Appendix

Inter-coder agreement in ATLAS.ti [i]

Prof. Klaus Krippendorff

What is reliability?

Reliability is the ability to rely on something, here on data generated by coding within ATLAS.ti for representing textual phenomena, their readings, of analytical interest, and only these.

Concerns with the reliability of data are **motivated** by the experience that unreliable data reduce the chance of their analysis to lead to valid conclusions; introduce uncertainty for researchers to know what they are analyzing, and make it difficult for other scholars, critics, and stakeholders of said phenomena to interpret or build on the published findings.

Reliability can be experienced only when the use of data did not lead to failures. Prior to such experiences, the reliability of data **needs to be inferred** from observable conditions that are known to reduce failures.

Three **kinds of reliabilities** can be distinguished by the sources of unreliability they respectively capture:

- **Stability of one coder** declines when that coder confuses given codes, use them inconsistently over time, or is unable to repeat the process of generating data
- **Replicability of the coding instructions** declines not only with intra-coder instabilities but also with inter-coder disagreements among several coders who interpret and apply them independent of each other to the same set of phenomena. Replicability has to be immune to all irrelevant influences, whether unequal coder qualifications, unlike literary competences, different recording instruments or variations in the circumstances of replication, and times
- **Accuracy of the coded data** refers to the correspondence of coding by one or more coders with an accepted standard. It declines with intra-coder instabilities, inter-coder disagreements, and shared coder preferences, biases, or prejudices.

Evidence of replicability is stronger than evidence of intra-coder stabilities but weaker than evidence of accuracy. However, standards for the coding of data are rarely available and when they are, coding efforts would be mute, except for testing small subsamples of reliability data. Therefore, replicability is the reliability measure of choice.

What data are needed to infer replicability?

Data that give rise to inter-coder agreements from which the replicability of a population of data can be inferred

- Have to **replicate** the very coding process on a sample of phenomena to be converted into reliable data, using different coders who apply identical coding instructions to the same set of phenomena of analytical interest
- Must be **informed by written coding instructions**, and only these
- The **sample** (volume of textual matter including videos) of phenomena to be coded must be large enough to represent the diversity of the phenomena of analytical interest
- **Coder qualifications** must be **sufficiently common** for coders to be freely replaceable.
- The **number of coders** employed needs to embrace various analysts' and diverse stakeholders' ability to understand the phenomena studied through the coding instructions applied. Two coders may not suffice
- **Coders must work independent of each other** and not communicate about their coding task
- Any **preparatory training** that coders received and the qualifications for which they were selected need to be **communicable for replication elsewhere.**

Deviations from these conditions tend to pollute the reliability data and inflate the observable inter-coder agreement, leading to mistaken assurances of their replicability. For example, selecting coders among close acquaintances, with a stake in the outcome of a study, receiving thorough but undocumented training, allowing them to discuss how to interpret given coding instructions, or settling emerging uncertainties by consensus, yield deceptively higher inter-coder agreements which is no longer indicative of the replicability of the generated data.

Perfect replicability means that data have the potential of leading to valid answers of given research questions, analysts can use the communicable coding instructions in reverse, to decode what their data represent, and the stakeholders in a research project can critically evaluate the analysis, talk about, or respond to the phenomena studied and build on published findings.

Reliability data in ATLAS.ti

After a principal investigator has developed suitable **coding instructions** in writing, without references to the textual matter to which they are to be applied, independently working coders need to apply these coding instructions to the same set of phenomena and return comparable data to the principal investigator. Although ATLAS.ti cannot prevent coders from introducing new codes, they will have to be ignored except as suggestions to improve the coding instructions for subsequent coding efforts.

Coding instructions must specify

- **The set of** relevant and logically or conceptually separate **semantic domains** with definitions and examples made readily available to coders.

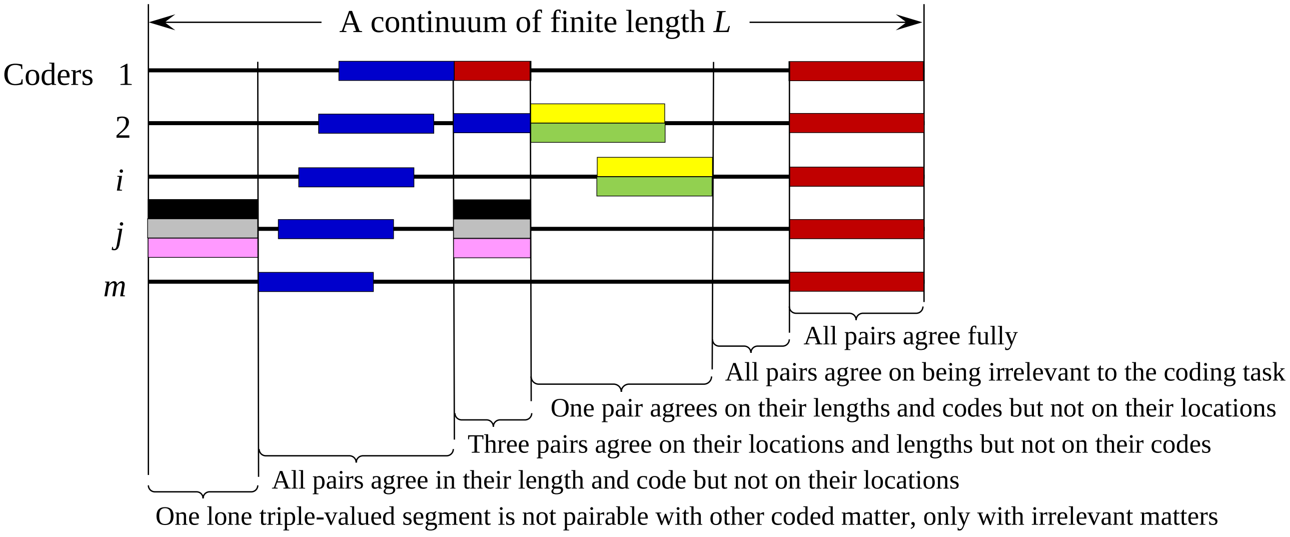
A semantic domain names a space of distinct concepts with shared meanings, e.g., “colors,” “mental illnesses,” “emotions,” “gender issues,” or “personalities.” Semantic domains may be named abstractly but are always context dependent. The concept of color is different when applied to the sky, a dress, a national flag, an ethnic group, or the state of a drunk. The gender of nouns is unlike the gender of living organisms. The contexts of semantic domains need to be preserved when coding texts. A single quote typically invokes several connected semantic domains. For example:

- “□ says □ to □ intending to □ but causing □“
  defines the semantic domains of speakers, utterances, addressees, and intended and unintended consequences.
- “ □ diagnoses □ to have □ .“
  If the first semantic domain concerns medical professionals, the second concerns patients and the third illnesses. If the first is a car mechanic, the other two relate to cars.

Most semantic domains concern attributes of objects, actions, people, or abstract ideas.

- **Each semantic domain contains a set of mutually exclusive codes** (at least one) by name with definitions and examples made readily available to coders

- **Coders highlight or identify segments of a given textual continuum**, e.g., quotes, propositions, or paragraphs and, following the written coding instructions, assign one code from each applicable semantic domain to them. For example, with colors identifying codes from separate semantic domains:


A continuum of finite length *L*
Coders 1
2
i
j
m
All pairs agree fully
All pairs agree on being irrelevant to the coding task
One pair agrees on their lengths and codes but not on their locations
Three pairs agree on their locations and lengths but not on their codes
All pairs agree in their length and code but not on their locations
One lone triple-valued segment is not pairable with other coded matter, only with irrelevant matters

Definitions of terms

Coders: $1, 2, \ldots, i, \ldots, j, \ldots m$

Segments:

Coder *i*’s segments: $S_{i1}, S_{i2}, \ldots S_{ig}, S_{ig+1}, \ldots, S_{i\text{last for } i}$

Coder *j*’s segments: $S_{j1}, S_{j2}, \ldots S_{ih}, S_{ih+1}, \ldots, S_{j\text{last for } j}$

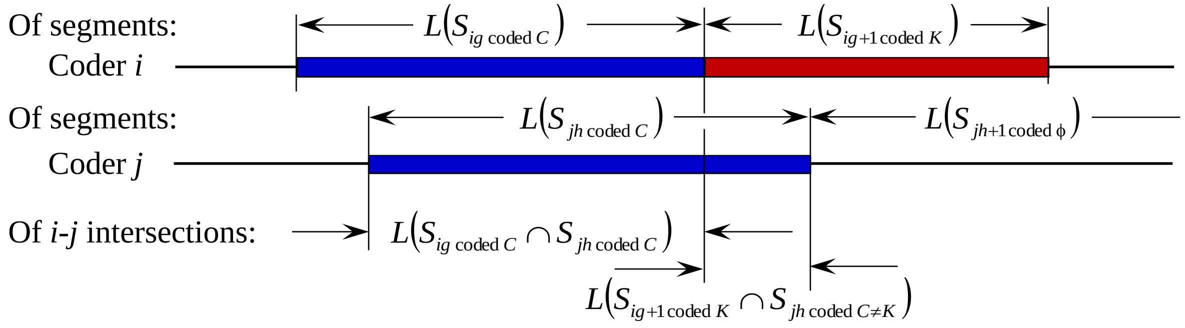
Coding of segments:

multi-valued sets C of codes $c \in C$ or K of codes $k \in K$

single-valued codes $C=c$ or $K=k$

uncoded matter, designated by $C=\phi$ or $K=\phi$

Lengths (in terms of the number of characters for texts or number of seconds for videos):


Of segments:
Coder *i*
$L(S_{ig\text{ coded }C})$
$L(S_{ig+1\text{ coded }K})$
Of segments:
Coder *j*
$L(S_{jh\text{ coded }C})$
$L(S_{jh+1\text{ coded }\phi})$
Of *i*-*j* intersections:
$L(S_{ig\text{ coded }C} \cap S_{jh\text{ coded }C})$
$L(S_{ig+1\text{ coded }K} \cap S_{jh\text{ coded }C \neq K})$

Of the continuum: $L = \sum_{g=1}^{\text{last for } i} L(S_{ig}) = \sum_{h=1}^{\text{last for } j} L(S_{jh}) L$

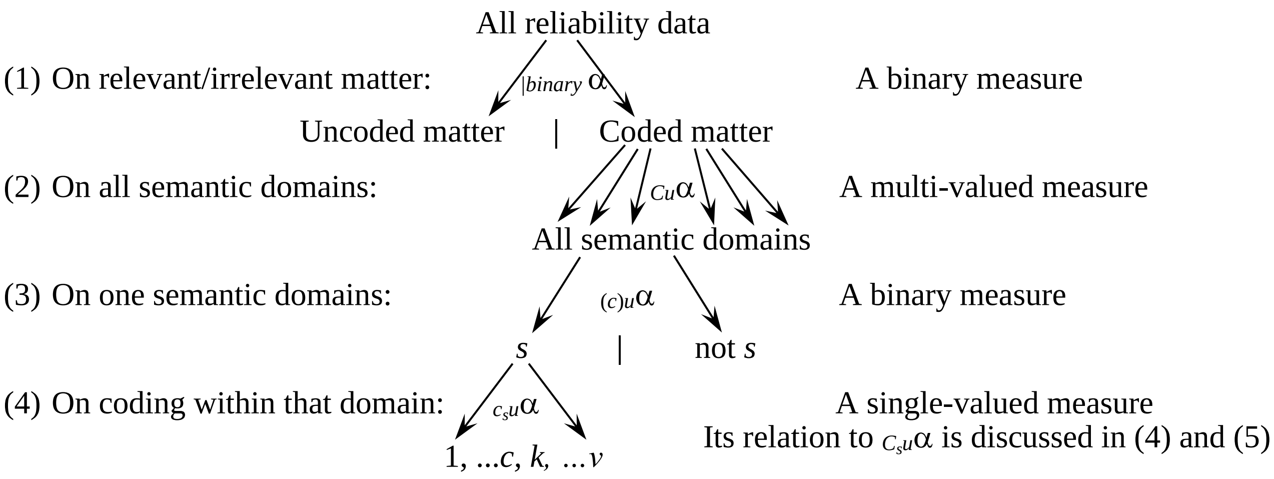
Differences:

With the number of elements in sets referred to as its cardinality $|C|$, differences

Between two sets C and K of codes: $\Delta_{CK} = |C||K - |C \cap K|^2$

Between two single codes c and k: $\Delta_{ck} = {}_{nominal}\delta^2_{ck} = \begin{cases} 0 \text{ iff } c = k \\ 1 \text{ iff } c \neq k \end{cases}$

ATLAS.ti currently provides three measures of inter-coder agreement



- *Alpha binary* indicates the extent to which coders agree on the relevance of texts for the research project.
- $_{Cu}\alpha$ indicates the extent to which coders agree on the presence or absence of sematic domains,
- $_{(c)u}\alpha$ indicates the degree to which coders identify a particular semantic domain *s*.

$_{c_s u}\alpha$ indicates the agreement on coding within a semantic domain *s*. When coders pollute the reliability data by intra-coder confusions, the multi-valued $_{C_s u}\alpha$ is computed. This coefficient is not yet implemented in ATLAS.ti

Alpha binary

The binary **agreement on distinguishing relevant + from irrelevant ϕ matter**

(Simplified by assuming all relevant matter as finely grained):

binary α’s coincidences are:

$$\ell_{\phi+} = \frac{1}{m-1}\sum_i \sum_{j\neq i}^m \sum_{g.h} L\left(S_{ig\ \text{coded}\ \phi} \cap S_{jh\ \text{coded}\ \neq\phi}\right)$$

$$\ell_{\phi\cdot} = \frac{1}{m-1}\sum_i^m \sum_g L\left(S_{ig\ \text{coded}\ \phi}\right)$$

$$\ell_{\cdot+} = \frac{1}{m-1}\sum_i^m \sum_g L\left(S_{ig\ \text{coded}\ \neq\phi}\right)$$

Its coincidence matrix is:

$$\begin{array}{c|cc|l}
 & \phi & + & \\
\hline
\phi & \ell_{\phi\phi} & \ell_{\phi+} & \ell_{\phi\cdot} \\
+ & \ell_{+\phi} & \ell_{++} & \ell_{\phi\cdot} \\
\hline
 & \ell_{\cdot\phi} & \ell_{\cdot+} & \ell_{\cdot\cdot} = mL
\end{array}$$

Its disagreements are: $${}_{|cu}D_o = \frac{\ell_{\phi+} + \ell_{+\phi}}{\ell_{\cdot\cdot}} \quad \text{and} \quad {}_{|cu}D_e = \frac{\ell_{\phi\cdot}\ell_{\cdot+} + \ell_{+\cdot}\ell_{\cdot\phi}}{\ell_{\cdot\cdot}(\ell_{\cdot\cdot}-1)}$$

The binary **α-agreement** is: $${}_{|cu}\alpha = 1 - \frac{{}_{|cu}D_o}{{}_{|cu}D_e} = 1 - (\ell_{\cdot\cdot}-1)\frac{\ell_{\phi+}}{\ell_{\phi\cdot}\ell_{\cdot+}}$$

Where: *u* stands for unitizing, *c* for single-valued coding, and | for its binary nature.

The alpha-binary coefficient measures inter-coder agreement at the most general level. It measures whether different coders identify the same sections in the data to be relevant for the topics of interest represented by the codes.

Cu-alpha for all semantic domains in the analysis

The **agreement on recognizing diverse semantic domains:**

Any segment may be described in terms of several semantic domains. Being logically or conceptually independent of each other, joint descriptions constitute multi-valued coding.

$_{Cu}\alpha$ assesses the agreement on recognizing semantic domains (not their distinct codes) within relevant matter. Accordingly, *C* and *K* are the sets of semantic domains applicable to segments of the textual continuum.

Its coincidences are:

$$\ell_{CK} = \frac{1}{m-1} \sum_i \sum_{j \neq i}^{m} \sum_{g.h} L\left(S_{ig\ \text{coded}\ C \neq \phi} \cap S_{jh\ \text{coded}\ K \neq \phi}\right)$$

Its coincidence matrix is:

$$\begin{array}{c|ccccc|l}
 & 1 & 2 & \ldots & K & v & \\
\hline
1 & \ell_{11} & \ell_{12} & \ldots & \ell_{1K} & \ell_{1v} & \ell_{1\cdot} \\
2 & \ell_{21} & \ell_{22} & \ldots & \ell_{2K} & \ell_{2v} & \ell_{2\cdot} \\
C & \ell_{C1} & \ell_{C2} & \ldots & \ell_{CK} & \ell_{Cv} & \ell_{C\cdot} = \sum_{K=1}^{v} \ell_{CK} \\
\vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots & \vdots & \vdots \\
v & \ell_{v1} & \ell_{v2} & \ldots & \ell_{vK} & \ell_{vv} & \ell_{v\cdot} \\
\hline
 & \ell_{\cdot 1} & \ell_{\cdot 2} & \ldots & \ell_{\cdot K} & \ell_{\cdot v} & \ell_{\cdot\cdot} = \sum_{C=1}^{v} \sum_{K=1}^{v} \ell_{CK}
\end{array}$$

Matching coincidences occupy its diagonal: $\ell_{11}, \ell_{22}, \ldots \ell_{CC}, \ldots \ell_{KK}, \ldots \ell_{vv}$.

Mismatching coincidences are located in its off-diagonal triangles: $\ell_{CK} = \ell_{KC}$.

Differences between two sets *C* and *K* are: $\Delta_{CK} = |C||K - |C \cap K|^2$

Its disagreements are:

$$_{Su}D_o = \frac{\sum_C \sum_K \ell_{CK} \Delta_{CK}}{\sum_C \sum_K \ell_{CK} |C||K|}$$

and:

$$_{Su}D_e = \frac{\sum_C \ell_{C\cdot} \sum_K \ell_{\cdot K} \Delta_{CK}}{(\sum_C \ell_{C\cdot} |C|)^2 - \sum_i^m \sum_g \left(L(S_{ig\ \text{coded}\ C \neq \phi})\right)^2 |C|}\ .$$

The $_{Cu}\alpha$**-agreement** is:

$$_{Su}\alpha = 1 - \frac{_{Su}D_o}{_{Su}D_e}$$

Where the capital *S* stands for multi-valued sets of references uses of semantic domains.

The (capital) Cu-alpha coefficient gives you a summary value for all semantic domains in the analysis.

cu-alpha for a particular semantic domain

The **agreement of identifying the applicability of any one semantic domain *s***:

A chosen semantic domain s may or may not occur in a set of semantic domains, $s \in C$ of the coincidences generated in (2): $\ell_{CK} = \frac{1}{m-1} \sum_i \sum_{j \neq i}^m \sum_{g.h} L\left(S_{ig \text{ coded } C \neq \phi} \cap S_{jh \text{ coded } K \neq \phi}\right)$

Its disagreements are:

$$ {}_{(s)u}D_o = \frac{\sum_s \sum_K \sum_{k \in K} \ell_{CK} \text{ iff } s \in C \text{ and } k \notin C}{\sum_s \sum_K \sum_{k \in K} \ell_{CK} \text{ iff } s \in C} $$

and from (2):

$$ {}_{Su}D_e = \frac{\sum_C \ell_{C\cdot} \sum_K \ell_{\cdot K} \Delta_{CK}}{(\sum_C \ell_{C\cdot} |C|)^2 - \sum_i^m \sum_g \left(L(S_{ig \text{ coded } C})\right)^2 |C|} $$

The **${}_{(c)u}\alpha$-agreement** is: ${}_{(s)u}\alpha = 1 - \frac{{}_{(s)u}D_o}{{}_{Su}D_e}$

Where (s) denotes the semantic domain singled out for attention.

The (lower case) cu-alpha coefficient gives you value for the performance of one selected semantic domain.

About the author

Klaus Krippendorff's research focuses on the role of language and dialogue in the social construction of reality: identities, institutions, cultural artifacts, power, Otherness, and meanings; emancipatory epistemology (hermeneutics) of human communication and the design of technology; content analysis, semantics, pragmatics of social interaction, and related research methods; conversation theory, information theory, and cyberspace; and second-order cybernetics of complex communication systems and their reflexive, self-organizing, and autopoietic properties.

Webseite: https://www.asc.upenn.edu/people/faculty/klaus-krippendorff-phd

Kontakt: Klaus.krippendorff@asc.upenn.edu

[i] Developed from Chapter 12 in Klaus Krippendorff (2018). *Content Analysis; An Introduction to Its Methodology, 4th Edition*. Thousand Oaks, CA: Sage. Partly implemented in Klaus Krippendorff; Yann Mathet; Stéphane Bouvry & Antoine Widlöcher (2016). On the Reliability of Unitizing Textual Continua: Further Developments. *Quality & Quantity 50*, 6: 2347-2364. Online since 2015.9.15 at https://link.springer.com/article/10.1007/s11135-015-0266-1