

Unit 9: Inter-Annotator Agreement

Statistics for Linguists with R – A SIGIL Course

Designed by Stefan Evert¹ and Marco Baroni²

¹Computational Corpus Linguistics Group
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

²Center for Mind/Brain Sciences (CIMeC)
University of Trento, Italy

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

Introduction

Manually annotated data will be used for ...

1. Linguistic analysis

- ▶ Which factors determine a certain choice or interpretation?
- ▶ Are there syntactic correlates of the container-content relation?

2. Machine learning (ML)

- ▶ Automatic semantic annotation, e.g. for text mining
- ▶ Extend WordNet with new entries & relations
- ▶ Online semantic analysis in NLP pipeline (e.g. dialogue system)

Crucial issue: **Are the annotations correct?**

☞ ML learns to make same mistakes as human annotator

☞ Inconclusive & misleading results from linguistic analysis

Validity vs. reliability

(terminology from Artstein & Poesio 2008)

- ▶ We are interested in the **validity** of the manual annotation
 - ▶ i.e. whether the annotated categories are **correct**
- ▶ But there is no “ground truth”
 - ▶ Linguistic categories are determined by human judgement
 - ▶ Consequence: we cannot measure correctness directly
- ▶ Instead measure **reliability** of annotation
 - ▶ i.e. whether human coders¹ consistently make same decisions
 - ▶ Assumption: high reliability implies validity
- ▶ How can reliability be determined?

¹The terms “annotator” and “coder” are used interchangeably in this talk.

Easy & hard tasks

(Brants 2000 for German POS/syntax, Véronis 1998 for WSD)

Objective tasks

- ▶ Decision rules, linguistic tests
- ▶ Annotation guidelines with discussion of boundary cases
- ▶ POS tagging, syntactic annotation, segmentation, phonetic transcription, ...

➡ IAA = 98.5% (POS tagging)
IAA ≈ 93.0% (syntax)

Subjective tasks

- ▶ Based on speaker intuitions
- ▶ Short annotation instructions
- ▶ Lexical semantics (subjective interpretation!), discourse annotation & pragmatics, subjectivity analysis, ...

➡ IAA = $\frac{48}{70} = 68.6\%$ (HW)
IAA ≈ 70% (word senses)

[NB: error rates around 5% are considered acceptable for most purposes]

Inter-annotator agreement

- ▶ Multiple coders annotate same data (with same guidelines)
- ▶ Calculate **Inter-annotator agreement (IAA)**

Sentence	A	B	agree?
Put tea in a heat-resistant jug and add the boiling water.	yes	yes	✓
Where are the batteries kept in a phone ?	no	yes	✗
Vinegar's usefulness doesn't stop inside the house .	no	no	✓
How do I recognize a room that contains radioactive materials ?	yes	yes	✓
A letterbox is a plastic, screw-top bottle that contains a small notebook and a unique rubber stamp.	yes	no	✗

➡ Observed agreement between A and B is 60%

👉 Is 70% agreement good enough?

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures


Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

But 90% agreement is certainly a good result?

 i.e. it indicates high reliability

Thought experiment 1

- ▶ Assume that A and B are lazy annotators, so they just marked sentences randomly as “yes” and “no”

[or they enjoyed too much sun & Bordeaux wine yesterday]

- ▶ How much agreement would you expect?

- ▶ Annotator decisions are like coin tosses:

25% both coders randomly choose “yes” ($= 0.5 \cdot 0.5$)

25% both coders randomly choose “no” ($= 0.5 \cdot 0.5$)

50% agreement purely by chance

- ➡ IAA = 70% is only mildly better than chance agreement

Thought experiment 2

- ▶ Assume A and B are lazy coders with a proactive approach

- ▶ They believe that their task is to find as many examples of container-content pairs as possible to make us happy

- ▶ So they mark 95% of sentences with “yes”

- ▶ But individual choices are still random

- ▶ How much agreement would you expect now?

- ▶ Annotator decisions are like tosses of a biased coin:

90.25% both coders randomly choose “yes” ($= .95 \cdot .95$)

0.25% both coders randomly choose “no” ($= .05 \cdot .05$)

90.50% agreement purely by chance

- ➡ IAA = 90% might be no more than chance agreement

Measuring inter-annotator agreement

(notation follows Artstein & Poesio 2008)

Agreement measures must be corrected for **chance agreement!**
(for computational linguistics: Carletta 1996)

Notation: A_o ... observed (or "percentage") agreement
 A_e ... expected agreement by chance

General form of chance-corrected agreement measure R :

$$R = \frac{A_o - A_e}{1 - A_e}$$

Measuring inter-annotator agreement

Some general properties of R :

- ▶ Perfect agreement: $R = 1 = \frac{1 - A_e}{1 - A_e}$
- ▶ Chance agreement: $R = 0 = \frac{A_e - A_e}{1 - A_e}$
- ▶ Perfect disagreement: $R = \frac{-A_e}{1 - A_e}$

Various agreement measures depending on precise definition of A_e :

- ▶ $R = S$ for random coin tosses (Bennett *et al.* 1954)
- ▶ $R = \pi$ for shared category distribution (Scott 1955)
- ▶ $R = \kappa$ for individual category distributions (Cohen 1960)

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

Contingency tables for annotator agreement

coder A	coder B		
	yes	no	
yes	24	8	32
no	14	24	38
	38	32	70

coder A	coder B		
	yes	no	
yes	n_{11}	n_{12}	$n_{1\cdot}$
no	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	N

coder A	coder B		
	yes	no	
yes	.343	.114	.457
no	.200	.343	.543
	.543	.457	1

coder A	coder B		
	yes	no	
yes	p_{11}	p_{12}	$p_{1\cdot}$
no	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	p

Contingency tables for annotator agreement

Contingency table of **proportions** $p_{ij} = \frac{n_{ij}}{N}$

coder A	coder B		
	yes	no	
yes	.343	.114	.457
no	.200	.343	.543
	.543	.457	1

coder A	coder B		
	yes	no	
yes	p_{11}	p_{12}	$p_{1\cdot}$
no	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	p

Relevant information can be read off from contingency table:

- Observed agreement $A_o = p_{11} + p_{22} = .686$
- Category distribution for coder A: $p_{i\cdot} = p_{i1} + p_{i2}$
- Category distribution for coder B: $p_{\cdot j} = p_{1j} + p_{2j}$

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

Calculating the expected chance agreement

- How often are annotators expected to agree if they make random choices according to their category distributions?
- Decisions of annotators are independent → multiply marginals

coder A	coder B		
	yes	no	
yes	.248	.209	.457
no	.295	.248	.543
	.543	.457	1

coder A	coder B		
	yes	no	
yes	$p_{1\cdot} \cdot p_{\cdot 1}$	$p_{1\cdot} \cdot p_{\cdot 2}$	$p_{1\cdot}$
no	$p_{2\cdot} \cdot p_{\cdot 1}$	$p_{2\cdot} \cdot p_{\cdot 2}$	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	p

→ Expected chance agreement:

$$A_e = p_{1\cdot} \cdot p_{\cdot 1} + p_{2\cdot} \cdot p_{\cdot 2} = 49.6\%$$

Sanity check: Is it plausible to assume that annotators always flip coins?

- No need to make such strong assumptions
- Annotations of individual coders may well be systematic
- We only require that choices of A and B are **statistically independent**, i.e. no common ground for their decisions

Definition of the Kappa coefficient

(Cohen 1960)

Formal definition of the **Kappa** coefficient:

$$A_o = p_{11} + p_{22}$$

$$A_e = p_{1\cdot} \cdot p_{\cdot 1} + p_{2\cdot} \cdot p_{\cdot 2}$$

$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

In our example:

$$A_o = .343 + .343 = .686$$

$$A_e = .248 + .248 = .496$$

$$\kappa = \frac{.686 - .496}{1 - .496} = 0.376 !!$$

Other agreement measures

(Scott 1955; Bennett *et al.* 1954)

1. π estimates a common category distribution \bar{p}_i
 - ▶ goal is to measure chance agreement between arbitrary coders, while κ focuses on a specific pair of coders

$$A_e = (\bar{p}_1)^2 + (\bar{p}_2)^2$$

$$\bar{p}_i = \frac{1}{2}(p_{i\cdot} + p_{\cdot i})$$

2. S assumes that coders actually flip coins ...

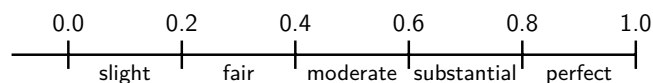
- ▶ i.e. equiprobable category distribution $\bar{p}_1 = \bar{p}_2 = \frac{1}{2}$

$$A_e = \frac{1}{2}$$

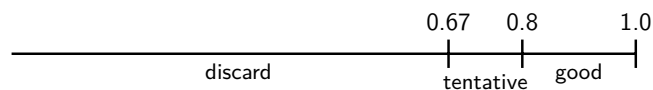
Much controversy whether π or κ is the more appropriate measure, but in practice they often lead to similar agreement values!

Scales for the interpretation of Kappa

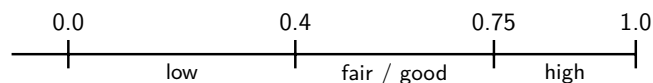
- ▶ Landis & Koch (1977)



- ▶ Krippendorff (1980)



- ▶ Green (1997)



- ▶ and many other suggestions ...

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

An example from Di Eugenio & Glass (2004)

coder A	coder B		
	yes	no	
yes	70	25	95
no	0	55	55
	70	80	150

coder A	coder B		
	yes	no	
yes	.467	.167	.633
no	.000	.367	.367
	.467	.533	1

- Cohen (1960): $A_o = .833$, $A_e = .491$, $\kappa = .672$
- Scott (1955): $A_o = .833$, $A_e = .505$, $\pi = .663$
- Krippendorff (1980): data show tentative agreement according to κ , but should be discarded according to π

What do you think?

More samples from the same annotators ...

coder A	coder B		
	yes	no	
yes	67	24	91
no	2	57	59
	69	81	150

$$A_0 = .827$$

$$\kappa = .659 \quad (A_e = .491)$$

$$\pi = .652 \quad (A_e = .502)$$

More samples from the same annotators ...

coder A	coder B		
	yes	no	
yes	70	20	90
no	4	56	60
	74	76	150

$$A_0 = .840$$

$$\kappa = .681 \quad (A_e = .499)$$

$$\pi = .677 \quad (A_e = .504)$$

We are not interested in a particular sample, but rather want to know how often coders agree in general (for this task).

► **Sampling variation** of κ

[NB: A_e is *expected* chance agreement, not value in specific sample]

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

Kappa is a sample statistic $\hat{\kappa}$

	+	-
+	π_{11}	π_{12}
-	π_{21}	π_{22}

population

$$\alpha_o = \pi_{11} + \pi_{12}$$

$$\alpha_e = \pi_{1\cdot} \cdot \pi_{\cdot 1} + \pi_{2\cdot} \cdot \pi_{\cdot 2}$$

$$\kappa = \frac{\alpha_o - \alpha_e}{1 - \alpha_e}$$

	+	-
+	p_{11}	p_{12}
-	p_{21}	p_{22}

sample

$$A_o = p_{11} + p_{12}$$

$$A_e = p_{1\cdot} \cdot p_{\cdot 1} + p_{2\cdot} \cdot p_{\cdot 2}$$

$$\hat{\kappa} = \frac{A_o - A_e}{1 - A_e}$$

Sampling variation of $\hat{\kappa}$

(Fleiss *et al.* 1969; Krenn *et al.* 2004)

- ▶ Standard approach: show (or hope) that $\hat{\kappa}$ approximately follows Gaussian distribution if samples are large enough
- ▶ Show (or hope) that $\hat{\kappa}$ is unbiased estimator: $E[\hat{\kappa}] = \kappa$
- ▶ Compute standard deviation of $\hat{\kappa}$ (Fleiss *et al.* 1969: 325):

$$(\hat{\sigma}_{\hat{\kappa}})^2 = \frac{1}{N \cdot (1 - A_e)^4} \cdot \left(\sum_{i=1}^2 p_{ii} [(1 - A_e) - (p_{\cdot i} + p_{i\cdot})(1 - A_o)]^2 + (1 - A_o)^2 \sum_{i \neq j} p_{ij} (p_{\cdot i} + p_{j\cdot})^2 - (A_o A_e - 2A_e + A_o)^2 \right)$$

Sampling variation of $\hat{\kappa}$

(Lee & Tu 1994; Boleda & Evert unfinished)

- ▶ Asymptotic 95% confidence interval:

$$\kappa \in [\hat{\kappa} - 1.96 \cdot \hat{\sigma}_{\hat{\kappa}}, \hat{\kappa} + 1.96 \cdot \hat{\sigma}_{\hat{\kappa}}]$$

- ▶ For the example from Di Eugenio & Glass (2004), we have

$$\kappa \in [0.562, 0.783] \quad \text{with} \quad \hat{\sigma}_{\hat{\kappa}} = .056$$

➡ comparison with threshold .067 is pointless!

- ▶ How accurate is the Gaussian approximation?
 - ▶ Simulation experiments indicate biased $\hat{\kappa}$, underestimation of $\hat{\sigma}_{\hat{\kappa}}$ and non-Gaussian distribution for skewed marginals
 - ▶ Confidence intervals are reasonable for larger samples
- ▶ Recent work on improved estimates (e.g. Lee & Tu 1994)

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

Extensions of Kappa: Multiple categories

- ▶ Straightforward extension to $C > 2$ categories
→ $C \times C$ contingency table of proportions p_{ij}
- ▶ Observed agreement: $A_o = \sum_{i=1}^C p_{ii}$
- ▶ Expected agreement: $A_e = \sum_{i=1}^C p_{i\cdot} \cdot p_{\cdot i}$
- ▶ Kappa: $\hat{\kappa} = \frac{A_o - A_e}{1 - A_e}$
- ▶ Equation for $\hat{\sigma}_{\hat{\kappa}}$ also extends to C categories
- ▶ Drawback: $\hat{\kappa}$ only uses diagonal and marginals of table, discarding most information from the off-diagonal cells

Extensions of Kappa: Weighted Kappa

- ▶ For multiple categories, some disagreements may be more “serious” than others → assign greater weight
- ▶ E.g. German PP-verb combinations (Krenn *et al.* 2004)
 1. figurative expressions (collocational)
 2. support-verb constructions (collocational)
 3. free combinations (non-collocational)
- ▶ Rewrite $\hat{\kappa}$ in terms of expected/observed **disagreement**

$$\hat{\kappa} = \frac{(1 - D_o) - (1 - D_e)}{1 - (1 - D_e)} = 1 - \frac{D_o}{D_e}$$

$$D_o = 1 - A_o = \sum_{i \neq j} p_{ij} \rightsquigarrow \sum_{i \neq j} w_{ij} p_{ij}$$

$$D_e = 1 - A_e = \sum_{i \neq j} p_{i\cdot} \cdot p_{\cdot j} \rightsquigarrow \sum_{i \neq j} w_{ij} (p_{i\cdot} \cdot p_{\cdot j})$$

Extensions of Kappa: Multiple annotators

(Krenn *et al.* 2004)

- ▶ Naive strategy: compare each annotator against selected “expert”, or consensus annotation after reconciliation phase

BK vs. NN	kappa value	homogeneity		interval size
		min	max	
7	.775	71.93%	82.22%	10.29
9	.747	68.65%	79.77%	11.12
10	.700	64.36%	75.85%	11.49
4	.696	64.09%	75.91%	11.82
1	.692	63.39%	75.91%	12.52
6	.671	61.05%	73.33%	12.28
5	.669	60.12%	72.75%	12.63
2	.639	56.14%	70.64%	14.50
11	.592	52.40%	65.65%	13.25
3	.520	51.70%	64.33%	12.63
8	.341	33.68%	49.71%	16.03
12	.265	17.00%	35.05%	18.05

Extensions of Kappa: Multiple annotators

- ▶ Better approach: compute $\hat{\kappa}$ for each possible pair of annotators, then report average and standard deviation
- ▶ Extensions of agreement coefficients to multiple annotators are mathematical implementations of this basic idea (see Artstein & Poesio 2008 for details)
- ▶ If sufficiently many coders (= test subjects) are available, annotation can be analysed as psycholinguistic experiment
 - ▶ ANOVA, logistic regression, generalised linear models
 - ▶ correlations between annotators → systematic disagreement

Outline

Reliability & agreement

Introduction

Observed vs. chance agreement

The Kappa coefficient

Contingency tables

Chance agreement & Kappa

Statistical inference for Kappa

Random variation of agreement measures

Kappa as a sample statistic

Outlook

Extensions of Kappa

Final remarks

Suggested reading & materials

Artstein & Poesio (2008)

Everyone should at least read this article.

R package **irr** (inter-rater reliability)

Lacks confidence intervals → to be included in corpora package.

Different types of non-reliability

1. Random errors (slips)
 - ▶ Lead to chance agreement between annotators
2. Different intuitions
 - ▶ Systematic disagreement
3. Misinterpretation of tagging guidelines
 - ▶ May not result in disagreement → not detected

References I

- Artstein, Ron and Poesio, Massimo (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–596.
- Bennett, E. M.; Alpert, R.; Goldstein, A. C. (1954). Communications through limited questioning. *Public Opinion Quarterly*, **18**(3), 303–308.
- Brants, Thorsten (2000). Inter-annotator agreement for a German newspaper corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.
- Carletta, Jean (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, **22**(2), 249–254.
- Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.
- Di Eugenio, Barbara and Glass, Michael (2004). The kappa statistic: A second look. *Computational Linguistics*, **30**(1), 95–101.
- Fleiss, Joseph L.; Cohen, Jacob; Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**(5), 323–327.
- Green, Annette M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the Twenty-Second Annual SAS Users Group International Conference (online)*, San Diego, CA.

References II

- Krenn, Brigitte; Evert, Stefan; Zinsmeister, Heike (2004). Determining intercoder agreement for a collocation identification task. In *Proceedings of KONVENS 2004*, pages 89–96, Vienna, Austria.
- Krippendorff, Klaus (1980). *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.
- Landis, J. Richard and Koch, Gary G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.
- Lee, J. Jack and Tu, Z. Nora (1994). A better confidence interval for kappa (κ) on measuring agreement between two raters with binary outcomes. *Journal of Computational and Graphical Statistics*, **3**(3), 301–321.
- Scott, William A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, **19**(3), 321–325.
- Véronis, Jean (1998). A study of polysemy judgements and inter-annotator agreement. In *Proceedings of SENSEVAL-1*, Herstmonceux Castle, Sussex, UK.