Proseminar Linguistische Annotation

Ines Rehbein und Josef Ruppenhofer

SS 2010

Wie kann man die Qualität der Annotationen messen?

- Reliabilität:
  - Wie hoch ist die Übereinstimmung zwischen den Annotator/innen? (Inter-Annotator Agreement (IAA), Inter-Coder Agreement)
  - Wie hoch ist die Übereinstimmung, wenn die gleiche Annotatorin die gleichen Daten nach einer gewissen Zeitspanne erneut annotiert?
- Validität:
  - sind die Annotationen korrekt? (Accuracy)

Annotator	**Dogan**	**Pat.**
$Sentence_1$	Ort	Ort
$Sentence_2$	Artefakt	Geschehen
$Sentence_3$	Artefakt	Artefakt
$Sentence_4$	Geschehen	Ort
$Sentence_5$	Artefakt	Ort
...	...	...
$Sentence_{50}$	Artefakt	Ort

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wovon wird Inter-Annotator Agreement beeinflusst?

- Anzahl der Annotator/innen
- Anzahl der Kategorien/Label/Klassen
- Verteilung der Instanzen auf die verschiedenen Kategorien (höhere Übereinstimmung für häufige/dominante Kategorien)
- Training und Annotationsrichtlinien
- Bias der einzelnen Annotator/innen
- Annotationstool
- Methode der Annotation (z.B. automatische Vorannotation)

Wie kann man die Qualität der Annotationen messen?

- **Prozentuale Übereinstimmung (percentage agreement)**: wieviel % der Label wurden von beiden Annotator/innen vergeben?

Annotator	**Dogan**	**Pat.**
$Sentence_1$	Ort	Ort
$Sentence_2$	Artefakt	Geschehen
$Sentence_3$	Artefakt	Artefakt
$Sentence_4$	Geschehen	Ort
$Sentence_5$	Artefakt	Ort
...	...	...
$Sentence_{50}$	Artefakt	Ort

$$agr_i = \begin{cases} 1 \textit{ if the two annotators assign } i \textit{ to the same category} \\ 0 \textit{ if the two annotators assign } i \textit{ to different categories} \end{cases}$$

Prozentuale Übereinstimmung

Annotator	**Dog.**	**Pat.**
$Sentence_1$	Ort	Ort
$Sentence_2$	Artefakt	Geschehen
$Sentence_3$	Artefakt	Artefakt
...	...	...
$Sentence_{50}$	Artefakt	Ort

Konfusionsmatrix

Annotator		**Dogan**			
		Artefakt	**Geschehen**	**Ort**	**Total**
Patricia	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

Prozentuale Übereinstimmung II

Annotator		Dogan			
		Artefakt	**Geschehen**	**Ort**	**Total**
Pat.	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

Observed Agreement:

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i \tag{1}$$

Prozentuale Übereinstimmung II

Annotator		**Dogan**			
		Artefakt	**Geschehen**	**Ort**	**Total**
Pat.	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

Observed Agreement:

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i = \frac{1}{50}(0 + 33 + 12) = \frac{45}{50} = 0.9 \qquad (2)$$

Prozentuale Übereinstimmung

- Nachteile:
  - berücksichtigt nicht den Anteil an Übereinstimmung, der durch Zufall erreicht wurde (**chance agreement**) und der sehr stark von der jeweiligen Annotationsaufgabe abhängt
  - Bias zugunsten einer kleinen Anzahl an Kategorien (Labeln)
  - berücksichtigt nicht die Verteilung an Instanzen in den jeweiligen Kategorien (höhere Übereinstimmung für die dominante Klasse)

Observed Agreement has to be corrected for chance agreement

Beobachtetes Agreement (Agr_o) muss korrigiert werden, um den Anteil an Übereinstimmung zu berücksichtigen, der durch Zufall bedingt ist.

Prozentuale Übereinstimmung

- Nachteile:
  - berücksichtigt nicht den Anteil an Übereinstimmung, der durch Zufall erreicht wurde (**chance agreement**) und der sehr stark von der jeweiligen Annotationsaufgabe abhängt
  - Bias zugunsten einer kleinen Anzahl an Kategorien (Labeln)
  - berücksichtigt nicht die Verteilung an Instanzen in den jeweiligen Kategorien (höhere Übereinstimmung für die dominante Klasse)

Observed Agreement has to be corrected for chance agreement

Beobachtetes Agreement (Agr_o) muss korrigiert werden, um den Anteil an Übereinstimmung zu berücksichtigen, der durch Zufall bedingt ist.

Prozentuale Übereinstimmung

- Nachteile:
  - berücksichtigt nicht den Anteil an Übereinstimmung, der durch Zufall erreicht wurde (**chance agreement**) und der sehr stark von der jeweiligen Annotationsaufgabe abhängt
  - Bias zugunsten einer kleinen Anzahl an Kategorien (Labeln)
  - berücksichtigt nicht die Verteilung an Instanzen in den jeweiligen Kategorien (höhere Übereinstimmung für die dominante Klasse)

Observed Agreement has to be corrected for chance agreement

Beobachtetes Agreement (Agr_o) muss korrigiert werden, um den Anteil an Übereinstimmung zu berücksichtigen, der durch Zufall bedingt ist.

Prozentuale Übereinstimmung

- Nachteile:
  - berücksichtigt nicht den Anteil an Übereinstimmung, der durch Zufall erreicht wurde (**chance agreement**) und der sehr stark von der jeweiligen Annotationsaufgabe abhängt
  - Bias zugunsten einer kleinen Anzahl an Kategorien (Labeln)
  - berücksichtigt nicht die Verteilung an Instanzen in den jeweiligen Kategorien (höhere Übereinstimmung für die dominante Klasse)

Observed Agreement has to be corrected for chance agreement

Beobachtetes Agreement (Agr_o) muss korrigiert werden, um den Anteil an Übereinstimmung zu berücksichtigen, der durch Zufall bedingt ist.

Prozentuale Übereinstimmung

- Nachteile:
  - berücksichtigt nicht den Anteil an Übereinstimmung, der durch Zufall erreicht wurde (**chance agreement**) und der sehr stark von der jeweiligen Annotationsaufgabe abhängt
  - Bias zugunsten einer kleinen Anzahl an Kategorien (Labeln)
  - berücksichtigt nicht die Verteilung an Instanzen in den jeweiligen Kategorien (höhere Übereinstimmung für die dominante Klasse)

Observed Agreement has to be corrected for chance agreement

Beobachtetes Agreement (Agr_o) muss korrigiert werden, um den Anteil an Übereinstimmung zu berücksichtigen, der durch Zufall bedingt ist.

Chance-corrected coefficients for measuring agreement between two annotators

- Welchen Anteil an Übereinstimmung können wir zufallsbedingt erwarten (A_e), wenn 2 Annotatoren unabhängig voneinander Kategorien zuweisen?
- 2 unabhängige Ereignisse (im Sinne der Wahrscheinlichkeitstheorie; Vergleich: Münzwurf (2 Kategorien), Würfel (6 Kategorien)
- Wenn wir diesen erwarteten Zufallsanteil wissen und herausrechnen können, bekommen wir ein verlässlicheres Maß der Übereinstimmung:
  - $1 - A_e$: Welche Übereinstimmung über Zufall hinaus ist möglich?
  - $A_o - A_e$: Welche Übereinstimmung über den Zufall hinausgehend wurde tatsächlich gefunden?
  - Maße, basierend auf dem Verhältnis zwischen gefundenem Agreement $A_o - A_e$ und möglichem Agreement $1 - A_e$:

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e} \tag{3}$$

Chance-corrected coefficients for measuring agreement between two annotators II

- Bekannte zufalls-korrigierte Koeffizienten:
  - S (Bennett, Alpert, and Goldstein 1954)
  - π (Scott 1955)
  - κ (Cohen 1960)
  - Krippendorff's α: basiert auf einer verwandten Formel, aber misst Disagreement

Chance-corrected coefficients for measuring agreement between two annotators III

- A_o - leicht zu berechnen (wie häufig stimmen 2 Annotatorinnen überein?)
- A_e (*chance agreement*): Wahrscheinlichkeit, dass 2 Annotatoren eine beliebige Instanz *zufällig* mit der gleichen Kategorie auszeichnen
- alle 3 Koeffizienten treffen die Unabhängigkeitsannahme (*indepencence assumption*)
  $\Rightarrow$ expected agreement: Wahrscheinlichkeit, dass a_1 und a_2 auf jeder beliebigen Kategorie k übereinstimmen:

$$A_e^S = A_e^\pi = A_e^\kappa = \sum_{k \in K} P(k|a_1) \cdot P(k|a_2) \tag{4}$$

Unterschiede zwischen den 3 Koeffizienten S, π, κ

- Unterschied zwischen S, π, κ: Berechnung von $P(k|a_i)$ (Wahrscheinlichkeit, dass Annotator a_i eine beliebige Instanz als Kategorie k annotiert)
  - S: $P(k_j|a_m) = P(k_l|a_n)$ (keine Unterscheidung zwischen Kategorien und Annotatorinnen)
  - π: $P(k_j|a_m) = P(k_j|a_n)$ (Unterscheidung zwischen Kategorien, aber nicht zwischen Annotatorinnen)
  - κ: (Unterscheidung zwischen Kategorien und zwischen Annotatorinnen: Annahme einer separaten Distribution für jede Annotatorin)
- Problem: wir haben kein Vorwissen über die Verteilung der Instanzen auf die verschiedenen Kategorien
- Verteilung der Kategorien (für π) und der Bias der Annotatoren (für κ) muss anhand der beobachteten Daten geschätzt werden

S - gleiche Verteilung für alle Kategorien

$$A_e^S = A_e^\pi = A_e^\kappa = \sum_{k \in K} P(k|a_1) \cdot P(k|a_2) \tag{5}$$

- Zufällige Auswahl einer Kategorie aus einer gleichmäßigen Verteilung (alle Kategorien sind gleich wahrscheinlich) $P(k|a_i) = \frac{1}{k}$

$$A_e^S = \sum_{k \in K} \frac{1}{k} \cdot \frac{1}{k} = k \cdot \left(\frac{1}{k}\right)^2 = \frac{1}{k} \tag{6}$$

- Probleme mit S:
  - Bias gegenüber feinkörnigen Annotationsschemata mit selten vorkommenden Kategorien
  - gleichmäßige Verteilung (uniform distribution) ist sehr unwahrscheinlich in Bezug auf natürliche Sprache

S - gleiche Verteilung für alle Kategorien: Beispiel

Annotator		Dogan			
		Artefakt	**Geschehen**	**Ort**	**Total**
Pat.	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i = \frac{1}{50}(0 + 33 + 12) = \frac{45}{50} = 0.9 \tag{7}$$

$$A_e^S = \sum_{k \in K} \frac{1}{k} \cdot \frac{1}{k} = 3 \cdot \left(\frac{1}{3}\right)^2 = 0.333 \tag{8}$$

$$S = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.333}{1 - 0.333} = 0.85 \tag{9}$$

S - gleiche Verteilung für alle Kategorien: Beispiel

Annotator		**Dogan**			
		Artefakt	**Geschehen**	**Ort**	**Total**
Pat.	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i = \frac{1}{50}(0 + 33 + 12) = \frac{45}{50} = 0.9 \tag{7}$$

$$A_e^S = \sum_{k \in K} \frac{1}{k} \cdot \frac{1}{k} = 3 \cdot \left(\frac{1}{3}\right)^2 = 0.333 \tag{8}$$

$$S = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.333}{1 - 0.333} = 0.85 \tag{9}$$

S - gleiche Verteilung für alle Kategorien: Beispiel

Annotator		**Dogan**			
		Artefakt	**Geschehen**	**Ort**	**Total**
Pat.	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i = \frac{1}{50}(0 + 33 + 12) = \frac{45}{50} = 0.9 \tag{7}$$

$$A_e^S = \sum_{k \in K} \frac{1}{k} \cdot \frac{1}{k} = 3 \cdot \left(\frac{1}{3}\right)^2 = 0.333 \tag{8}$$

$$S = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.333}{1 - 0.333} = 0.85 \tag{9}$$

S - gleiche Verteilung für alle Kategorien: Beispiel

Annotator		**Dogan**			
		Artefakt	**Geschehen**	**Ort**	**Total**
Pat.	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i = \frac{1}{50}(0 + 33 + 12) = \frac{45}{50} = 0.9 \tag{7}$$

$$A_e^S = \sum_{k \in K} \frac{1}{k} \cdot \frac{1}{k} = 3 \cdot \left(\frac{1}{3}\right)^2 = 0.333 \tag{8}$$

$$S = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.333}{1 - 0.333} = 0.85 \tag{9}$$

S - gleiche Verteilung für alle Kategorien: Fazit

- Zufalls-korrigiertes Agreement S ist kleiner als prozentuales Agreement (0.85 versus 0.9)
- Aber: S nimmt an, dass alle Kategorien gleich verteilt sind - stimmt das?
- Häufig auftretende Kategorien haben höhere Wahrscheinlichkeit, von den Annotatorinnen zugewiesen zu werden
- Sollte bei der Berechnung von A_e mit einbezogen werden

π - jede Kategorie hat eine eigene Verteilung

- Die zufällige Zuweisung von Kategorien zu Instanzen wird von der Verteilung der Instanzen auf die verschiedenen Kategorien beeinflusst
- Schätzung von $\hat{P}(k)$ (beobachteter Anteil an Instanzen, die von beiden Annotatoren mit Kategorie k annotiert wurden)
  $P(k|a_1) = P(k|a_2) = \hat{P}(k)$

$$\hat{P}(k) = \frac{n_k}{2_i} \tag{10}$$

(Anzahl an Instanzen, die von beiden Annotatorinnen n_k mit Kategorie k ausgezeichnet wurden, geteilt durch die Gesamtanzahl an zugewiesenen Kategorien)

- *Expected agreement:*

$$A_e^{\pi} = \sum_{k \in K} \hat{P}(k) \cdot \hat{P}(k) = \sum_{k \in K} \left(\frac{n_k}{2_i} \right)^2 = \frac{1}{4i^2} \sum_{k \in K} n_k^2 \tag{11}$$

π - jede Kategorie hat eine eigene Verteilung: Beispiel

$$
\begin{array}{lll}
P(Artefakt|Annot_1) & P(Artefakt|Annot_2) & = \hat{P}(Artefakt) = 0.05 \\
P(Geschehen|Annot_1) & P(Geschehen|Annot_2) & = \hat{P}(Geschehen) = 0.70 \\
P(Ort|Annot_1) & P(Ort|Annot_2) & = \hat{P}(Ort) = 0.25
\end{array}
$$

$$
A_e^{\pi} = \sum_{k \in K} \hat{P}(k) \cdot \hat{P}(k) = 0.05^2 + 0.70^2 + 0.25^2 = 0.0025 + 0.49 + 0.0625 = 0.555 \tag{12}
$$

$$
\pi = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.555}{1 - 0.555} = 0.775 \tag{13}
$$

π - jede Kategorie hat eine eigene Verteilung: Beispiel

$$
\begin{array}{lll}
P(Artefakt|Annot_1) & P(Artefakt|Annot_2) & = \hat{P}(Artefakt) = 0.05 \\
P(Geschehen|Annot_1) & P(Geschehen|Annot_2) & = \hat{P}(Geschehen) = 0.70 \\
P(Ort|Annot_1) & P(Ort|Annot_2) & = \hat{P}(Ort) = 0.25
\end{array}
$$

$$
A_e^{\pi} = \sum_{k \in K} \hat{P}(k) \cdot \hat{P}(k) = 0.05^2 + 0.70^2 + 0.25^2 = 0.0025 + 0.49 + 0.0625 = 0.555 \tag{12}
$$

$$
\pi = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.555}{1 - 0.555} = 0.775 \tag{13}
$$

π - jede Kategorie hat eine eigene Verteilung: Beispiel

$$
\begin{array}{lll}
P(Artefakt|Annot_1) & P(Artefakt|Annot_2) & = \hat{P}(Artefakt) = 0.05 \\
P(Geschehen|Annot_1) & P(Geschehen|Annot_2) & = \hat{P}(Geschehen) = 0.70 \\
P(Ort|Annot_1) & P(Ort|Annot_2) & = \hat{P}(Ort) = 0.25
\end{array}
$$

$$
A_e^{\pi} = \sum_{k \in K} \hat{P}(k) \cdot \hat{P}(k) = 0.05^2 + 0.70^2 + 0.25^2 = 0.0025 + 0.49 + 0.0625 = 0.555 \tag{12}
$$

$$
\pi = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.555}{1 - 0.555} = 0.775 \tag{13}
$$

π - jede Kategorie hat eine eigene Verteilung: Beispiel

$$
\begin{array}{lll}
P(Artefakt|Annot_1) & P(Artefakt|Annot_2) & = \hat{P}(Artefakt) = 0.05 \\
P(Geschehen|Annot_1) & P(Geschehen|Annot_2) & = \hat{P}(Geschehen) = 0.70 \\
P(Ort|Annot_1) & P(Ort|Annot_2) & = \hat{P}(Ort) = 0.25
\end{array}
$$

$$
A_e^{\pi} = \sum_{k \in K} \hat{P}(k) \cdot \hat{P}(k) = 0.05^2 + 0.70^2 + 0.25^2 = 0.0025 + 0.49 + 0.0625 = 0.555 \tag{12}
$$

$$
\pi = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.555}{1 - 0.555} = 0.775 \tag{13}
$$

π - jede Kategorie hat eine eigene Verteilung: Fazit

- Zufalls-korrigiertes Agreement π ist kleiner als s (0.775 versus 0.85)
- Aber: π nimmt an, dass alle Annotatoren gleich entscheiden - stimmt das?
- Manche Annotatorinnen sind besser als andere (Training/Eignung/Lust und Laune)
- Annotatoren können Vorlieben für bestimmte Kategorien haben (Bias)
- Sollte bei der Berechnung von A_e mit einbezogen werden

κ - eigene Verteilung für Kategorien und Annotatoren

- Individuelle Verteilung für jede Annotatorin:
  - zufällige Zuweisung von Kategorien zu Instanzen wird von A-priori-Wahrscheinlichkeit (Anfangswahrscheinlichkeit) gelenkt, die für jede Annotatorin bestimmt wird und den Bias der einzelnen Annotatorinnen reflektiert
  - Schätzung von $\hat{P}(k|a_i)$ (beobachtete Anzahl an Instanzen, die vonAnnotatorin i Kategorie k zugewiesen wurden)

$$P(k|a_i) = \hat{P}(k|a_i) = \frac{n_{a_i k}}{i} \tag{14}$$

(Anzahl an mit k annotierten Instanzen durch Annotator n_{ak}, geteilt durch die Anzahl an Instanzen i)

$$A_e^{\kappa} = \sum_{k \in K} \hat{P}(k|a_1) \cdot \hat{P}(k|a_2) = \sum_{k \in K} \frac{n_{a_1 k}}{i} \cdot \frac{n_{a_2 k}}{i} = \frac{1}{i^2} \sum_{k \in K} n_{a_1 k} n_{a_2 k} \tag{15}$$

κ - eigene Verteilung für Kategorien und Annotatoren: Beispiel

$$
\begin{array}{llll}
P(Artefakt|Annot_1) & = 0.08 & P(Artefakt|Annot_2) & = 0.02 \\
P(Geschehen|Annot_1) & = 0.68 & P(Geschehen|Annot_2) & = 0.72 \\
P(Ort|Annot_1) & = 0.24 & P(Ort|Annot_2) & = 0.26
\end{array}
$$

$$A_e^\kappa = \sum_{k \in K} \hat{P}(k|a_1) \cdot \hat{P}(k|a_2) = 0.08 \cdot 0.02 + 0.68 \cdot 0.72 + 0.24 \cdot 0.26 = 0.5536 \quad (16)$$

$$\kappa = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.5536}{1 - 0.5536} = 0.776 \quad (17)$$

κ - eigene Verteilung für Kategorien und Annotatoren: Beispiel

$$
\begin{array}{llll}
P(Artefakt|Annot_1) & = 0.08 & P(Artefakt|Annot_2) & = 0.02 \\
P(Geschehen|Annot_1) & = 0.68 & P(Geschehen|Annot_2) & = 0.72 \\
P(Ort|Annot_1) & = 0.24 & P(Ort|Annot_2) & = 0.26
\end{array}
$$

$$
A_e^\kappa = \sum_{k \in K} \hat{P}(k|a_1) \cdot \hat{P}(k|a_2) = 0.08 \cdot 0.02 + 0.68 \cdot 0.72 + 0.24 \cdot 0.26 = 0.5536 \tag{16}
$$

$$
\kappa = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.5536}{1 - 0.5536} = 0.776 \tag{17}
$$

κ - eigene Verteilung für Kategorien und Annotatoren: Beispiel

$$
\begin{array}{llll}
P(Artefakt|Annot_1) & = 0.08 & P(Artefakt|Annot_2) & = 0.02 \\
P(Geschehen|Annot_1) & = 0.68 & P(Geschehen|Annot_2) & = 0.72 \\
P(Ort|Annot_1) & = 0.24 & P(Ort|Annot_2) & = 0.26
\end{array}
$$

$$
A_e^\kappa = \sum_{k \in K} \hat{P}(k|a_1) \cdot \hat{P}(k|a_2) = 0.08 \cdot 0.02 + 0.68 \cdot 0.72 + 0.24 \cdot 0.26 = 0.5536 \tag{16}
$$

$$
\kappa = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.5536}{1 - 0.5536} = 0.776 \tag{17}
$$

κ - eigene Verteilung für Kategorien und Annotatoren: Beispiel

$$
\begin{array}{llll}
P(Artefakt|Annot_1) & = 0.08 & P(Artefakt|Annot_2) & = 0.02 \\
P(Geschehen|Annot_1) & = 0.68 & P(Geschehen|Annot_2) & = 0.72 \\
P(Ort|Annot_1) & = 0.24 & P(Ort|Annot_2) & = 0.26
\end{array}
$$

$$A_e^\kappa = \sum_{k \in K} \hat{P}(k|a_1) \cdot \hat{P}(k|a_2) = 0.08 \cdot 0.02 + 0.68 \cdot 0.72 + 0.24 \cdot 0.26 = 0.5536 \tag{16}$$

$$\kappa = \frac{A_o - A_e}{1 - A_e} = \frac{0.9 - 0.5536}{1 - 0.5536} = 0.776 \tag{17}$$

π und κ - Fazit

- π besser für Generalisierungen
- κ besser, wenn auf konkrete Annotationsaufgabe und individuelle Annotator/innen Bezug genommen wird

Was tun mit mehr als 2 Annotatoren?

- π und κ können generalisiert werden $\rightarrow$ Anwendung auf mehr als 2 Annotatoren
- Vorher: A_o: prozentualer Anteil an Instanzen, für die die 2 Annotatorinnen übereinstimmen
- Jetzt: nicht mehr möglich (was, wenn nur 2 von 3 Annotatoren übereinstimmen?)
- Konfusionsmatrix für mehr als 2 Annotatoren - wie?

Annotator		**Dogan**			
		Artefakt	**Geschehen**	**Ort**	**Total**
Patricia	**Artefakt**	0	1	0	1
	Geschehen	3	33	0	36
	Ort	1	0	12	13
	Total	4	34	12	50

Was tun mit mehr als 2 Annotatoren? II

- *Pairwise agreement*:
  - *Fleiss (1971)*: Agreement für eine bestimmte Instanz = Anteil an übereinstimmenden Beurteilungspaaren $Bp(a_i, a_j)$ aus der Gesamtanzahl an Beurteilungspaaren für diese Instanz

Instanz	**Anot1**	**Anot2**	**Anot3**
S_1	Ort	Ort	Geschehen
S_2	Artefakt	Artefakt	Artefakt
S_3	Geschehen	Artefakt	Geschehen

$S_1 : Bp(a_1, a_2), Bp(a_1, a_3), Bp(a_2, a_3) =$
$(Ort, Ort), (Ort, Geschehen), (Ort, Geschehen)$

$S_2 : Bp(a_1, a_2), Bp(a_1, a_3), Bp(a_2, a_3) =$
$(Artefakt, Artefakt), (Artefakt, Artefakt), (Artefakt, Artefakt)$

$S_3 : Bp(a_1, a_2), Bp(a_1, a_3), Bp(a_2, a_3) =$
$(Geschehen, Artefakt), (Geschehen, Geschehen), (Artefakt, Geschehen)$

Agreement Table

Annotationen

Instanz	Anot1	Anot2	Anot3
S_1	Ort	Ort	Geschehen
S_2	Artefakt	Artefakt	Artefakt
S_3	Geschehen	Artefakt	Geschehen

Agreement table

Instanz	Artefakt	Geschehen	Ort
S_1	0	1	2
S_2	3	0	0
S_3	1	2	0
total	4 (0.44)	3 (0.33)	2 (0.22)

- listet jede Instanz mit der Anzahl, wie oft diese Instanz einer bestimmten Kategorie zugewiesen wurde
- keine Information, welche Annotatorin die Zuweisung gemacht hat

Multi-π für mehr als 2 Annotatoren

$$agr_i = \frac{1}{\binom{a}{2}} \sum_{k \in K} \binom{n_{ik}}{2} = \frac{1}{a(a-1)} \sum_{k \in K} n_{ik}(n_{ik} - 1) \qquad (18)$$

- n_{ik}: gibt an, wie oft eine Instanz i als Kategorie k ausgezeichnet wurde (Anzahl der Annotatorinnen, die i als k annotieren)
- $\binom{n_{ik}}{2}$: Paare von übereinstimmenten Beurteilungen für Instanz i
- agr_i: Summe aller übereinstimmender Beurteilungspaare $\binom{n_{ik}}{2}$ für alle Kategorien, geteilt durch die Gesamtanzahl an Beurteilungspaaren $\binom{a}{2}$ für Instanz i

Multi-π für mehr als 2 Annotatoren II

$$agr_i = \frac{1}{\binom{a}{2}} \sum_{k \in K} \binom{n_{ik}}{2} = \frac{1}{a(a-1)} \sum_{k \in K} n_{ik}(n_{ik}-1) \tag{19}$$

Instanz	**Artefakt**	**Geschehen**	**Ort**
S_1	0	1	2
S_2	3	0	0
S_3	1	2	0
total	4 (0.44)	3 (0.33)	2 (0.22)

$$agr_1 = \frac{1}{\binom{3}{2}} \left(\binom{n_{S_1\,Artefakt}}{2} + \binom{n_{S_1\,Geschehen}}{2} + \binom{n_{S_1\,Ort}}{2} \right) = \frac{1}{3}(0+0+1) = 0.33 \tag{20}$$

Multi-π für mehr als 2 Annotatoren II

$$agr_i = \frac{1}{\binom{a}{2}} \sum_{k \in K} \binom{n_{ik}}{2} = \frac{1}{a(a-1)} \sum_{k \in K} n_{ik}(n_{ik}-1) \tag{19}$$

Instanz	**Artefakt**	**Geschehen**	**Ort**
S_1	0	1	2
S_2	3	0	0
S_3	1	2	0
total	4 (0.44)	3 (0.33)	2 (0.22)

$$agr_1 = \frac{1}{\binom{3}{2}} \left(\binom{n_{S_1\,Artefakt}}{2} + \binom{n_{S_1\,Geschehen}}{2} + \binom{n_{S_1\,Ort}}{2} \right) = \frac{1}{3}(0+0+1) = 0.33 \tag{20}$$

Multi-π für mehr als 2 Annotatoren III

- *Observed Agreement*

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i = \frac{1}{ia(a-1)} \sum_{i \in I} \sum_{k \in K} n_{ik}(n_{ik} - 1) \tag{21}$$

- *Expected Agreement*
  - Wahrscheinlichkeit, dass Beurteilungspaare für eine Instanz zufällig übereinstimmen (Wahrscheinlichkeit, dass 2 zufällig ausgewählte Annotatoren einer Instanz zufällig die gleiche Kategorie zuweisen)
  - gleiche Verteilung für alle Annotatoren (wie π)

$$\hat{P}(k) = \frac{1}{ia} n_k \tag{22}$$

- Anteil an Instanzen, die als Kategoie k annotiert wurden (Anzahl an mit Kategorie k annotierten Instanzen, geteilt durch die Gesamtanzahl an Annotationen)

Multi-π für mehr als 2 Annotatoren III

- A_e^π: Wahrscheinlichkeit, dass zwei zufällig ausgewählte Annotatorinnen eine Instanz i mit Kategorie k annotieren
- A_e^π ist die gemeinsame Wahrscheinlichkeit (joint probability), dass jede Annotatorin eine Instanz unabhängig voneinander mit Kategorie k annotiert

$$A_e^\pi = \sum_{k \in K} (\hat{P}(k))^2 = \sum_{k \in K} \left(\frac{1}{ia} n_k \right)^2 = \frac{1}{(ia)^2} \sum_{k \in K} n_k^2 \tag{23}$$

- multi-π wird in der Literatur auch als K bezeichnet (Siegel & Castellan, 1988)

Multi-κ für mehr als 2 Annotatoren

- Generalisierung von κ
- Separate Wahrscheinlichkeitsverteilung für Annotatoren und Kategorien
- Für Details siehe Arnstein & Posio. Inter-coder agreement for computational linguistics (survey article). *Computational Linguistics* 34(4): 555-596, 2008. `http://ron.artstein.org/publications.html`

Multi-π und multi-κ für mehr als 2 Annotatoren: Fazit

- Maße für Interannotator-Agreement, die durch Zufall erreichte Übereinstimmung mit einbeziehen
- (multi-)π: Unterscheidung zwischen Wahrscheinlichkeit der Zuweisung bestimmter Kategorien, generalisiert über alle Annotator/innen
- (multi-)κ: Unterscheidung zwischen Wahrscheinlichkeit der Zuweisung bestimmter Kategorien, bezieht Bias der Annotator/innen mit ein
- Aber - keine Unterscheidung von fehlerhaften Zuweisungen (manche Fehler sind gravierender als andere)
  z.B.: Verwechslung von FrameNets DEPARTING und PORTAL nicht so schlimm wie Verwechslung von PORTAL und STARTING_POINT
- Sollte bei der Berechnung von A_e mit einbezogen werden

Gewichtetes Agreement - Krippendorff's α

- Krippendorff's α:
  - kann auf mehr als 2 Annotatoren angewendet werden
  - bezieht mit ein, wie schwerwiegend die fehlende Übereinstimmung ist
  - kann mit fehlenden Werten umgehen
- α basiert auf ähnlichen Annahmen wie π (Generalisierung über alle Annotatoren)
- wird die fehlende Übereinstimmung zwischen den verschiedenen Kategorien gleich gewichtet, dann ist *alpha* (fast) identisch zu multi-π
- α misst Disagreement (nicht Agreement)
- α basiert auf **Varianz** als Maß der Reliabilität von Annotationen (nur möglich für numerische Werte für Kategorien)

Gewichtetes Agreement - Krippendorff's α II

- Varianz eines Samples: Summe der quadrierten Unterschiede vom Mittelwert $SS = \sum(x - \bar{x})^2$, geteilt durch die Anzahl der Freiheitsgrade *df (degrees of freedom)*
- *df*: Anzahl der frei wählbaren Elemente in einer bestimmten Berechnung
  Beispiel: Mittelwert aus 3 Zahlen $\rightarrow$ 2 Freiheitsgrade
- je kleiner die Varianz für jede Instanz, desto größer die Reliabilität der Annotation
- um Vergleichbarkeit zwischen verschiedenen Studien zu ermöglichen, muss die Varianz für jede Instanz (s^2_{within}) in Hinblick auf die erwartete Varianz skaliert werden
- erwartete Varianz: kann geschätzt werden (Gesamtvarianz der Daten (s^2_{total})

Gewichtetes Agreement - Krippendorff's α III

- Eigenschaften von s^2_{within}/s^2_{total}:
  $s^2_{within}/s^2_{total} = 0$ bei völliger Übereinstimmung
  (keine Varianz auf den Instanzen)
  $s^2_{within}/s^2_{total} = 1$ wenn alle Übereinstimmung durch Zufall bedingt ist
  $s^2_{within}/s^2_{total} > 1$ bei systematischem Disagreement

- $1 - (s^2_{within}/s^2_{total})$

ähnlich skaliertes Maß wie die vorher beschriebenen Maße

  - 1: perfekte Übereinstimmung
  - 0: durch Zufall bedingte Übereinstimmung

$$\alpha = 1 - \frac{s^2_{within}}{s^2_{total}} = 1 - \frac{SS_{within}/df_{within}}{SS_{total}/df_{total}} \quad (24)$$

Gewichtetes Agreement - Krippendorff's α III

- Eigenschaften von s^2_{within}/s^2_{total}:
  $s^2_{within}/s^2_{total} = 0$ bei völliger Übereinstimmung
  (keine Varianz auf den Instanzen)
  $s^2_{within}/s^2_{total} = 1$ wenn alle Übereinstimmung durch Zufall bedingt ist
  $s^2_{within}/s^2_{total} > 1$ bei systematischem Disagreement

  - $1 - (s^2_{within}/s^2_{total})$

  ähnlich skaliertes Maß wie die vorher beschriebenen Maße

  - 1: perfekte Übereinstimmung
  - 0: durch Zufall bedingte Übereinstimmung

$$\alpha = 1 - \frac{s^2_{within}}{s^2_{total}} = 1 - \frac{SS_{within}/df_{within}}{SS_{total}/df_{total}} \quad (24)$$

Gewichtetes Agreement - Krippendorff's α III

- Eigenschaften von s^2_{within}/s^2_{total}:
  $s^2_{within}/s^2_{total} = 0$ bei völliger Übereinstimmung
  (keine Varianz auf den Instanzen)
  $s^2_{within}/s^2_{total} = 1$ wenn alle Übereinstimmung durch Zufall bedingt ist
  $s^2_{within}/s^2_{total} > 1$ bei systematischem Disagreement

  - $1 - (s^2_{within}/s^2_{total})$

  ähnlich skaliertes Maß wie die vorher beschriebenen Maße

  - 1: perfekte Übereinstimmung
  - 0: durch Zufall bedingte Übereinstimmung

$$\alpha = 1 - \frac{s^2_{within}}{s^2_{total}} = 1 - \frac{SS_{within}/df_{within}}{SS_{total}/df_{total}} \quad (24)$$

Gewichtetes Agreement - Krippendorff's α IV

- Generalisierung von α für nicht-numerische Werte
  - Entfernen des arithmetischen Mittelwerts:

$$SS = \sum (x - \bar{x})^2 = \frac{1}{2N} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_n - x_m)^2 \tag{25}$$

(für jede Menge an Zahlen kann die Summe der quadrierten Differenzen vom Mittelwert SS auch ausgedrückt werden durch die Summme der quadrierten Differenzen zwischen allen (geordneten) Zahlenpaaren, skaliert mit einem Faktor von $\frac{1}{2}N$;

für Details siehe Arnstein & Poesio (2008), Seite 15 ff)

  - Definition einer Distanzfunktion $d_{ab} = (a - b)^2$

Gewichtetes Agreement - Krippendorff's α IV

- Generalisierung von α für nicht-numerische Werte
  - Entfernen des arithmetischen Mittelwerts:

$$SS = \sum (x - \bar{x})^2 = \frac{1}{2N} \sum_{n=1}^{N} \sum_{m=1}^{M} (x_n - x_m)^2 \tag{25}$$

(für jede Menge an Zahlen kann die Summe der quadrierten Differenzen vom Mittelwert SS auch ausgedrückt werden durch die Summme der quadrierten Differenzen zwischen allen (geordneten) Zahlenpaaren, skaliert mit einem Faktor von $\frac{1}{2}N$;

für Details siehe Arnstein & Poesio (2008), Seite 15 ff)

  - Definition einer Distanzfunktion $d_{ab} = (a - b)^2$

Gewichtetes Agreement - Krippendorff's α V

- *Observed disagreement*:

$$D_o^\alpha = 2s^2_{within} = \frac{1}{ia(a-1)} \sum_{i \in I}^{k} \sum_{j=1}^{k} \sum_{l=1} n_{ik_j} n_{ik_l} d_{k_j k_l} \tag{26}$$

- *Expected disagreement*:

$$D_e^\alpha = 2s^2_{total} = \frac{1}{ia(ia-1)} \sum_{j=1}^{k} \sum_{l=1}^{k} n_{k_j} n_{k_l} d_{k_j k_l} \tag{27}$$

- α: jede Instanz wird als separate Ebene in einer Varianzanalyse betrachtet
  - Anzahl an Ebenen (der Varianzanalyse) entspricht Anzahl an Instanzen
  - Anzahl an Beobachtungen entspricht der Anzahl an Annotatorinnen
- $Varianz_{within}$: Summe der quadrierten Unterschiede vom Mittelwert für eine Instanz, geteilt durch $df_{within} = i(a-1)$ (entspricht der Summe der quadrierten Unterschiede zwischen allen Beurteilungspaaren für diese Instanz, summiert über alle Instanzen)
- $Varianz_{total}$: Summe der quadrierten Unterschiede vom Gesamt-Mittelwert für alle Instanzen, geteilt durch $df_{total} = ia - 1$ (entspricht der Summe der quadrierten Unterschiede zwischen allen Beurteilungspaaren, summiert über alle Instanzen)

Gewichtetes Agreement - Krippendorff's α VI

$$\alpha = 1 - \frac{D_o}{D_e} \tag{28}$$

- Distanzfunktion kann durch beliebige Funktionen für verschiedene Skalen ersetzt werden
- Distanzfunktion für nominale Skalen

$$d_{ab} = \begin{cases} 0 \ \textit{wenn} \ k_a = k_b \\ 1 \ \textit{wenn} \ k_a \neq k_b \end{cases} \tag{29}$$

$\rightarrow \alpha$ ist (fast) equivalent zu multi-π

- Nachteil: schwer zu interpretieren (verschiedene Distanzfunktionen geben verschiedene Werte auf den gleichen Daten)

Maße für Inter-Annotator (Dis)Agreement - Fazit

- Verschiedene Maße mit verschiedenen Eigenschaften: percentage agreement, S, π, κ, multi-π, multi-κ, Krippendorff's α
- Wichtige Unterschiede:

Maß	**Zufalls-korr.**	**Unterscheidung Kat.**	**Unterscheidung Annot.**	**mehr als 2 Annot.**	**Gew.**
percentage agr.	-	-	-	-	-
S	x	-	-	-	-
π	x	x	-	-	-
κ	x	x	x	-	-
multi-π	x	x	-	x	-
multi-κ	x	x	x	x	-
Krippendorff's α	x	x	-	x	x