# Assessing the Reliability of an Annotation Scheme for Indefinites

## Measuring Inter-annotator Agreement

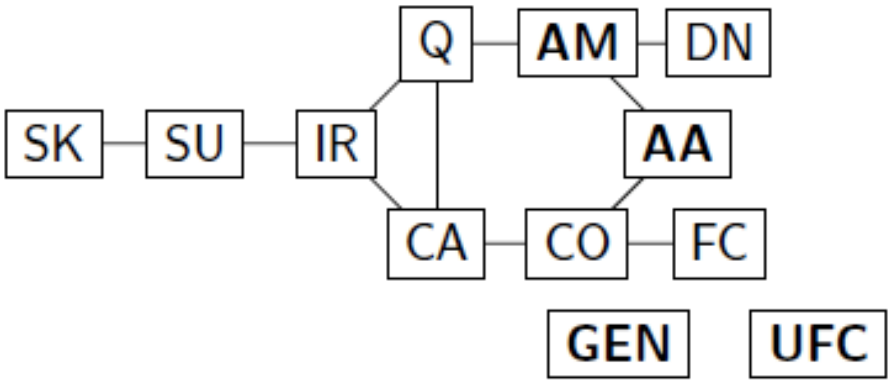
Raquel Fernández

Institute for Logic, Language & Computation
University of Amsterdam

# Semantic Judgements

Theories of linguistic phenomena are typically based on speakers' *judgements* (regarding e.g. acceptability, category, etc.).

For instance, consider Haspelmath's proposal:




- Hypothesis: an indefinite will always express a set of functions that are contiguous on the map.

# Semantic Judgements

What do we need to confirm this hypothesis? At least, the following:

- data: a set of indefinites in context;
- judgements indicating the function of each indefinite.

This raises several issues, among others:

- how much data? what kind of data - constructed examples?
- whose judgements? the investigator's? those of native speakers - how many? what if judgements differ among speakers?

# Semantic Judgements

How to overcome the difficulties associated with subjective judgements?

- Option 1: forget about judgements and work with raw data
- Option 2: gather evidence that speakers other than the investigators' themselves can make similar judgements
  - take judgements from several speakers and measure their agreement.

**From Carletta (1996):**

"At one time, it was considered sufficient when working with such judgments to show examples based on the authors' interpretation. Research was judged according to whether or not the reader found the explanation plausible. Now, researchers are beginning to require evidence that people besides the authors themselves can understand, and reliably make, the judgments underlying the research. This is a reasonable requirement, because if researchers cannot even show that people can agree about the judgments on which their research is based, then there is no chance of replicating the research results."

# Annotations and their Reliability

When data and judgements are stored in a computer-readable format, judgements are typically called *annotations*.

- Since annotations correspond to speakers' judgements, there isn't an objective way of establishing the *validity* of an annotation.
- Instead, we aim to measure the *reliability* of an annotation:
  - annotations are reliable if annotators agree *sufficiently for relevant purposes* – they consistently make the same decisions.
  - high reliability is a prerequisite for validity.
- How can the reliability of an annotation be determined?
  - several coders annotate the same data with the same guidelines
  - calculate *inter-annotator agreement*

# Inter-annotator Agreement

How can inter-annotator agreement be calculated?

- Some terminology and notation:
  - set of items $\{i \mid i \in I\}$, with cardinality $\mathbf{i}$.
  - set of categories $\{k \mid k \in K\}$, with cardinality $\mathbf{k}$.
  - set of coders $\{c \mid c \in C\}$, with cardinality $\mathbf{c}$.

# Observed Agreement

The simplest measure of agreement is *observed agreement* $A_o$:

- the percentage of judgements on which the coders agree, that is the number of items on which coders agree divided by total number of items.

Binary classification task: content-container relation

| items | coder A | coder B | agr |
|---|---|---|---|
| Put *tea* in a *heat-resistant jug* and ... | true | true | ✓ |
| The *kitchen* holds patient *drinks* and snacks. | true | false | × |
| Where are the *batteries* kept in a *phone*? | true | false | × |
| ...the *robber* was inside the *office* when ... | false | false | ✓ |
| Often the *patient* is kept in the *hospital* ... | false | false | ✓ |
| *Batteries* stored in *contact* with one another... | false | false | ✓ |

- $A_o = 4/6 = 66.6\%$

Contingency table:

| | coder B | | |
|---|---|---|---|
| **coder A** | true | false | |
| true | 1 | 2 | 3 |
| false | 0 | 3 | 3 |
| | 1 | 5 | 6 |

Contingency table with proportions:
(each cell divided by total # of items **i**)

| | coder B | | |
|---|---|---|---|
| **coder A** | true | false | |
| true | .166 | .333 | .5 |
| false | 0 | .5 | .5 |
| | .166 | .833 | 1 |

- $A_o = .166 + .5 = .666 = 66.6\%$

# Observed vs. Chance Agreement

Problem: using observed agreement to measure reliability does not take into account agreement that is due to *chance*.

- In the above example, if annotators make random choices the expected agreement due to chance is 50%:
  - both coders randomly choose `true` $(.5 \times .5 = .25)$
  - both coders randomly choose `false` $(.5 \times .5 = .25)$
  - expected agreement by chance: $.25 + .25 = 50\%$
- An observed agreement of $66.6\%$ is only mildly better than $50\%$

# Observed vs. Chance Agreement

- *Number of categories*: fewer categories will result in higher observed agreement by chance.
  $\mathbf{k} = 2 \rightarrow 50\%$ $\mathbf{k} = 3 \rightarrow 33\%$ $\mathbf{k} = 4 \rightarrow 25\%$ ...
- *Distribution of items among categories*: if some categories are very frequent, observed agreement will be higher by chance.
  - both coders randomly choose `true` $(.95 \times .95 = 90.25\%)$
  - both coders randomly choose `false` $(.05 \times .05 = 0.25\%)$
  - expected agreement by chance $90.25 + 0.25 = 90.50\%$

  $\Rightarrow$ Observed agreement of 90% may be less than chance agreement.

Observed agreement does not take these factors into account and hence is not a good measure of reliability.

# Measuring Reliability

$\Rightarrow$ Reliability measures must be corrected for *chance agreement*.

- Let $A_o$ be observed agreement, and $A_e$ expected agreement by chance.
- $1 - A_e$: how much agreement beyond chance is attainable.
- $A_o - A_e$: how much agreement beyond chance was found.
- General form of chance-corrected agreement measure of reliability:

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

The ratio between $A_o - A_e$ and $1 - A_e$ tells us which proportion of the possible agreement beyond chance was actually achieved.

- Some general properties of $R$:

| **perfect agreement** | **chance agreement** | **perfect disagreement** |
|---|---|---|
| $R = 1 = \frac{A_o - A_e}{1 - A_e}$ | $R = 0 = \frac{0}{1 - A_e}$ | $R = \frac{0 - A_e}{1 - A_e}$ |

# Measuring Reliability: *kappa*

Several agreement measures have been proposed in the literature (see Arstein & Poesio 2008 for details)

- The general form of $R$ is the same for several measures $R = \frac{A_o - A_e}{1 - A_e}$
- They all compute $A_o$ in the same way:
  - proportion of agreements over total number of items
- They differ on the precise definition of $A_e$.

We'll focus on the *kappa* ($\kappa$) coefficient (Cohen 1960; see also Carletta 1996)

- $\kappa$ calculates $A_e$ considering *individual* category distributions:
  - they can be read off from the marginals of contingency tables:

| | **coder B** | | |
|---|---|---|---|
| **coder A** | true | false | |
| true | 1 | 2 | 3 |
| false | 0 | 3 | 3 |
| | 1 | 5 | 6 |

| | **coder B** | | |
|---|---|---|---|
| **coder A** | true | false | |
| true | .166 | .333 | .5 |
| false | 0 | .5 | .5 |
| | *.166* | *.833* | 1 |

category distribution for coder A: $P(c_A|\texttt{true}) = .5$ ; $P(c_a|\texttt{false}) = .5$
category distribution for coder B: $P(c_B|\texttt{true}) = .166$ ; $P(c_B|\texttt{false}) = .833$

# Chance Agreement for *kappa*

$A_e$: how often are annotators expected to agree if they make random choices according to their individual category distributions?

- we assume that the decisions of the coders are independent: need to multiply the marginals
- Chance of $c_A$ and $c_B$ agreeing on category $k$: $P(c_A|k) \cdot P(c_B|k)$
- $A_e$ is then the chance of the coders agreeing on any $k$:

$$A_e = \sum_{k \in K} P(c_A|k) \cdot P(c_B|k)$$

| | **coder B** | | |
|---|---|---|---|
| **coder A** | true | false | |
| true | 1 | 2 | 3 |
| false | 0 | 3 | 3 |
| | 1 | 5 | 6 |

| | **coder B** | | |
|---|---|---|---|
| **coder A** | true | false | |
| true | .166 | .333 | .5 |
| false | 0 | .5 | .5 |
| | *.166* | *.833* | 1 |

- $A_e = (.5 \cdot .166) + (.5 \cdot .833) = .083 + .416 = 49.9\%$

# Kappa for our Example

| items | coder A | coder B | agr |
|---|---|---|---|
| Put *tea* in a *heat-resistant jug* and ... | true | true | ✓ |
| The *kitchen* holds patient *drinks* and snacks. | true | false | × |
| Where are the *batteries* kept in a *phone*? | true | false | × |
| ...the *robber* was inside the *office* when ... | false | false | ✓ |
| Often the *patient* is kept in the *hospital* ... | false | false | ✓ |
| *Batteries* stored in *contact* with one another... | false | false | ✓ |

| | coder B | | |
|---|---|---|---|
| coder A | true | false | |
| true | 1 | 2 | 3 |
| false | 0 | 3 | 3 |
| | 1 | 5 | 6 |

| | coder B | | |
|---|---|---|---|
| coder A | true | false | |
| true | **.166** | .333 | .5 |
| false | 0 | **.5** | .5 |
| | .166 | .833 | 1 |

- $A_o = .166 + .5 = .666 = 66.6\%$
- $A_e = (.5 \cdot .166) + (.5 \cdot .833) = .083 + .416 = 49.9\%$

$$\kappa = \frac{66.6 - 49.9}{1 - 49.9} = \frac{16.7}{50.1} = \mathbf{33.3\%}$$

- *Kappa* for multiple annotators: compute $\kappa$ for each possible pair of annotators, then report average (and standard deviation).

# Scales for the Interpretation of Kappa

- Landis and Koch (1977)

```
0.0 - 0.2 :   slight
      0.2 - 0.4 :   fair
            0.4 - 0.6 :   moderate
                  0.6 - 0.8:   substantial
                        0.8 - 1.0 :   perfect
```

- Krippendorff (1980)

```
0.0 - 0.67 :   discard
      0.67 - 0.8 : tentative
             0.8 - 1.0:   good
```

- Green (1997)

```
0.0 - 0.4 :   low
      0.4 - 0.75 : fair / good
            0.75 - 1.0:   high
```

- There are many other suggestions as well. . .

# Weighted Disagreements

- The classic version of $\kappa$ considers all types of disagreements equally.
- However, we may want to treat some disagreements as more important than others – some categories may be more similar than others.
- We can use *weighted coefficients*: Krippendorff's $\alpha$ and *weighted kappa* $\kappa_w$.
  - The formula for $\kappa_w$ derives agreement from disagreement:

$$\kappa_w = 1 - \frac{D_o}{D_e}$$

  - We'll see how to derive $D_o$ and $D_e$ from the confusion matrices; for details of the formulas see Arstein & Poesio (2008).

# Weighted Disagreements – An Example

Consider this confusion matrix from Arstein & Poesio (2008):

| | **coder B** | | | |
|---|---|---|---|---|
| **coder A** | Stat | IReq | Chck | |
| Stat | 46 | 6 | 0 | 52 |
| IReq | 0 | 32 | 0 | 32 |
| Chck | 0 | 6 | 10 | 16 |
| | 46 | 44 | 10 | 100 |

We can calculate *unweighted* $\kappa$ as described before:

- $A_o$ : the sum of the cells in the diagonal
  $A_o = .46 + .32 + .10 = .88$
- $A_e$ : the sum of the marginals for each category (multiplied)
  $A_e = .46 \times .52 + .44 \times .32 + .10 \times .16 = .396$
- $\kappa = (A_o - A_e)/(1 - A_e)$
  $\kappa = (.88 - .396)/(1 - .396) = .8013$

# Weighted Disagreements – An Example

Suppose we weight the distances between the categories as shown in the RHS table: identical categories have 0 disagreement, while 1 denotes maximal disagreement.

| | coder B | | | |
|---|---|---|---|---|
| **coder A** | Stat | IReq | Chck | |
| Stat | 46 | 6 | 0 | 52 |
| IReq | 0 | 32 | 0 | 32 |
| Chck | 0 | 6 | 10 | 16 |
| | 46 | 44 | 10 | 100 |

| | coder B | | |
|---|---|---|---|
| **coder A** | Stat | IReq | Chck |
| Stat | 0 | 1 | 0.5 |
| IReq | 1 | 0 | 0.5 |
| Chck | 0.5 | 0.5 | 0 |

To calculate $\kappa_w$, we can derive $D_o$ and $D_e$ as follows:

- $D_o$ : the sum of all cells multiplying each cell by each weight (and dividing by total of items if not working with proportions).
- $D_e$ : the sum of $D_e^{k_i k_j}$ for each category pair $k_i, k_j$, where
  - $D_e^{k_i k_j}$ : the product of the marginals for $k_i$ and $k_j$ divided by the total of items (or the square of the total of items if not working with proportions), multiplying each cell by each weight.

# Weighted Disagreements – An Example

| coder A | coder B: Stat | IReq | Chck | |
|---|---|---|---|---|
| Stat | 46 | 6 | 0 | 52 |
| IReq | 0 | 32 | 0 | 32 |
| Chck | 0 | 6 | 10 | 16 |
| | 46 | 44 | 10 | 100 |

| coder A | coder B: Stat | IReq | Chck |
|---|---|---|---|
| Stat | 0 | 1 | 0.5 |
| IReq | 1 | 0 | 0.5 |
| Chck | 0.5 | 0.5 | 0 |

- $D_o$ : the sum of all cells multiplying each cell by each weight (and dividing by total of items if not working with proportions).

$$\mathrm{D_o} = \frac{46\times 0 + 6\times 1 + 32\times 0 + 6\times 0.5 + 10\times 0}{100} = \frac{6+3}{100} = 0.09$$

- $D_e$ : the sum of $D_e^{k_i k_j}$ for each category pair $k_i, k_j$, where
  - $D_e^{k_i k_j}$ : the product of the marginals for $k_i$ and $k_j$ divided by the total of items (or the square of the total of items if not working with proportions), multiplying each cell by each weight.

$$\begin{aligned} &\tfrac{46\times 52}{100\times 100}\times 0 + \tfrac{44\times 52}{100\times 100}\times 1 + \tfrac{10\times 52}{100\times 100}\times \tfrac{1}{2} \\ &+ \tfrac{46\times 32}{100\times 100}\times 1 + \tfrac{44\times 32}{100\times 100}\times 0 + \tfrac{10\times 32}{100\times 100}\times \tfrac{1}{2} \\ &+ \tfrac{46\times 16}{100\times 100}\times \tfrac{1}{2} + \tfrac{44\times 16}{100\times 100}\times \tfrac{1}{2} + \tfrac{10\times 16}{100\times 100}\times 0 \end{aligned} \qquad 0.49$$

# Weighted Disagreements – An Example

| | coder B | | | |
|---|---|---|---|---|
| **coder A** | Stat | IReq | Chck | |
| Stat | 46 | 6 | 0 | 52 |
| IReq | 0 | 32 | 0 | 32 |
| Chck | 0 | 6 | 10 | 16 |
| | 46 | 44 | 10 | 100 |

| | coder B | | |
|---|---|---|---|
| **coder A** | Stat | IReq | Chck |
| Stat | 0 | 1 | 0.5 |
| IReq | 1 | 0 | 0.5 |
| Chck | 0.5 | 0.5 | 0 |

$$\kappa_w = 1 - \frac{D_o}{D_e}$$

$\kappa_w = 1 - (.09/.49) = .8163$

$\kappa = (.88 - .396)/(1 - .396) = .8013$

# For Our Project

- Guidelines
- Set of data
- Set of annotators annotating the full set of data
- Extract the confusion matrices of the resulting annotations and analyse them
  - need to work out the technicalities involved in doing the annotation in a reliable way and extracting the confusion matrices, plus calculating agreement from them (possibly with online calculators)
- Pilot study
- Final annotation experiment from which we can draw conclusions on the reliability of the scheme for indefinites
- Different types of non-reliability:
  - Random slips: lead to chance agreement between annotators
  - Different intuitions: lead to systematic disagreements
  - Misinterpretation of annotation guidelines: may not result in disagreement $\rightarrow$ may not be detected

# References


- Artstein, Ron and Poesio, Massimo (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4), 555–596.
- 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.
- Green, Annette M. (1997). Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the Twenty-Second Annual SAS Users Group International Conference*, San Diego, CA.
- 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.