Annotator agreement in the anonymization of court decisions

CL2021 in Limerick

Philipp Heinrich,
Natalie Dykes, Stefan Evert

Chair of Computational Corpus Linguistics
Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

recorded on
June 27, 2021

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

Why bother?

Legal documents are interesting!

- Publication of court decisions necessary for transparent legal system
- Electronic legal documents essential as training data for legal tech applications
- Corpus-linguistic research (e.g. terminology, comprehensibility, . . .)

Why bother?

Legal documents are interesting!

- Publication of court decisions necessary for transparent legal system
- Electronic legal documents essential as training data for legal tech applications
- Corpus-linguistic research (e.g. terminology, comprehensibility, . . .)

Individuals (e. g. witnesses) have a right to remain anonymous!

- Compliance with constitutional data protection rights (i.a. GDPR)

⇒ Remove any information that might be used for de-anonymization

Why bother?

Legal documents are interesting!

- Publication of court decisions necessary for transparent legal system
- Electronic legal documents essential as training data for legal tech applications
- Corpus-linguistic research (e.g. terminology, comprehensibility, . . .)

Individuals (e. g. witnesses) have a right to remain anonymous!

- Compliance with constitutional data protection rights (i.a. GDPR)

$\Rightarrow$ Remove any information that might be used for de-anonymization

$\Downarrow$

Funding by the **Bavarian State Ministry of Justice** for research on the **automatic anonymization of court decisions**

Project and team

- **Goal**: Evaluation of the legal and technical issues concerning the ability to automatically anonymize (and pseudonymize) court decisions
- **Interdisciplinary project**: legal theory and methodology (guidelines), computational corpus linguistics (annotation, automatization)

CCL

- Prof. Dr. Stefan Evert
- Natalie Dykes
- Philipp Heinrich

Law School

- Prof. Dr. Axel Adrian
- Michael Keuchen

+ 4–8 student assistants

Tag set

direct identifiers

- names (natural and legal persons)
- addresses
- registration numbers
- dates of events

Tag set

direct identifiers

- names (natural and legal persons)
- addresses
- registration numbers
- dates of events

⋮

indirect identifiers

- profession details, academic titles, health conditions
- descriptive information about local conditions or companies
- unique features (e.g. the only red house in a small village)

Ortsangabe [HI]

75 die Beklagten zu verurteilen, das Anwesen Feldstraße 4 d, 91096 Möhrendorf zu räumen und an die Kläger herauszugeben,

Ortsangabe [HI] Prozess [LO]

77 hilfsweise die Beklagten zu verurteilen, das Anwesen Feldstraße 4 d, 91096 Möh

79 Die Beklagten haben beantragt,

81 die Klage abzuweisen.

83 Das Gericht hat Beweis erhoben durch Inaugenscheinnahme des Anwesens der … r Inaugenscheinnahme

Prozess[LO]

21.05.2020 (Bl. 83 d. A.) und hinsichtlich der Zeugenvernehmung auf das der Sit … n bestellten und vereid

Name[HI]

Prof. Dr. Ing. Helmut Zöller erhoben.

85 ENTSCHEIDUNGSGRÜNDE

87 Die zulässige Klage ist im Hauptantrag begründet.

89 I.

91 Die Klage ist zulässig.

93 Zwar ist die Beklagte zu 2) nicht Vertragspartner des Mietvertrages. Sie lebt jedo … fnis besteht.

95 II.

97 Die Klage ist begründet.

99 Die Kläger können von dem Beklagten zu 1) die Räumung des Anwesens verlan

Sachverhalt [LO]

101 1. Der auf unbestimmte Zeit geschlossene Mietvertrag vom 01.10.2014 wurde … s aus wichtigem Grund wenn dem Kündigenden unter Berücksichtigung aller Umstände des Einzelfalls – … etverhältnisses bis zur sonstigen Beendigung des Mietverhältnisses nicht zugemutet werden kann.

Merkmal [LO]keep

103 2. Einen ausreichenden Kündigungsgrund stellen die in den Kündigungsschreiben der Kläger vorgebrachten Sachverhalte nur im Hinblick auf das unerlaubte Handeltreiben mit Betäubungsmittel im streitgegenständlichen Anwesen dar.

Edit Annotation

Text

Feldstraße 4 d, 91096 Möhrendorf — Link

Search

google, wikipedia

Entity type

- Name
- Merkmal
- Sonstiges
- ⊟ juristische Person
  - Name
  - Merkmal
  - Sonstiges
- ⊟ Adresse
  - ◉ Ortsangabe
  - Merkmal
  - Sonstiges
- ⊟ Datum
  - Prozessgeschichte
  - Sachverhalt

Entity attributes

Risiko: hoch ☐ informationserhaltend ☐ unsicher ☐ uneinig

Notes

×

Add Frag. | Delete | Move | OK | Cancel

Corpus & data

Tabular data analyzed here

- based on 513 verdicts
- law of tenancy and traffic law
- 917,163 tokens
- 24,972 sensitive text spans after adjudication
- columns:
  - document ID
  - character span (start, end)
  - category tag
  - risk level

Tags and risk assessment (selection)

	high	medium	low
address (indirect)	1	182	1046
address (exact)	2317	1127	1028
date (fact)	0	7	4212
date (process)	0	0	3704
formal (court)	0	0	2120
formal (reference number)	6	18	1833
legal person (indirect)	0	14	38
legal person (name)	51	695	27
natural person (indirect)	0	19	306
judges, lawyers, ... (name)	1765	2	228
natural person (name)	3333	0	2
car (indirect)	0	0	670
car (registration number)	209	1	0

Tags and risk assessment (selection)

	high	medium	low
address (indirect)	1	182	1046
address (exact)	2317	1127	1028
date (fact)	0	7	4212
date (process)	0	0	3704
formal (court)	0	0	2120
formal (reference number)	6	18	1833
legal person (indirect)	0	14	38
legal person (name)	51	695	27
natural person (indirect)	0	19	306
judges, lawyers, . . . (name)	1765	2	228
natural person (name)	3333	0	2
car (indirect)	0	0	670
car (registration number)	209	1	0

Estimating recall for selected categories

- Restriction to selected categories: $n = 15{,}005$ sensitive text spans

- Focus on precision and recall
  - standard IAA measures (Cohen's κ, Krippendorff's α) problematic (overlaps)
  - here: annotator has successfully identified text span if their annotation (regardless of selected category) overlaps with the respective span in the adjudicated data

Estimating recall for selected categories

- Restriction to selected categories: $n = 15{,}005$ sensitive text spans
- Focus on precision and recall
  - standard IAA measures (Cohen's κ, Krippendorff's α) problematic (overlaps)
  - here: annotator has successfully identified text span if their annotation (regardless of selected category) overlaps with the respective span in the adjudicated data
- Probably safe to say there are no **false positives** in the adjudicated data set
  - precision can be estimated reliably (but is not really of interest for our project)

Estimating recall for selected categories

- Restriction to selected categories: $n = 15{,}005$ sensitive text spans
- Focus on precision and recall
  - standard IAA measures (Cohen's κ, Krippendorff's α) problematic (overlaps)
  - here: annotator has successfully identified text span if their annotation (regardless of selected category) overlaps with the respective span in the adjudicated data
- Probably safe to say there are no **false positives** in the adjudicated data set
  - precision can be estimated reliably (but is not really of interest for our project)
- But did we miss any sensitive text spans?
  - Let $n_0 \geq n$ be the true number of text spans that need to be anonymized, i.e. we assume there are still $n_0 - n$ **false negatives** in the adjudicated data!

Estimating recall for selected categories

- But did we miss any sensitive text spans?
  - Let $n_0 \geq n$ be the true number of text spans that need to be anonymized, i.e. we assume there are still $n_0 - n$ **false negatives** in the adjudicated data!

- Naïve estimates for individual *success probabilities* (= recall)

RB	LT	HS	MP
97.42%	97.37%	96.54%	99.20%

  will inevitably overestimate real recall

- NB: annotator MP best, even when not the final adjudicator

A naïve statistical model

Assumptions

a) coders only make random errors with probability of failure $q = 1 - p$

b) $q = 1 - p$ is the same for all text spans

c) $q = 1 - p$ is the same for all coders

d) errors made by the different coders on different items are independent

A naïve statistical model

Assumptions

a) coders only make random errors with probability of failure $q = 1 - p$

b) $q = 1 - p$ is the same for all text spans

c) $q = 1 - p$ is the same for all coders

d) errors made by the different coders on different items are independent

Random variable

- N: number of coders needed until a given text span is found for the first time
- $N = k$: first $k - 1$ coders have missed the span, but coder k has annotated it
- $N \sim \text{Geo}(p)$, i. e. for $k \in \{1, 2, \ldots\}$:

$$\mathbb{P}\{N = k\} = (1 - p)^{k-1} \cdot p \quad \text{and} \quad \mathbb{P}\{N \leq k\} = 1 - (1 - p)^k$$

Observable variables

definition

I_k : number of text spans with $N = k$ (found for the first time by k-th coder)

C_k : number of text spans with $N \leq k$ (found by a set of k coders)

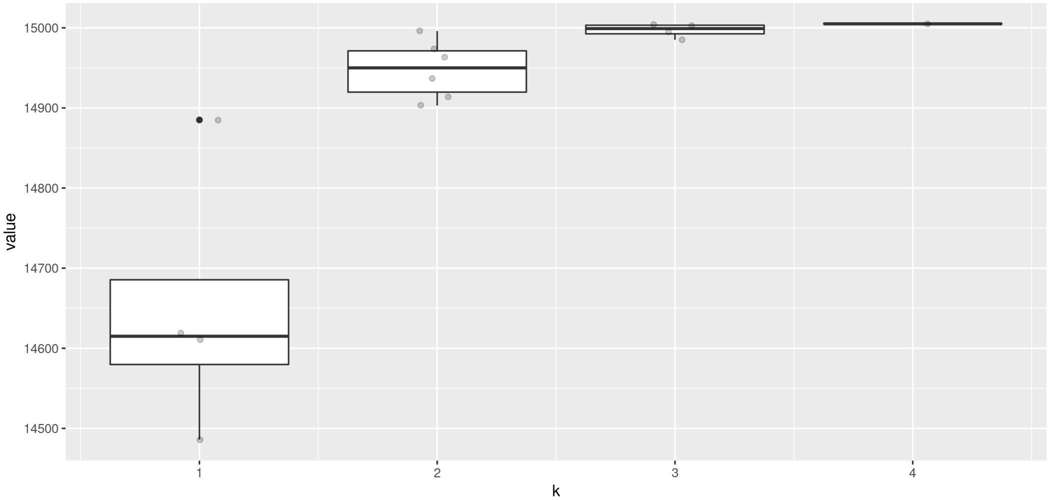
distribution

probabilites are equal for all spans and errors are assumed to be independent:

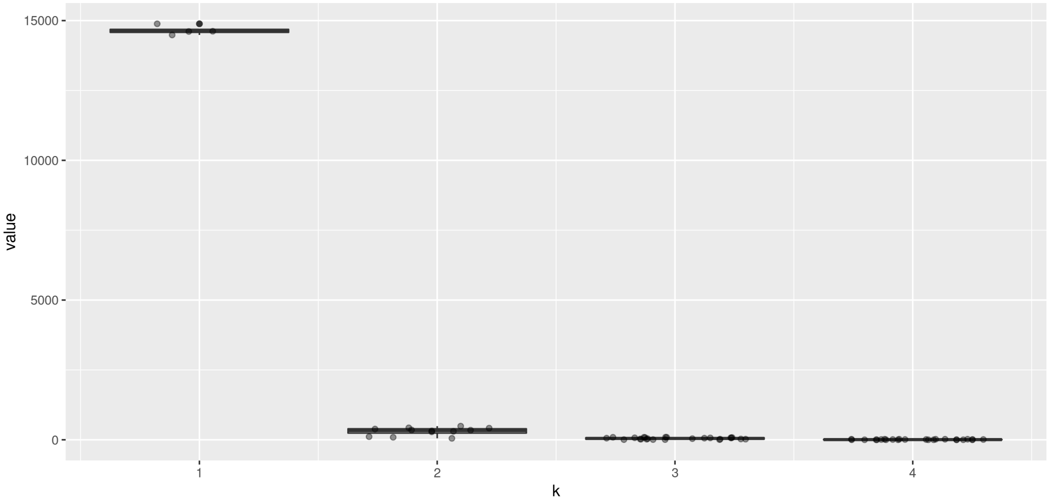
$$I_k \sim \mathrm{Bin}\left(n_0, \mathbb{P}\left\{N = k\right\}\right)$$
$$\mathbb{E}\left[I_k\right] = n_0 \cdot \mathbb{P}\left\{N = k\right\}$$
$$\mathbb{P}\left\{N = k\right\} = (1 - p)^{k-1} \cdot p$$

$$C_k \sim \mathrm{Bin}\left(n_0, \mathbb{P}\left\{N \leq k\right\}\right)$$
$$\mathbb{E}\left[C_k\right] = n_0 \cdot \mathbb{P}\left\{N \leq k\right\}$$
$$\mathbb{P}\left\{N \leq k\right\} = 1 - (1 - p)^k$$

Empirical distribution of C_k for $k = \{1, 2, 3, 4\}$

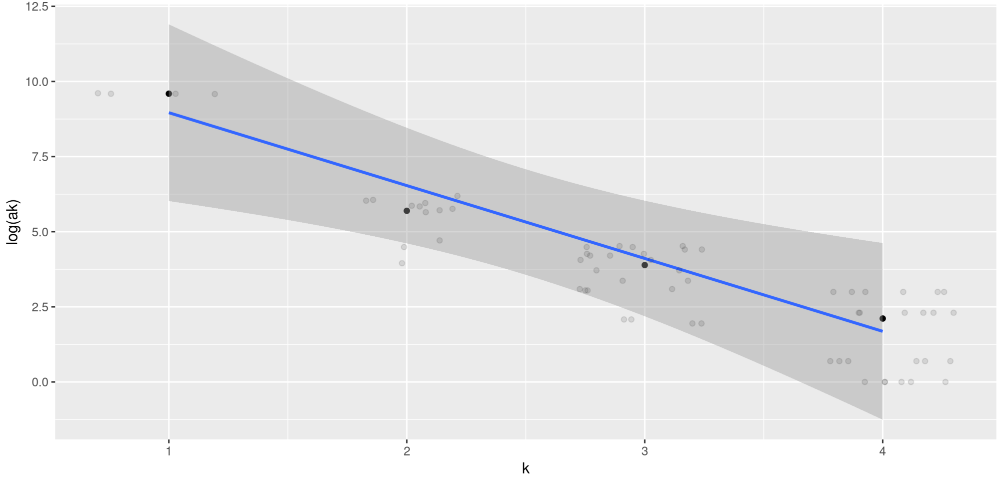


Empirical distribution of l_k for $k = \{1, 2, 3, 4\}$



Empirical distribution of l_k for $k = \{1, 2, 3, 4\}$

Logarithmic y-axis, taking the means



Parameter estimation

$$I_k \sim \mathrm{Bin}\left(n_0, (1-p)^{k-1} \cdot p\right)$$

$$\mathbb{E}[I_k] = n_0 \cdot p \cdot (1-p)^{k-1}$$

$$\log\left(\mathbb{E}[I_k]\right) = [\log(n_0) + \log(p) - \log(1-p)] + [\log(1-p)] \cdot k$$

Parameter estimation

$$I_k \sim \mathrm{Bin}\left(n_0, (1-p)^{k-1} \cdot p\right)$$

$$\mathbb{E}\left[I_k\right] = n_0 \cdot p \cdot (1-p)^{k-1}$$

$$\log\left(\mathbb{E}\left[I_k\right]\right) = \left[\log(n_0) + \log(p) - \log(1-p)\right] + \left[\log(1-p)\right] \cdot k$$

- $\hat{p} \approx 91.15\%$
- $\hat{n}_0 \approx 8539.18$
- $\hat{\mathbb{E}}\left[I_5\right] \approx 0.48$

Parameter estimation

$$I_k \sim \text{Bin}\left(n_0, (1-p)^{k-1} \cdot p\right)$$

$$\mathbb{E}[I_k] = n_0 \cdot p \cdot (1-p)^{k-1}$$

$$\log(\mathbb{E}[I_k]) = [\log(n_0) + \log(p) - \log(1-p)] + [\log(1-p)] \cdot k$$

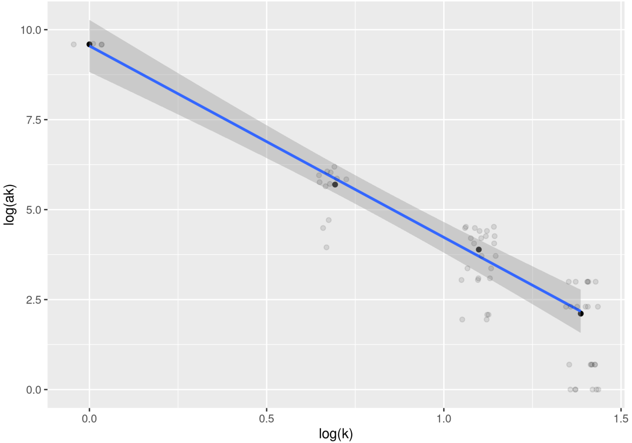
- $\hat{p} \approx 91.15\%$
- $\hat{n}_0 \approx 8539.18$
- $\hat{\mathbb{E}}[I_5] \approx 0.48$

however: not a perfect fit!

- estimates imply negative number of FNs ($\hat{n}_0 \not\geq n = 15{,}005$)
- individual success probability $\hat{p}$ wildly underestimated

An allometric model inspired by the plot

Double-logarithmic axes



$$\log(\mathbb{E}[I_k]) = \log(t) + m \cdot \log(k)$$
$$\Leftrightarrow \mathbb{E}[I_k] = t \cdot k^m$$

- $\hat{p} \approx 97.19\%$
- $\hat{n}_0 \approx 14421.22$
- $\hat{\mathbb{E}}[I_5] \approx 2.68$
- $\hat{\mathbb{E}}[I_6] \approx 1.02$
- $\hat{\mathbb{E}}[I_7] \approx 0.49$

Conclusion

- Anonymization of **sensitive legal documents** (data protection)
- Here: 4-fold annotation by trained student assistants following detailed **guidelines**
- Estimation of expected **false negatives** to ensure annotation quality
- Interim result: **4-7 annotators sufficient**
- **Statistical model** not entirely satisfying

Future work

- Individual success probability p_i for each annotator ($\neq$ assumption c)
- Correlations between annotators ($\neq$ assumption b / d)

	miss-miss	hit-miss	miss-hit	hit-hit
RB-LT	42	352	344	14267
RB-HS	102	417	284	14202
RB-MP	68	52	318	14567
LT-HS	91	428	303	14183
LT-MP	9	111	385	14500
HS-MP	31	89	488	14397

Thank you for your attention!