Reliability of Annotations

- The performance of an algorithm has to be evaluated against some kind of correct solution, the *key*.
- For most linguistic tasks *correct* can be defined using human performance (not linguists intuition!).
- However, if different humans get different solutions for the same task, it is questionable which solution is correct and whether the task can be solved by humans at all.
- Therefore, measures of reliability have to be used to test whether human performance is reliable.
- If human performance is indeed reliable, the solution produced by human can be used as a key against which an algorithm can be evaluated.

How to Measure Reliability?

- Kowtko, Isard, Doherty (1992) and Litman & Hirschberg (1990) use pairwise agreement between naive annotators.

- Silverman et al. (1992) have two groups of annotators: a small group of experienced annotators and a larger group of annotators with less experience. – Silverman et al. (1992) argue that the annotations are reliable, if there is only a small difference between the groups.

However, what does reliability mean in these cases?

Agreement

	A	B	C	S
1	2	0	0	1
2	2	0	0	1
3	2	0	0	1
4	0	2	0	1
5	0	2	0	1
6	0	2	0	1
7	0	0	2	1
8	0	0	2	1
9	0	0	2	1
10	1	1	0	0
	7	7	6	9

Balanced distribution

N = 10
T = 20
Z = 9

PA = $\frac{Z}{N}$ = $\frac{9}{10}$ = 0,9

What does an agreement of 90% mean?

Agreement

	A	B	C	S
1	2	0	0	1
2	2	0	0	1
3	2	0	0	1
4	2	0	0	1
5	2	0	0	1
6	2	0	0	1
7	2	0	0	1
8	2	0	0	1
9	2	0	0	1
10	0	1	1	0
	18	1	1	9

Skewed distribution

N = 10
T = 20
Z = 9

PA = $\frac{Z}{N}$ = $\frac{9}{10}$ = 0,9

Agreement by chance not considered!

Agreement by chance

	A	B	C	S
1	2	0	0	1
2	2	0	0	1
3	2	0	0	1
4	0	2	0	1
5	0	2	0	1
6	0	2	0	1
7	0	0	2	1
8	0	0	2	1
9	0	0	2	1
10	1	1	0	0
	7	7	6	9

Balanced distribution

N = 10

T = 20

Z = 9

PA = $\frac{Z}{N}$ = $\frac{9}{10}$ = $0,9$

PE = $(\frac{A}{T})^2+(\frac{B}{T})^2+(\frac{C}{T})^2$ = $(\frac{7}{20})^2+(\frac{7}{20})^2+(\frac{6}{20})^2$ = $\frac{134}{400}$ = $0,335$

Agreement by chance

	A	B	C	S
1	2	0	0	1
2	2	0	0	1
3	2	0	0	1
4	2	0	0	1
5	2	0	0	1
6	2	0	0	1
7	2	0	0	1
8	2	0	0	1
9	2	0	0	1
10	0	1	1	0
	18	1	1	9

Skewed distribution

N = 10

T = 20

Z = 9

PA = $\frac{Z}{N}$ = $\frac{9}{10}$ = $0,9$

PE = $(\frac{A}{T})^2+(\frac{B}{T})^2+(\frac{C}{T})^2$ = $(\frac{18}{20})^2+(\frac{1}{20})^2+(\frac{1}{20})^2$ = $\frac{326}{400}$ = $0,815$

We look for a statistic/measure which considers agreement between annotators as well as agreement by chance.

The Kappa Statistic as a Measure of Reliability

- The kappa statistic (Cohen 1960, Siegel & Castellan 1988, Carletta 1996) can be used when multiple annotators have to assign markables to one of a set of nonordered classes.

- Kappa computes a coefficient among annotators and takes into account the chance agreement (which makes it far more suitable than just computing the level of agreement in percent).

- Kappa is defined as:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

  where *P(A)* is the actual agreement between annotators, *P(E)* the agreement by chance.

Kappa

	A	B	C	S
1	2	0	0	1
2	2	0	0	1
3	2	0	0	1
4	0	2	0	1
5	0	2	0	1
6	0	2	0	1
7	0	0	2	1
8	0	0	2	1
9	0	0	2	1
10	1	1	0	0
	7	7	6	9

Balanced distribution

N = 10
T = 20
Z = 9

PA = $\frac{Z}{N}$ = $\frac{9}{10}$ = $0,9$

PE = $(\frac{A}{T})^2+(\frac{B}{T})^2+(\frac{C}{T})^2$ = $(\frac{7}{20})^2+(\frac{7}{20})^2+(\frac{6}{20})^2$ = $\frac{134}{400}$ = $0,335$

K = $\frac{PA-PE}{1-PE}$ = $\frac{0,9-0,335}{1-0,335}$ = $\frac{0,565}{0,665}$ = $0,85$

Kappa

	A	B	C	S
1	2	0	0	1
2	2	0	0	1
3	2	0	0	1
4	2	0	0	1
5	2	0	0	1
6	2	0	0	1
7	2	0	0	1
8	2	0	0	1
9	2	0	0	1
10	0	1	1	0
	18	1	1	9

Skewed distribution

N = 10
T = 20
Z = 9

PA = $\frac{Z}{N}$ = $\frac{9}{10}$ = $0,9$

PE = $(\frac{A}{T})^2+(\frac{B}{T})^2+(\frac{C}{T})^2$ = $(\frac{18}{20})^2+(\frac{1}{20})^2+(\frac{1}{20})^2$ = $\frac{326}{400}$ = $0,815$

K = $\frac{PA-PE}{1-PE}$ = $\frac{0,9-0,815}{1-0,815}$ = $\frac{0,085}{0,185}$ = $0,46$

Three Annotators

	A	B	C	S
1	3	0	0	1
2	3	0	0	1
3	3	0	0	1
4	0	3	0	1
5	0	3	0	1
6	0	3	0	1
7	0	0	3	1
8	0	0	3	1
9	0	0	3	1
10	1	1	1	0
	10	10	10	9

Balanced Distribution

N = 10
T = 30
Z = 9

PA = $\frac{Z}{N} = \frac{9}{10} = 0,9$

PE = $(\frac{A}{T})^2 + (\frac{B}{T})^2 + (\frac{C}{T})^2 = (\frac{10}{30})^2 + (\frac{10}{30})^2 + (\frac{10}{30})^2 = \frac{300}{900} = 0,\bar{3}$

K = $\frac{PA - PE}{1 - PE} = \frac{0,9 - 0,\bar{3}}{1 - 0,\bar{3}} = \frac{0,5\bar{6}}{0,\bar{6}} = 0,85$

Three Annotators

	A	B	C	S
1	3	0	0	1
2	3	0	0	1
3	3	0	0	1
4	3	0	0	1
5	3	0	0	1
6	3	0	0	1
7	3	0	0	1
8	3	0	0	1
9	3	0	0	1
10	1	1	1	0
	28	1	1	9

Skewed distribution

N = 10
T = 30
Z = 9

PA = $\frac{Z}{N} = \frac{9}{10} = 0,9$

PE = $(\frac{A}{T})^2 + (\frac{B}{T})^2 + (\frac{C}{T})^2 = (\frac{28}{30})^2 + (\frac{1}{30})^2 + (\frac{1}{30})^2 = \frac{786}{900} = 0,87\bar{3}$

K = $\frac{PA - PE}{1 - PE} = \frac{0,9 - 0,875}{1 - 0,875} = \frac{0,02\bar{6}}{0,12\bar{6}} = 0,21$

Further Notes on Kappa

- When there is complete agreement between annotators, then $K = 1$. If there is no agreement besides chance agreement, then $K = 0$.

- In the field of content analysis (Krippendorf 1980), $K > 0.8$ indicates good reliability, $0.68 \leq K \leq 0.8$ allows to draw tentative conclusion.

- In particular for small datasets the significance of the values computed by kappa should be reported (see Fleiss (1971), Siegel & Castellan (1988) for the formula).

- Passonneau (1997) showed how to apply kappa to the problem of coreference resolution, i.e., how to measure agreement among annotators assigning markables to coreference classes.

Example: Sortal Classes in Texts

- Task: Assign sortal classes to noun phrases in texts of different genres
- First attempt: Ten different classes which were manually annotated.

Example: Sortal Classes in Texts

Person	one or more human beings
Group	institutionalized group of human beings
PhysObj	physical object
Concept	abstract concept
Loc	geographical location
Time	date, time span
Event	sth. which takes place in space and time
Action	sth. which is done
State	state of affairs, feeling, . . .
Property	characteristic or attribute of sth.

Sortal Classes – CG11

exprtype file: /home/strube/refer/bin/.ref.types.naacl00
attribute class: Sortal Class
files: .ms .mwo

	none	Pers	Group	Loc	Time	PhysObj	Event	Act	State	Prop	Concept	S
1	0	0	0	0	0	2	0	0	0	0	0	1
2	0	0	0	0	0	1	0	0	0	0	1	0
3	0	0	0	0	0	2	0	0	0	0	0	1
4	0	0	0	2	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	2	0	1
6	0	0	0	1	0	1	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	2	1
8	0	2	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0	2	1
10	0	0	0	0	1	0	0	0	1	0	0	0

Kappa for 10 Sortal Classes

$$N = 50$$

$$T = 100$$
$$Z = 30$$

$$PA = \frac{30}{50} = 0,6$$

$$PE = (\frac{15}{100})^2 + (\frac{14}{100})^2 + \ldots + (\frac{8}{100})^2 = \frac{1712}{10000} = 0,1712$$

$$K = \frac{0,6-0,1712}{1-0,1712} = 0,5174$$

Second Attempt: Three Sortal Classes

Since the annotations for ten sortal classes were not reliable we combined several classes to one:

Person: Person, Group

PhysObj: PhysObj, Loc

Abstract: Concept, Time, Event, Action, State, Property

Sortal Classes, just 3 – CG11

exprtype file: /home/strube/refer/bin/.ref.types.naacl00
attribute class: Sortal Class
files: .ms .mwo

		none	Pers	PhysObj	Abs	S
1	10030001008	0	0	2	0	1
2	10240001033	0	0	1	1	0
3	10360001040	0	0	2	0	1
4	10470002034	0	0	0	2	1
5	30000003055	0	0	0	2	1
6	30370003055	0	0	1	1	0
7	30590005035	0	0	0	2	1
8	40140004037	0	2	0	0	1
9	40540004061	0	0	0	2	1
10	40650005035	0	0	0	2	1

Kappa for 3 Sortal Classes

$$N = 50$$

$$T = 100$$
$$Z = 44$$

$$PA = \frac{44}{50} = 0,88$$

$$PE = (\frac{29}{100})^2 + (\frac{16}{100})^2 + (\frac{55}{100})^2 = \frac{4122}{10000} = 0,4122$$

$$K = \frac{0,88-0,4122}{1-0,4122} = 0,7958$$