DKPro Agreement
An Open-Source Java Library for Measuring Inter-Rater Agreement

Christian M. Meyer, Margot Mieskes, Christian Stab, and Iryna Gurevych
Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt / German Institute for Educational Research

Summary
DKPro Agreement is an open-licensed Java library for computing inter-rater agreement using a shared interface and data model. Highlights:
• Support for all commonly used inter-rater agreement measures
• Calculation of multiple coefficients using the same data model
• Both coding and unitizing setups are possible
• Multiple diagnostic devices and visual aids for analyzing disagreement
• Thoroughly tested on a wide range of examples from the literature over 60 test cases for annotation studies including citation of original source
• Available as open source software under the Apache License 2.0 (ASL)
• Extensions and comments welcome!
• Integrates well with existing Java-based NLP frameworks
• Ready-to-use via Maven Central – simply specify the dependency: g: groupid: dkpro-dkpro.statistics artifactId: dkpro-statistics-agreement version: 2.0.0 Part of DKPro Statistics collection

Motivation
• Reliability is a necessary precondition of high quality datasets
• Long tradition of assessing inter-rater agreement in psychology, medicine, content analysis
• In NLP/CL often ignored or limited
• Researchers rely on manual calculations
• Hasty implementation
• Insufficiently documented online calculators
• Measures are often not comparable
• Urgent need for software that implements the most important measures allows for diagnosing disagreement integrates with existing projects and annotation workbenches (e.g., WebAnno, CSniper)

Coding Setup
Raters assign categories to fixed items.
• Document classification
• POS tagging
• Dialog act tagging
• etc.

Step 1: Represent the Annotated Data
Define annotations manually or load from flat files/DB or use UIMA annotations

Available coefficients:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Type</th>
<th>Raters</th>
<th>Chance-corr.</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage agreement β</td>
<td>coding</td>
<td>≥ 2</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Bennett et al.'s β1 (1954)</td>
<td>coding</td>
<td>2</td>
<td>uniform –</td>
<td>–</td>
</tr>
<tr>
<td>Scott's β1 (1955)</td>
<td>coding</td>
<td>2</td>
<td>study-specific –</td>
<td>–</td>
</tr>
<tr>
<td>Cohen's κ (1960)</td>
<td>coding</td>
<td>2</td>
<td>rater-specific –</td>
<td>–</td>
</tr>
<tr>
<td>Randolph's β (2005) [MS]</td>
<td>coding</td>
<td>≥ 2</td>
<td>uniform –</td>
<td>–</td>
</tr>
<tr>
<td>Fleiss's κ1 (1971) [MS]</td>
<td>coding</td>
<td>≥ 2</td>
<td>study-specific –</td>
<td>–</td>
</tr>
<tr>
<td>Hubert's κ1 (1977) [MS]</td>
<td>coding</td>
<td>≥ 2</td>
<td>rater-specific –</td>
<td>–</td>
</tr>
<tr>
<td>Krippendorff's α (1980)</td>
<td>coding</td>
<td>≥ 2</td>
<td>study-specific –</td>
<td>–</td>
</tr>
<tr>
<td>Cohen's weighted κ (1968)</td>
<td>coding</td>
<td>≥ 2</td>
<td>rater-specific –</td>
<td>–</td>
</tr>
<tr>
<td>Krippendorff's α (1995)</td>
<td>unitizing</td>
<td>≥ 2</td>
<td>study-specific –</td>
<td>–</td>
</tr>
</tbody>
</table>

Code example:

```java
PercentageAgreement pa = new PercentageAgreement(study); System.out.println(pa.calculateAgreement());
FleissKappaAgreement kappa = new FleissKappaAgreement(study); System.out.println(kappa.calculateAgreement());
KrippendorffFAlphAgreement alpha = new KrippendorffFAlphAgreement(study, new NominalDistanceFunction()); System.out.println(alpha.calculateObservedDisagreement());
```

Step 2: Measure the Inter-Rater Agreement

Step 3: Analyze the Disagreement

Raw agreement scores are of limited help for diagnosing the main sources of disagreement. DKPro Agreement provides multiple diagnostic devices:

Agreement insights
• Observed agreement
• Expected agreement
• Rater-specific agreement
• Category-specific agreement
• Item-specific agreement

Formatted output and visual aids
• Coincidence matrix
• Contingency matrix
• Reliability matrix
• Continuum of a unitizing study
• Planned: Hinton diagrams

http://www.ukp.tu-darmstadt.de