	1.64	1.35
Chekhov	1.66	1.72	1.67	1.34	1.76	1.88	1.57	1.30
Henry	1.51	1.91	1.63	1.77	1.48	1.70	1.60	1.44
Kafka	1.77	1.80	1.64		1.68	2.09	1.54	
Kleist	1.70	1.74	1.61	1.43	1.78	1.82	1.52	1.51
Lagerloef	1.40	1.94	1.68	1.19	1.43	1.41	1.37	1.50
Storm	1.49	1.48	1.77	1.50	1.67	1.70	1.66	1.55
Tieck	1.49	1.45	1.88	1.36	1.74	1.64	1.71	1.44
Min	1.40	1.45	1.61	1.19	1.43	1.41	1.37	1.30
Mean	1.58	1.72	1.72	1.43	1.65	1.75	1.58	1.44
Max	1.77	1.94	1.88	1.77	1.78	2.09	1.71	1.55
Stddev.	0.18	0.28	0.19	0.13	0.15	0.34	0.31	0.47

Table: Expected Disorder (the lower the better)

Results

Inter-Annotator Agreement Gamma (own vs. student)

SANTA_	1	2	3	4	5	6	7	8
Buechner	0.07	0.39	0.05		0.07	−0.14	−0.14	0.01
Chekhov	−0.02	0.02	−0.11	0.01	0.03	−0.33	0.47	−0.02
Henry	−0.05	−0.10	0.02	0.06	−0.14	−0.03	0.39	−0.04
Kafka	0.25	0.44	−0.22	1.00	0.31	−0.07	0.42	
Kleist	0.05	0.25	−0.21	0.30	0.02	0.32	0.11	0.50
Lagerloef	−0.13	−0.03	0.06	0.44	0.56	−0.43	0.31	1.00
Storm	0.01	0.12	−0.03	0.62	−0.02	−0.01	0.37	0.73
Tieck	−0.02	−0.03	0.18	0.03	−0.09	−0.04	0.47	0.30
Min	−0.13	−0.10	−0.22	0.01	−0.14	−0.43	−0.14	−0.04
Mean	0.02	0.13	−0.03	0.35	0.09	−0.09	0.30	0.35
Max	0.25	0.44	0.18	1.00	0.56	0.32	0.47	1.00
Stddev.	0.11	0.15	0.14	0.12	0.15	0.21	0.18	0.12

Table: Inter-Annotator Agreement Gamma (the higher the better)

Foreign vs. Student Annotations

Results

Observed Disorder (foreign vs. student)

SANTA_	1	2	3	4	5	6	7	8
Buechner	1.67	1.67	1.97	1.86	1.68	1.88	1.52	1.02
Chekhov	1.38	1.69	1.61	1.75	2.17	2.07	0.88	1.51
Henry	1.52	2.17	2.00	1.77	1.83	1.66	1.20	1.52
Kafka	1.17	1.08	2.00	1.50	1.41	2.33	1.18	1.00
Kleist	1.48	0.91	1.50	1.25	1.68	1.21	0.00	0.00
Lagerloef	1.06	1.96	2.00	1.79	1.73	1.84	0.84	1.00
Storm	1.89	1.56	2.02	1.93	1.80	1.50	1.31	1.38
Tieck	1.41	1.91	2.00	2.00	1.65	1.73	1.83	1.20
Min	1.06	0.91	1.50	1.25	1.41	1.21	0.00	0.00
Mean	1.45	1.62	1.89	1.73	1.74	1.78	1.09	1.08
Max	1.89	2.17	2.02	2.00	2.17	2.33	1.83	1.52
Stddev.	0.22	0.57	0.21	0.48	0.40	0.49	0.46	0.61

Table: Observed Disorder (the lower the better)

Results

Expected Disorder (foreign vs. student)

SANTA_	1	2	3	4	5	6	7	8
Buechner	1.61	1.82	1.52	1.43	1.76	1.85	1.51	1.20
Chekhov	1.57	1.68	1.67	1.77	1.89	1.88	1.55	
Henry	1.55	2.06	1.55	1.72	1.60	1.49	1.70	1.52
Kafka	1.57	2.43		1.61	1.78	2.17	1.65	
Kleist	1.69	1.65		1.39		1.78	0.00	0.00
Lagerloef	1.14	1.80	1.46	1.53	1.61	1.81	1.44	
Storm	1.63	1.74	1.65	1.65	1.71	1.50	1.69	1.71
Tieck	1.63	1.73	1.86	1.51	1.84	1.55	1.79	1.49
Min	1.14	1.65	1.46	1.39	1.60	1.49	0.00	0.00
Mean	1.55	1.86	1.62	1.58	1.74	1.75	1.42	1.19
Max	1.69	2.43	1.86	1.77	1.89	2.17	1.79	1.71
Stddev.	0.18	0.28	0.19	0.13	0.15	0.34	0.31	0.47

Table: Expected Disorder (the lower the better)

Results

Inter-Annotator Agreement Gamma (foreign vs. student)

SANTA_	1	2	3	4	5	6	7	8
Buechner	−0.04	0.09	−0.29	−0.30	0.04	−0.01	−0.01	0.15
Chekhov	0.12	0.00	0.04	0.01	−0.15	−0.10	0.44	
Henry	0.02	−0.05	−0.29	−0.03	−0.15	−0.11	0.29	0.01
Kafka	0.25	0.56		0.07	0.21	−0.08	0.29	
Kleist	0.13	0.45		0.10		0.32	1.00	1.00
Lagerloef	0.06	−0.09	−0.37	−0.17	−0.08	−0.02	0.42	
Storm	−0.16	0.10	−0.23	−0.17	−0.05	0.00	0.23	0.19
Tieck	0.13	−0.10	−0.08	−0.32	0.10	−0.12	−0.02	0.20
Min	−0.16	−0.10	−0.37	−0.32	−0.15	−0.12	−0.02	0.01
Mean	0.07	0.12	−0.20	−0.10	−0.01	−0.01	0.33	0.31
Max	0.25	0.56	0.04	0.10	0.21	0.32	1.00	1.00
Stddev.	0.11	0.15	0.14	0.12	0.15	0.21	0.18	0.12

Table: Inter-Annotator Agreement Gamma (the higher the better)

Results

Comments

- Obvious ways to boost scores
  - Simple schemes without many features get higher scores (in addition to being easier to annotate)
- Some guidelines include more narrative phenomena than 'levels'
  - Can also lead to lower IAA