UNIVERSITÄT
ZU KÖLN

Annotation

VL Sprachliche Informationsverarbeitung

Nils Reiter
nils.reiter@uni-koeln.de

November 9, 2023
Winter term 2023/24

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Hausaufgabe 2

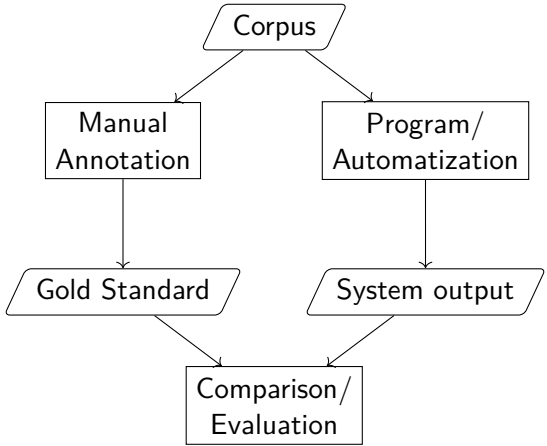
- Reden von Politiker:innen herunterladen
- Type-Token-Ratio berechnen
- Was kam raus?

Hausaufgabe 2

- Reden von Politiker:innen herunterladen
- Type-Token-Ratio berechnen
- Was kam raus?

Meine Kommentare zu den Ergebnissen

- Plattformen: Java (2x), Python (5x), Python Jupyter NB (8x), je einmal: Google Colab, HTML+Javascript, R, TypeScript
- Absolute Pfade in Programmcode ☹
- Encoding 😐
- Je länger der Text, desto geringer die TTR – deswegen besser STTR verwenden
- TTR *kann* eine interessante Unterschiede zeigen, aber meistens in Kombination mit anderen Indikatoren


Corpus
Manual Annotation
Program/ Automatization
Gold Standard
System output
Comparison/ Evaluation

Was sind Annotationen?

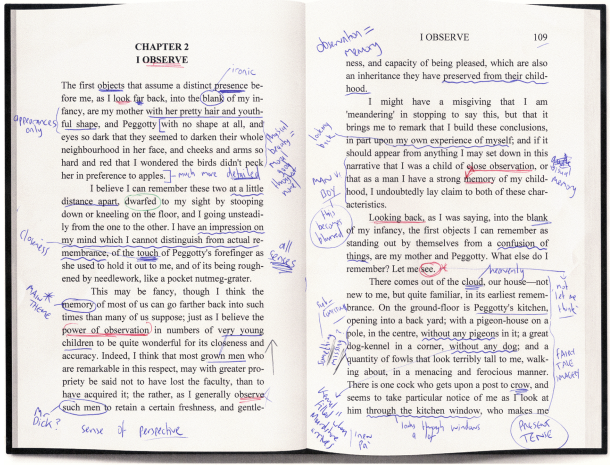
Annotation bedeutet "Anmerkung", "Beifügung", "Hinzufügung". In diesem Sinn haben Annotationen bei Stichworten, Begriffsklärungen oder ausführlichen Texten den Charakter der Erklärung beziehungsweise Ergänzung.

WP: Annotation, Version 134526826

Was sind Annotationen?

Annotation bedeutet "Anmerkung", "Beifügung", "Hinzufügung". In diesem Sinn haben Annotationen bei Stichworten, Begriffsklärungen oder ausführlichen Texten den Charakter der Erklärung beziehungsweise Ergänzung.

WP: Annotation, Version 134526826



CHAPTER 2
I OBSERVE

The first objects that assume a distinct presence before me, as I look far back, into the blank of my infancy, are my mother with her pretty hair and youthful shape, and Peggotty with no shape at all, and eyes so dark that they seemed to darken their whole neighbourhood in her face, and cheeks and arms so hard and red that I wondered the birds didn't peck her in preference to apples.

I believe I can remember these two at a little distance apart, dwarfed to my sight by stooping down or kneeling on the floor, and I going unsteadily from the one to the other. I have an impression on my mind which I cannot distinguish from actual remembrance, of the touch of Peggotty's forefinger as she used to hold it out to me, and of its being roughened by needlework, like a pocket nutmeg-grater.

This may be fancy, though I think the memory of most of us can go farther back into such times than many of us suppose; just as I believe the power of observation in numbers of very young children to be quite wonderful for its closeness and accuracy. Indeed, I think that most grown men who are remarkable in this respect, may with greater propriety be said not to have lost the faculty, than to have acquired it; the rather, as I generally observe such men to retain a certain freshness, and gentle-

ness, and capacity of being pleased, which are also an inheritance they have preserved from their childhood.

I might have a misgiving that I am 'meandering' in stopping to say this, but that it brings me to remark that I build these conclusions, in part upon my own experience of myself; and if it should appear from anything I may set down in this narrative that I was a child of close observation, or that as a man I have a strong memory of my childhood, I undoubtedly lay claim to both of these characteristics.

Looking back, as I was saying, into the blank of my infancy, the first objects I can remember as standing out by themselves from a confusion of things, are my mother and Peggotty. What else do I remember? Let me see.

There comes out of the cloud, our house—not new to me, but quite familiar, in its earliest remembrance. On the ground-floor is Peggotty's kitchen, opening into a back yard; with a pigeon-house on a pole, in the centre, without any pigeons in it; a great dog-kennel in a corner, without any dog; and a quantity of fowls that look terribly tall to me, walking about, in a menacing and ferocious manner. There is one cock who gets upon a post to crow, and seems to take particular notice of me as I look at him through the kitchen window, who makes me

Figure: Handschriftliche Annotationen auf Papier

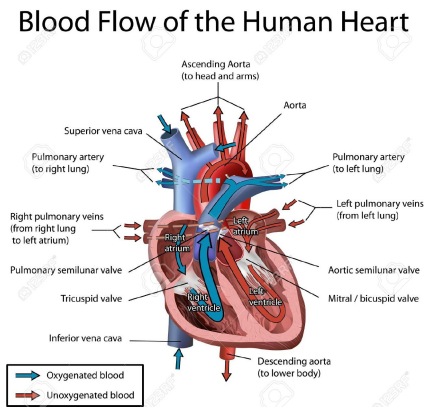


Figure: Bild-Annotationen

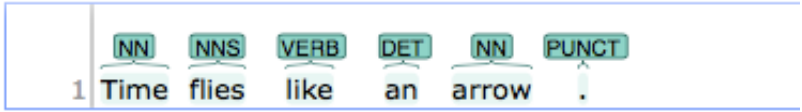


Figure: Digitale Annotationen von Parts of Speech

Annotationen

- Bestimmte Stellen des Werkes
- Verschiedene Arten
  - Werk: Text vs. Bild
  - Handschriftlich vs. Digital
  - Automatisch vs. Manuell
  - Frei vs. Fest
  - Subjektiv vs. Objektiv

Annotationen

- Bestimmte Stellen des Werkes
- Verschiedene Arten
  - Werk: Text vs. Bild
  - Handschriftlich vs. Digital
  - Automatisch vs. Manuell
  - Frei vs. Fest
  - Subjektiv vs. Objektiv

CL und DH

- Digital (ggf. Handschriftlich)
- Text-Annotationen
- Manuell und Automatisch
- Feste und freie Kategorien

Two motivations

Theory Development

- A number of theoretical answers, all with good arguments
- Let's see what people can distinguish in real texts

Two motivations

Theory Development

- A number of theoretical answers, all with good arguments
- Let's see what people can distinguish in real texts

Data creation

- Training/testing data for machine learning

Background Reading

Eduard Hovy/Julia Lavid (2010). "Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics". In: *International Journal of Translation Studies* 22.1

Nils Reiter (2020). "Anleitung zur Erstellung von Annotationsrichtlinien". In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 193–202. DOI: 10.1515/9783110693973-009

Empirische Validierung und Ergänzung von Theorien

- Theorien machen Aussagen über bestimmte Kategorien
  - "In narrativen Texten variiert die Erzählgeschwindigkeit"
  - "Artikel und Nomen bilden zusammen eine Nominalphrase"
- Annotation: Anwendung der Theorie auf einen Text

Empirische Validierung und Ergänzung von Theorien

- Theorien machen Aussagen über bestimmte Kategorien
  - "In narrativen Texten variiert die Erzählgeschwindigkeit"
  - "Artikel und Nomen bilden zusammen eine Nominalphrase"
- Annotation: Anwendung der Theorie auf einen Text
- Validierung
  - Annotator:innen können auf Kategorien hinweisen, die nicht von der Theorie abgedeckt sind
  - Systematisch verwechselte Kategorien sind mglw. unklar definiert
    - The Duke was pretty last night.
    - The Duchess was entertaining last night.
    - Adjektiv oder Verb (gerund)?

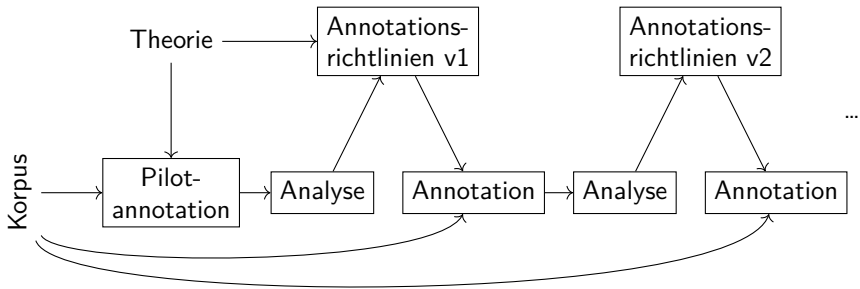
Empirische Validierung und Ergänzung von Theorien

- Annotationen als Mittel zum Zweck
  - Theorien müssen adaptiert werden, um quantitativ mit ihnen zu arbeiten
  - Adaptation: "Operationalisierung" Krautter et al. (2023)
  - Annotation kann ein Werkzeug dazu sein
    - Zwingt zur Genauigkeit
    - Erlaubt Vergleich von verschiedenen Lesarten der Theorie
- Wortarten: Weitgehend gelöst (STTS, Penn Treebank)

Datenerzeugung für automatische Verarbeitung

- Testdaten für automatische Verfahren
- Trainingsdaten für maschinelle Lernverfahren
  - $\rightarrow$ Trainiertes Modell einsetzbar auf neuen, noch nicht annotierten Daten

Workflow zu Annotationen



Parallele Annotationen

- Annotation der gleichen Text(stellen) durch verschiedene, unabhängige Annotator:innen
- Erlaubt Gegenprobe
- Erhöht insgesamt Zuverlässigkeit der Annotationen
- Deckt Probleme mit Kategorien/Richtlinien/Definitonen auf

Parallele Annotationen

- Annotation der gleichen Text(stellen) durch verschiedene, unabhängige Annotator:innen
- Erlaubt Gegenprobe
- Erhöht insgesamt Zuverlässigkeit der Annotationen
- Deckt Probleme mit Kategorien/Richtlinien/Definitonen auf

Wer annotiert?

Idealerweise: Unabhängige Personen

- Hiwis
- Crowd Sourcing
- Studierende

demo

Annotationsrichtlinien

- Instanziierung der Theorie
- Objektivierung
- Annotatorinnen sollen (ausschließlich) auf Basis der Richtlinien annotieren

Stuttgart-Tübingen Tagset (STTS)

- Annotationsrichtlinien für Wortarten, die in großen Projekten verwendet wurden
- 11 Oberkategorien (Haupt-tags):
  Nomen, Verben, Artikel, Adjektive, Pronomina, Kardinalzahlen, Adverbien, Konjunktionen, Adpositionen, Interjektionen, Partikeln

Hausaufgabe 3: Annotation!

- Ziel: Annotation eines linguistischen und eines nicht-linguistischen Phänomens
  - Linguistisch: Koreferenz
  - Nicht-Linguistisch: Indizien für problematisches Glücksspielverhalten in Online-Foren

    Smith et al. (2023)
- To do:
  - In Ilias einen Account "studentXX" aussuchen
  - Annotationsrichtlinien herunterladen
  - In Annotationstool Inception einloggen und
    - den Text Klauschke mit Koreferenz und
    - die anderen mit Indizien für "Problem Gambling" annotieren

Annotation Analysis

- Multiple annotators annotate the same text(s)
- Annotations are compared
- Disagreements can be quantified ("Inter-Annotator-Agreement", IAA)

Cohen, 1960; Fleiss, 1971; Fournier, 2013; Mathet et al., 2015,

Annotation Analysis

- Multiple annotators annotate the same text(s)
- Annotations are compared
- Disagreements can be quantified ("Inter-Annotator-Agreement", IAA)

  Cohen, 1960; Fleiss, 1971; Fournier, 2013; Mathet et al., 2015,
- Goal: High agreement
  - Based on the same guideline, different annotators should annotate the same categories
  - IAA: Metric for the quality of the annotation guidelines
    - Not: ...of the annotators

Annotation Analysis

- Multiple annotators annotate the same text(s)
- Annotations are compared
- Disagreements can be quantified ("Inter-Annotator-Agreement", IAA)

  Cohen, 1960; Fleiss, 1971; Fournier, 2013; Mathet et al., 2015,
- Goal: High agreement
  - Based on the same guideline, different annotators should annotate the same categories
  - IAA: Metric for the quality of the annotation guidelines
    - Not: ...of the annotators
- Manual inspection helpful, but not reliable
  - Researchers are *biased*: They have interests in a high agreement

Annotation Analysis

- Multiple annotators annotate the same text(s)
- Annotations are compared
- Disagreements can be quantified ("Inter-Annotator-Agreement", IAA)

  Cohen, 1960; Fleiss, 1971; Fournier, 2013; Mathet et al., 2015,
- Goal: High agreement
  - Based on the same guideline, different annotators should annotate the same categories
  - IAA: Metric for the quality of the annotation guidelines
    - Not: ...of the annotators
- Manual inspection helpful, but not reliable
  - Researchers are *biased*: They have interests in a high agreement
- Quantitative metrics: Free from personal influence

Inter-Annotator Agreement

First attempt

- Metric: Percentage agreement
- Portion of items for which annotators agree

Inter-Annotator Agreement

First attempt

- Metric: Percentage agreement
- Portion of items for which annotators agree

Example

	A1	A2
1	A	A
2	B	C
3	C	C
4	A	C
5	C	B

Inter-Annotator Agreement

First attempt

- Metric: Percentage agreement
- Portion of items for which annotators agree

Example

	A1	A2
1	A	A
2	B	C
3	C	C
4	A	C
5	C	B

- 5 instances, 2 annotators (A1, A2), 3 categories (A, B, C)
- Percentage agreement: $\frac{2}{5} = 0.2 = 20\%$

Inter-Annotator Agreement

First attempt

	A1	A2
1	A	A
2	B	C
3	C	C
4	A	C
5	C	B

- 5 instances,
  2 annotators (A1, A2),
  3 categories (A, B, C)
- Percentage agreement:
  $\frac{2}{5} = 0.2 = 20\%$

	A1	A2	A3
1	A	A	A
2	B	C	A
3	C	C	C
4	A	C	C
5	C	B	B

- 5 instances
  3 annotators (A1, A2, A3),
  3 categories (A, B, C)
- Percentage agreement:
  $\frac{2}{5} = 0.2 = 20\%$

Are they equally good?

Inter-Annotator Agreement

Data Structures

Word	A1	A2	A3
Zwei	Artikel (ART)	Artikel (ART)	Kardinalzahl (CARD)
Hunde	Nomen (NN)	Nomen (plural, NNS)	Nomen (plural, NNS)
bellen	Finites Vollverb (VVFIN)	Infinites Verb (VVINF)	Imperatives Verb (VVIMP)
.	$.	$.	$.

Inter-Annotator Agreement

Data Structures

Word	A1	A2	A3
Zwei	Artikel (ART)	Artikel (ART)	Kardinalzahl (CARD)
Hunde	Nomen (NN)	Nomen (plural, NNS)	Nomen (plural, NNS)
bellen	Finites Vollverb (VVFIN)	Infinites Verb (VVINF)	Imperatives Verb (VVIMP)
.	$.	$.	$.

⇓ conversion ⇓

Word	ART	CARD	NN	NNS	VVFIN	VVINF	VVIMP	$.
Zwei	2	1						
Hunde			1	2				
bellen					1	1	1	
.								3

Inter-Annotator Agreement

Pairs

Word	ART	CARD	NN	NNS	VVFIN	VVINF	VVIMP	$.
Zwei	2	1						
Hunde			1	2				
bellen					1	1	1	
.								3

- How many pairwise agreements are there?

Inter-Annotator Agreement

Pairs

Word	ART	CARD	NN	NNS	VVFIN	VVINF	VVIMP	$.
Zwei	2	1						
Hunde			1	2				
bellen					1	1	1	
.								3

- How many pairwise agreements are there?
  - $1 + 1 + 0 + 3 = 5$
- Why 3? binomial coefficient!

Binomial Coefficient / ‘n choose k’ / ‘n über k’

Number of sub sets with k elements from a set of n elements

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial Coefficient / 'n choose k' / 'n über k'

Number of sub sets with k elements from a set of n elements

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Factorial:

$$n! = n(n-1)(n-2)(n-3)\cdots 1$$

Binomial Coefficient / 'n choose k' / 'n über k'

Number of sub sets with k elements from a set of n elements

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Factorial:

$$n! = n(n-1)(n-2)(n-3)\cdots 1$$

for $k = 2$:

$$\binom{n}{2} = \frac{n!}{(n-2)!2!}$$

Binomial Coefficient / 'n choose k' / 'n über k'

Number of sub sets with k elements from a set of n elements

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Factorial:

$$n! = n(n-1)(n-2)(n-3)\cdots 1$$

for $k = 2$:

$$\binom{n}{2} = \frac{n!}{(n-2)!2!} = \frac{1}{2}\frac{n!}{(n-2)!}$$

Binomial Coefficient / 'n choose k' / 'n über k'

Number of sub sets with k elements from a set of n elements

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Factorial:

$$n! = n(n-1)(n-2)(n-3)\cdots 1$$

for $k = 2$:

$$\begin{aligned}\binom{n}{2} &= \frac{n!}{(n-2)!2!} = \frac{1}{2}\frac{n!}{(n-2)!} \\ &= \frac{1}{2}\frac{n(n-1)(n-2)(n-3)\cdots 1}{(n-2)(n-3)\cdots 1}\end{aligned}$$

Binomial Coefficient / 'n choose k' / 'n über k'

Number of sub sets with k elements from a set of n elements

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Factorial:

$$n! = n(n-1)(n-2)(n-3)\cdots 1$$

for $k = 2$:

$$\begin{aligned}\binom{n}{2} &= \frac{n!}{(n-2)!2!} = \frac{1}{2}\frac{n!}{(n-2)!} \\ &= \frac{1}{2}\frac{n(n-1)(n-2)(n-3)\cdots 1}{(n-2)(n-3)\cdots 1} \\ &= \frac{1}{2}n(n-1)\end{aligned}$$

Inter-Annotator Agreement

Problems

Situation 1:

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

- 3 annotators (=n)
- 5 pairwise agreements

Inter-Annotator Agreement

Problems

Situation 1:

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

- 3 annotators (=n)
- 5 pairwise agreements

Situation 2:

Word	A	B	C	D
1	2	1	1	
2		1	1	2
3	1	1	1	1
4	3	1		

- 4 annotators (=n)
- 5 pairwise agreements

Inter-Annotator Agreement

Problems

Situation 1:

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

- 3 annotators (=n)
- 5 pairwise agreements

Situation 2:

Word	A	B	C	D
1	2	1	1	
2		1	1	2
3	1	1	1	1
4	3	1		

- 4 annotators (=n)
- 5 pairwise agreements

How much worse is Situation 2 compared to 1?
$\rightarrow$ Scaling!

Inter-Annotator Agreement

Scaling

- Sometimes, values have different scales
  - i.e., different min and max values
- Scaling: Apply a function to values so that they are comparable
  - Simplest way: Divide by the (theoretical) maximum

Inter-Annotator Agreement

Scaling

- Sometimes, values have different scales
  - i.e., different min and max values
- Scaling: Apply a function to values so that they are comparable
  - Simplest way: Divide by the (theoretical) maximum
- What's the theoretical maximum here?
  - If all annotators agree: $\binom{n}{2} = \frac{1}{2}n(n-1)$ (on a single item)

Inter-Annotator Agreement

Situation 1:

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

- 3 annotators (=n)
- 5 pairwise agreements
- Scaled: $\frac{5}{4\binom{3}{2}} = \frac{5}{4\times 3} = \frac{5}{12} = 0.416$

Situation 2:

Word	A	B	C	D
1	2	1	1	
2		1	1	2
3	1	1	1	1
4	3	1		

- 4 annotators (=n)
- 5 pairwise agreements
- Scaled: $\frac{5}{4\binom{4}{2}} = \frac{5}{4\times 6} = \frac{5}{24} = 0.208$

Inter-Annotator Agreement

Observed Agreement

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

$N = 4, k = 4, n = 3$

Normalized observed agreement for item i

Problem: k categories, n annotators, N items

Inter-Annotator Agreement

Observed Agreement

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

$N = 4, k = 4, n = 3$

Normalized observed agreement for item i

Problem: k categories, n annotators, N items

$$\underbrace{\sum_{j=1}^{k} n_{i,j}(n_{i,j} - 1)}_{\text{abs. pairwise agr. for item } i}$$

Inter-Annotator Agreement

Observed Agreement

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

$N = 4, k = 4, n = 3$

Normalized observed agreement for item i

Problem: k categories, n annotators, N items

$$\underbrace{\frac{1}{n(n-1)}}_{\text{scaling for annotators}} \times \underbrace{\sum_{j=1}^{k} n_{i,j}(n_{i,j}-1)}_{\text{abs. pairwise agr. for item } i}$$

Inter-Annotator Agreement

Observed Agreement

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

$N = 4, k = 4, n = 3$

Normalized observed agreement for item i

Problem: k categories, n annotators, N items

$$O_i = \underbrace{\frac{1}{n(n-1)}}_{\text{scaling for annotators}} \times \underbrace{\sum_{j=1}^{k} n_{i,j}(n_{i,j}-1)}_{\text{abs. pairwise agr. for item } i}$$

Inter-Annotator Agreement

Observed Agreement

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

$N = 4, k = 4, n = 3$

Normalized observed agreement for item i

Problem: k categories, n annotators, N items

$$O_i = \underbrace{\frac{1}{n(n-1)}}_{\text{scaling for annotators}} \times \underbrace{\sum_{j=1}^{k} n_{i,j}(n_{i,j}-1)}_{\text{abs. pairwise agr. for item } i}$$

Normalized observed agreement for all items

$$O = \frac{1}{N}\sum_{i=1}^{N} \hat{P}_i \qquad \text{(arithmetic mean / average)}$$

Inter-Annotator Agreement

Expected Agreement

Situation 1:

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

- $n = 3$ annotators
- $k = 4$ categories
- 5 pairwise agreements

Situation 2:

Word	A	B	C
1	2	1	
2	2		1
3	1	1	1
4	3		

- $n = 3$ annotators
- $k = 3$ categories
- 5 pairwise agreements

What situation had the better agreement? How much better?

Inter-Annotator Agreement

Expected Agreement

Situation 1:

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

Situation 2:

Word	A	B	C
1	2	1	
2	2		1
3	1	1	1
4	3		

- $p(\text{"selecting category A"}) = \frac{\text{Number of positive events}}{\text{Number of possible events}} = \frac{\text{How often was category A selected?}}{\text{How many decisions were made / How often could it have been selected?}}$

Inter-Annotator Agreement

Expected Agreement

Situation 1:

Word	A	B	C	D
1	2	1		
2			1	2
3	1	1		1
4	3			

Situation 2:

Word	A	B	C
1	2	1	
2	2		1
3	1	1	1
4	3		

- $p(\text{"selecting category A"}) = \frac{\text{Number of positive events}}{\text{Number of possible events}} = \frac{\text{How often was category A selected?}}{\text{How many decisions were made / How often could it have been selected?}}$
- $p(\text{"selecting category A"}) =$
  - Situation 1: $\frac{6}{12}$
  - Situation 2: $\frac{8}{12}$

Inter-Annotator Agreement

Expected Agreement

- Probability that category j gets selected (by one annotator)

$$\underbrace{\sum_{i=1}^{N} n_{i,j}}_{\text{positive events (= annotations with cat. } j\text{)}}$$

Inter-Annotator Agreement

Expected Agreement

- Probability that category j gets selected (by one annotator)

$$\underbrace{\frac{1}{nN}}_{\text{Possible events (all annotations)}} \times \underbrace{\sum_{i=1}^{N} n_{i,j}}_{\text{positive events (= annotations with cat. } j\text{)}}$$

Inter-Annotator Agreement

Expected Agreement

- Probability that category j gets selected (by one annotator)

$$p_j = \underbrace{\frac{1}{nN}}_{\text{Possible events (all annotations)}} \times \underbrace{\sum_{i=1}^{N} n_{i,j}}_{\text{positive events (= annotations with cat. } j\text{)}}$$

Inter-Annotator Agreement

Expected Agreement

- Probability that category j gets selected (by one annotator)

$$p_j = \underbrace{\frac{1}{nN}}_{\text{Possible events (all annotations)}} \times \underbrace{\sum_{i=1}^{N} n_{i,j}}_{\text{positive events (= annotations with cat. } j\text{)}}$$

- Probability that two annotators select category j

$$p_j \times p_j = p_j^2$$

Inter-Annotator Agreement

Expected Agreement

- Probability that category j gets selected (by one annotator)

$$p_j = \underbrace{\frac{1}{nN}}_{\text{Possible events (all annotations)}} \times \underbrace{\sum_{i=1}^{N} n_{i,j}}_{\text{positive events (= annotations with cat. } j\text{)}}$$

- Probability that two annotators select category j

$$p_j \times p_j = p_j^2$$

- Probability that two annotators are in agreement (over all categories):

$$E = \sum_{j=1}^{k} p_j^2$$

Inter-Annotator Agreement

Expected and Observed Agreement

- Expected agreement
  - What level of agreement do we get by chance?
  - Statements about categories – calculated over columns
- Observed agreement
  - What level of agreement did we actually get?
  - Statements about items – calculated over rows

Inter-Annotator Agreement

Expected and Observed Agreement

- Expected agreement
  - What level of agreement do we get by chance?
  - Statements about categories – calculated over columns
- Observed agreement
  - What level of agreement did we actually get?
  - Statements about items – calculated over rows
- We are interested in the agreement *above chance*

Fleiss' Kappa (Fleiss, 1971)

p_j	Probability for j
O_i	Observed agreement for i

Fleiss' Kappa (Fleiss, 1971)

p_j Probability for j

O_i Observed agreement for i

$$O = \frac{1}{N}\sum_{i=1}^{N} O_i$$

Fleiss' Kappa (Fleiss, 1971)

p_j Probability for j

O_i Observed agreement for i

$$O = \frac{1}{N}\sum_{i=1}^{N} O_i$$

$$E = \sum_{j=1}^{k} p_j^2$$

Fleiss' Kappa (Fleiss, 1971)

p_j Probability for j

O_i Observed agreement for i

$$O = \frac{1}{N}\sum_{i=1}^{N} O_i$$

$$E = \sum_{j=1}^{k} p_j^2$$

$$\kappa = \frac{O - E}{1 - E}$$

Fleiss' Kappa (Fleiss, 1971)

$$
\begin{aligned}
p_j &\quad \text{Probability for } j \\
O_i &\quad \text{Observed agreement for } i \\
O &= \frac{1}{N}\sum_{i=1}^{N} O_i \\
E &= \sum_{j=1}^{k} p_j^2 \\
\kappa &= \frac{O-E}{1-E}
\end{aligned}
$$

- $O - E$: Tatsächlich erreichtes, nicht-zufälliges Agreement
- $1 - E$: Maximal erreichbares, nicht-zufälliges Agreement

Fleiss' Kappa (Fleiss, 1971)

$$
\begin{aligned}
p_j &\quad \text{Probability for } j \\
O_i &\quad \text{Observed agreement for } i \\
O &= \frac{1}{N}\sum_{i=1}^{N} O_i \\
E &= \sum_{j=1}^{k} p_j^2 \\
\kappa &= \frac{O-E}{1-E}
\end{aligned}
$$

- $O - E$: Tatsächlich erreichtes, nicht-zufälliges Agreement
- $1 - E$: Maximal erreichbares, nicht-zufälliges Agreement
- $-\infty < \kappa < 1$: Je höher desto besser
  - Extremfälle?

Fleiss' Kappa (Fleiss, 1971)

Boundary cases

Example (best case scenario)

Word	A	B	C
1	3		
2		3	
3	3		
4			3

Fleiss' Kappa (Fleiss, 1971)

Boundary cases

Example (best case scenario)

Word	A	B	C
1	3		
2		3	
3	3		
4			3

Observed agreement:

$$O_1 = O_2 = O_3 = O_4 = \frac{3}{3} = 1$$

$$O = 1$$

Fleiss' Kappa (Fleiss, 1971)

Boundary cases

Example (best case scenario)

Word	A	B	C
1	3		
2		3	
3	3		
4			3

Observed agreement:

$$O_1 = O_2 = O_3 = O_4 = \frac{3}{3} = 1$$

$$O = 1$$

Expected agreement:

$$p_1 = \frac{1}{12}(3 + 0 + 3 + 0) = \frac{1}{2}$$

$$p_2 = p_3 = \frac{1}{4} = 0.25$$

$$E = \frac{1}{4}(p_1^2 + p_2^2 + p_3^2) = 0.09375$$

Fleiss' Kappa (Fleiss, 1971)

Boundary cases

Example (best case scenario)

Word	A	B	C
1	3		
2		3	
3	3		
4			3

Observed agreement:

$$O_1 = O_2 = O_3 = O_4 = \frac{3}{3} = 1$$

$$O = 1$$

Expected agreement:

$$p_1 = \frac{1}{12}(3+0+3+0) = \frac{1}{2}$$

$$p_2 = p_3 = \frac{1}{4} = 0.25$$

$$E = \frac{1}{4}(p_1^2 + p_2^2 + p_3^2) = 0.09375$$

Fleiss' Kappa:

$$\kappa = \frac{1 - 0.09375}{1 - 0.09375} = 1$$

Fleiss' Kappa (Fleiss, 1971)

Boundary cases

Example (worst case scenario)

Word	A	B	C
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1

Observed agreement:

$$O_1 = O_2 = O_3 = O_4 = \frac{0}{3} = 0$$

$$O = 0$$

Fleiss' Kappa (Fleiss, 1971)

Boundary cases

Example (worst case scenario)

Word	A	B	C
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1

Observed agreement:

$$O_1 = O_2 = O_3 = O_4 = \frac{0}{3} = 0$$

$$O = 0$$

Expected agreement:

$$p_1 = p_2 = p_3 = \frac{1}{12}(1+1+1+1) = \frac{1}{3}$$

$$E = \frac{1}{4}(p_1^2 + p_2^2 + p_3^2) = 0.083333333$$

Fleiss' Kappa (Fleiss, 1971)

Boundary cases

Example (worst case scenario)

Word	A	B	C
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1

Observed agreement:

$$O_1 = O_2 = O_3 = O_4 = \frac{0}{3} = 0$$

$$O = 0$$

Expected agreement:

$$p_1 = p_2 = p_3 = \frac{1}{12}(1+1+1+1) = \frac{1}{3}$$

$$E = \frac{1}{4}(p_1^2 + p_2^2 + p_3^2) = 0.083333333$$

Fleiss' Kappa:

$$\kappa = \frac{0 - 0.083333333}{1 - 0.083333333} = -0.090909091$$

Summary

- Annotation
  - to operationalise theoretic concepts
  - to create reference data for training and testing
- Annotation guidelines describe how to annotate in practice
  - Developing them is usually an iterative process
  - Applying them is a lot work and time-consuming
- Annotation analysis
  - Talking to annotators, comparing results manually
  - Quantification: Inter-annotator agreement