

AnGer: Automatic anonymisation of German legal documents

Prof. Dr. Stephanie Evert

Korpus- und Computerlinguistik

Friedrich-Alexander-Universität Erlangen-Nürnberg

Prof. Dr. Axel Adrian

Rechtstheorie und -gestaltung



**Funded by
the European Union**
NextGenerationEU



Legal obligation of courts to publish (all) verdicts

Total volume: ca. **1.5 million verdicts** / year

Fundamental right to informational self-determination

→ verdicts need to be **anonymised** for publication

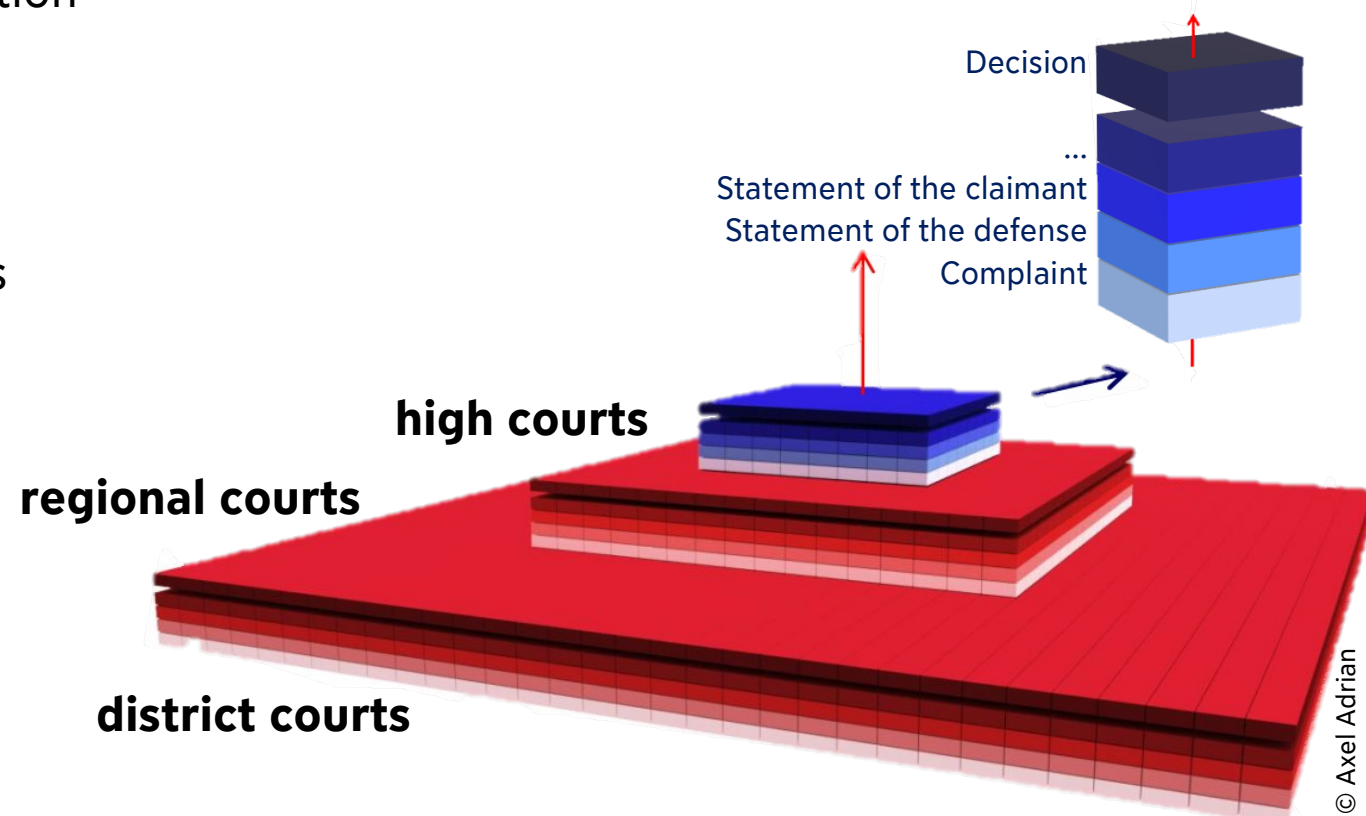
Current situation: < 3% of verdicts are published,
mostly from high courts and higher regional courts

– because manual anonymisation is too expensive

First-instance decisions are underrepresented

– i.e. those discussing the facts of a case

– would be most useful data for legal tech



Relevant legal norms:

- fundamental right to informational self-determination (from Art. 2 Abs. 1 GG)
- GDPR focused on protection of natural persons and their individual rights, legal persons are not included (recital 14 second sentence GDPR)
- § 30 AO (German Fiscal Code)
- §§ 203 ff StGB (German Criminal Code), ...
- successful anonymisation must ensure that **re-identification would require unreasonable effort** (wrt. time, cost, technology)
 - OLG Karlsruhe vom 22.12.2020, 6 VA 24/20
 - VGH Baden-Württemberg vom 23.7.2010, 1 S 501/10

Personally identifying information (PII)

- name (natural and legal persons)
- address, telephone number, license plate, ...
- date of birth & other key dates

Pseudo-identifiers

- profession details
- academic titles
- health data
- information about local environment
- unique features (e.g. only red house in village)
- de-anonymisation by **cross-referencing**

1. Manual anonymisation

- supported by text editor plugins (“office tech”)

2. Semi-automatic anonymisation

- manual validation of computer-generated suggestions → avoid AI act compliance
- self-learning AI improves suggestions over time

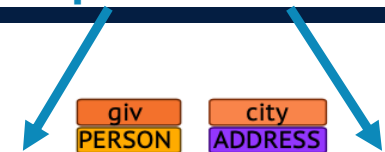
3. Fully automatic anonymisation

- only approach that scales to 1.5M verdicts / year
- existing solutions based on named entity recognition (NER) → ignore pseudo-identifiers

Some existing tools

- [A-Tool](#) (BALO.AI)
 - MS Word plugin used by courts in Switzerland
- EU-funded research project [MAPA](#)
 - **fully automatic anonymisation** based on NER technology
- [OpenRedact](#) (BMBF Prototype Fund)
 - semi-automatic open-source tool with adapted NER
- Text Anonymization Benchmark ([TAB](#))
 - gold standard: EGMR verdicts with semi-automatic annotation
- [HILANO](#) (BMBF kmu-Innovativ)
 - self-learning AI with human-in-the-loop approach
- [JANO](#) (IBM for Hessen & Baden-Württemberg)
 - semi-automatic: suggestions for manual anonymisation

partial match



false negative (FN)

1 Az. : 28 C 45/17 [image8.png] IM NAMEN DES VOLKES In dem Rechtsstreit 1) GROHMANN Hulda , Badener Ring 62 , 94060 Berg - Klägerin - 2) GROHMAN

Alf **Badener Ring 62 , 94060 Berg** Kläger - Prozessbevollmächtigte zu 1 und 2 : Rechtsanwälte SAMMER , MARKUS & KOLLEGEN ,
Kolpingstraße 11 , 49328 Westendorf , Gz . : 48182/52 Qk / Xan , Gerichtsfach-Nr : 44 gegen 1) SCHALLER Hailey , Charlottenstraße 57 , 94513 Schönberg -
Beklagte - 2) SCHALLER Reinhold , Charlottenstraße 57 , 94513 Schönberg - Beklagter - Prozessbevollmächtigte zu 1 und 2 : Rechtsanwältin DR. SCHLICHT
Bettina , Blumenweg 75 , 90763 Fürth , Gz . : 78/221865 wegen Forderung erlässt das Amtsgericht Freyung durch den Richter am Amtsgericht Dittmann am
19. 10. 2017 aufgrund der mündlichen Verhandlung vom 22. 08. 2017 folgendes Endurteil

true positive (TP)

false positive (FP)

https://mapa-demo.pangeamt.com/mapa/1.0/anonymization/model_showcase

Highest quality is required



EU AI act → automatic anonymisation must be provably fit for purpose, accurate, robust, ...

- Adrian/Evert/Heinrich/Keuchen (2024). In *Proceedings of IRIS 2024*. LexisNexis Best Paper Award.

Highest demands on reliability of systems:

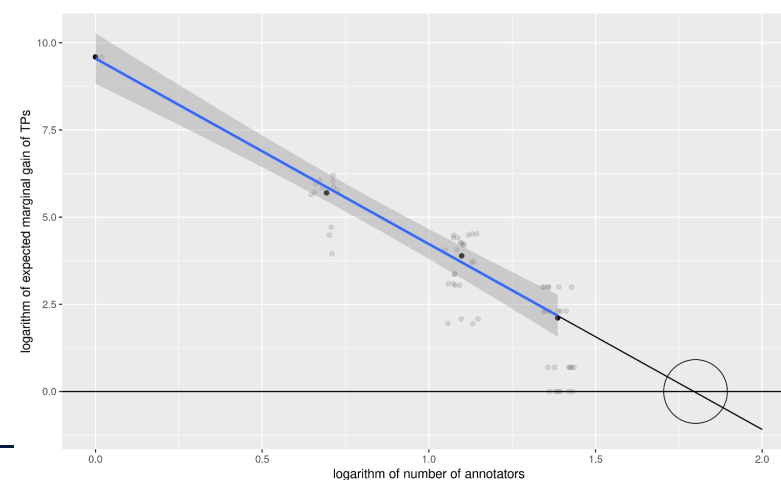
- 95% accuracy sounds great in NLP and AI, but doesn't scale: 5% errors → 75,000 leaks / year
- consequence: need **> 99% recall** for PII
 - key reason why no fully automatic systems are available yet
- relaxed demands on pseudo-identifiers
 - depending on their selectivity & cross-referencing

Gold standard needed

- for training machine learning algorithms (deep learning)
- for evaluation → proven reliability

Extremely high quality demands on gold standard

- can't test > 99% recall if gold standard itself is only 97% correct (which is common in NLP, probably lower for TAB)
- we **need a virtually error-free gold standard**
- empirical model with mathematical extrapolation → need 5-6 annotators for all texts (Heinrich et al. 2021)



LeAK

- funded 2020–2023 by BayStMJ
- team: Nathan Dykes, Philipp Heinrich, Michael Keuchen, Thomas Proisl + student annotators

Goals

- feasibility study on fully automatic anonymisation of court decisions
- focused on district courts (AG) and two legal domains (tenancy, traffic)
- exploration of masking techniques and empirical analysis of de-anonymisation risk

AnGer

- funded 2023–2025 by BMBF/NextGenerationEU
- team: Bao Minh Doan Dang, Philipp Heinrich, Michael Keuchen, Melanie Rosa, Julian Werner, Leonardo Zilio + many student annotators

Goals

- research towards application-ready prototype
- across courts at all levels
- across many legal domains
- domain-adaptation, data augmentation, etc.
- comprehensive de-anonymisation experiments

- manual annotation of sensitive text spans and their information category
- based on detailed annotation guidelines
- 4 independent annotators + 2 adjudicators
- realistic pseudonymisation enables experiments outside protected environment (e.g. HPC)
- based on detailed pseudonymisation guidelines

Here: two gold standards

- district courts (**AG**): tenancy & traffic law
570 decisions, ca. 1M tokens
- higher regional court (**OLG**): 11 legal domains
362 decisions, ca. 1M tokens

Label	Beschreibung / Hinweise	Beispiele	Informationserhaltung
Geschäftszeichen	Akten- und Geschäftszeichen in der Gerichtsentscheidung Nicht gerichtliches Aktenzeichen	Positiv Flurnummern Registernummern Urkundennummern Gz. Gerichtsfach	Nein IdR bei anderen Zeichen, Nr., Gz., usw.

Richtlinien für die Pseudonymisierung von Gerichtsentscheidungen – OLG
Version 6, Stand 31.08.2023

A. Übergeordnete Ziele bei der Pseudonymisierung

- I. An erster Stelle steht der **Schutz der berechtigten Interessen aller Beteiligten und Betroffenen** in den annotierten Gerichtsentscheidungen. Durch die Pseudonymisierung muss mit an Sicherheit grenzender Wahrscheinlichkeit gewährleistet sein, dass die ursprünglich direkt wie indirekt bezeichneten Personen nicht mehr identifizierbar sind.
Ergeben sich Zweifelsfälle oder Konflikte mit den weiteren nachfolgenden Zielen, so steht diese Regel nach einem möglichst maximalen Schutz an erster Stelle.
- II. Als zweites gilt es einen möglichst **realistischen und vergleichbaren Urteilsdatensatz** zu erstellen. Dieser soll die natürlichen Eigenarten, Besonderheiten in der Schreibweise, dem Auftreten von Merkmalen, aber auch typische Fehler, wie Schreibfehler enthalten.
- III. Weiter müssen die Pseudonyme konsistent über die gesamte Entscheidung eingesetzt werden. Der Datensatz soll nach der Pseudonymisierung weiterhin die **logische Konsistenz und Verknüpfungen enthalten**. Einem durchschnittlichen Leser sollte nicht auffallen, dass der Datensatz einer Pseudonymisierung unterzogen wurde.
- IV. Des Weiteren sollen die pseudonymisierten Entscheidungen zu keiner anderen rechtlichen Bewertung führen als dies bei den ursprünglichen Entscheidungen der Fall ist. Durch die Pseudonymisierung soll möglichst keine Veränderung der Rechtslage herbeigeführt werden, die zu einer anderen Entscheidung führen. Die pseudonymisierten Entscheidungen sollen daher auch in rechtlicher Hinsicht logisch und konsistent sein.

B. Vorgehensweise bei der Wahl der Pseudonyme für die einzelnen Merkmalkategorien (Labels aus der Annotation)

- I. Natürliche Personen
 - a. Namen
 - i. Vor- und Nachnamen werden mit Hilfe der Vorschläge aus den Namenslisten für natürliche Personen ersetzt. Auch die Reihenfolge Vorname NACHNAME oder NACHNAME, Vorname usw. wird beibehalten. Nachnamen werden durch Nachnamen und Vornamen durch Vornamen ersetzt. Bei der Auswahl der Namen aus den Listen wird versucht, die ethnische Namensherkunft beizubehalten. Dennoch ist zu beachten, dass die gewählten Kombinationen aus Vor- und Nachnamen realistisch klingen. Gleiches gilt auch für Namensstrukturen, wie Doppelnamen.

Is fully automatic anonymisation possible?

LeAK project (2021)



System	Alle Textstellen			Recall nach Risiko		
	Precision	Recall	F ₁	<i>hoch</i>	<i>mittel</i>	<i>niedrig</i>
Standard-NER (Flair)	0.14	0.12	0.13	0.39	0.31	0.01
Legal-NER (Flair)	0.26	0.16	0.19	0.42	0.28	0.05
OpenRedact	0.49	0.81	0.61	0.87	0.82	0.78
OpenNLP	0.88	0.80	0.84	0.85	0.45	0.83
Riedl & Padó	0.80	0.83	0.82	0.90	0.52	0.85
Fine-tuned GottBERT	0.80	0.90	0.84	0.96	0.80	0.89

Tabelle 5: Evaluation der korrekten Erkennung von Textstellen (Testset: pseudonymisierte Urteile zum Mietrecht)

Results: AG (district courts)

model trained & evaluated on AG gold standard (50% train, 50% test)



Legal domain	Precision	Recall	Recall (high-risk)
tenancy law	97.04%	96.05%	98.90%
traffic law	97.41%	97.38%	99.11%
complete gold standard	97.25%	96.79%	99.03%

Evaluation on OLG (higher regional court)

different text type + wider range of legal domains



Legal domain	Precision	Recall	Recall (high-risk)
Allgemeine Zivilsachen	90.32%	88.91%	98.89%
Bankensachen	94.24%	91.11%	98.70%
Bausachen	90.34%	96.14%	95.71%
Beschwerdeverfahren	86.49%	95.41%	100.0%
Beschwerden (Straf-/Bußgeld)	85.34%	96.81%	95.68%
Familiensachen	85.31%	89.38%	92.81%
Handelssachen	94.57%	95.63%	99.13%
Immaterialgüter	83.86%	81.97%	91.74%
Kostensachen	85.23%	91.84%	100.0%
Schiedssachen	88.31%	85.27%	100.0%
Verkehrsunfallsachen	87.48%	92.74%	98.72%
complete gold standard	88.99%	90.54%	96.98%

Domain adaptation: OLG (higher regional court)

adapted model fine-tuned with OLG training data (50%)



Legal domain	Precision	Recall	OLG-adapted	Recall (high-risk)	OLG-adapted
Allgemeine Zivilsachen	90.32%	88.91%	91.87%	98.89%	99.72%
Bankensachen	94.24%	91.11%	94.61%	98.70%	98.70%
Bausachen	90.34%	96.14%	97.43%	95.71%	98.16%
Beschwerdeverfahren	86.49%	95.41%	97.25%	100.0%	100.0%
Beschwerden (Straf-/Bußgeld)	85.34%	96.81%	97.34%	95.68%	95.68%
Familiensachen	85.31%	89.38%	91.44%	92.81%	92.16%
Handelssachen	94.57%	95.63%	98.19%	99.13%	100.0%
Immaterialgüter	83.86%	81.97%	89.04%	91.74%	93.91%
Kostensachen	85.23%	91.84%	97.96%	100.0%	100.0%
Schiedssachen	88.31%	85.27%	91.43%	100.0%	96.10%
Verkehrsunfallsachen	87.48%	92.74%	97.52%	98.72%	100.0%
complete gold standard	88.99%	90.54%	94.08%	96.98%	97.69%

Results: OLG-adapted model on AG verdicts



Legal domain	Precision	Recall	Recall (high-risk)	OLG-adapted
tenancy law	97.04%	96.05%	98.90%	99.23%
traffic law	97.41%	97.38%	99.11%	99.34%
complete gold standard	97.25%	96.79%	99.03%	99.29%

Hilfe

Risiko

Wie gefährlich die Veröffentlichung der Textstelle wäre.

Niedrig Mittel Hoch

Mindestkonfidenz 0%

Je höher der Wert, desto sicherer ist die automatische Erkennung.

Tags

Nur verwendete Tags anzeigen

Sie können ganze Kategorien ("Tags") hier ausschließen.

- Formales • Aktenzeichen k
- Formales • Gericht g
- Natürliche Person • Name n
- Natürliche Person • Juristischer Funktionsträger t
- Natürliche Person • Identifizierendes Merkmal
- Juristische Person • Name j
- Adresse • Ortsangabe a
- Adresse • Identifizierendes Merkmal
- Fahrzeug • Identifizierendes Merkmal
- Datum • Prozessgeschichte d
- Datum • Sachverhalt s
- Datum • Weltwissen

Amtsgericht Erlangen

Az :11 C 122/20

Mozartstraße 23, 91052 Erlangen

Telefon: 09131/782-01

Telefax: 09131/782-105

Verkündet am: 23.7.2020

(Schneider), JAng.

Urkundsbeamtin d. Geschäftsst.

IM NAMEN DES VOLKES

In dem Rechtsstreit

Patrick Müller, Schillerstraße 24, 91054 Erlangen

- Klägerin zu 1) -

Patricia Müller, Schillerstraße 24, 91054 Erlangen

- Kläger zu 2) -

Prozessbevollmächtigte zu 1), 2):

Rechtsanwälte Heinrich & Kollegen, Züricher Straße 10, 90431 Nürnberg

gegen

Thomas Schütz, Feldstraße 4 d, 91096 Möhrendorf

- Beklagter zu 1) -

Luise Schütz, Feldstraße 4 d, 91096 Möhrendorf

- Beklagte zu 2) -

Prozessbevollmächtigte zu 1), 2):

Steinbrecher + Amberger Rechtsanwälte PartGmbH, Gothestraße 25, 91054 Erlangen

wegen Räumung

Ausgewählter Abschnitt

Mozartstraße

annotieren

Ausgewählte Annotationen

Mozartstraße 23 , 91052 Erlangen

Adresse • Ortsangabe

Akzeptiert Nicht beibehalten alt + b Niedriges Risiko alt + r 99%

Demonstrator

AnGer 2023

corpora.linguistik.uni-erlangen.de/leak/

Amtsgericht Ingolstadt

Az : 13 C 510/20

Haakonweg 61, 85053 Ingolstadt

Telefon: 52838 / 528 - 74

Telefax: 42557 / 137 - 487

Verkündet am: 20. 8. 2020

(Schneider), JAng.

Urkundsbeamtin d. Geschäftsst.

IM NAMEN DES VOLKES

In dem Rechtsstreit

Xaver Heuer , Tuplenstraße 54, 85053 Ingolstadt

- Klägerin zu 1) -

Elodie Heuer , Tuplenstraße 54, 85053 Ingolstadt

- Kläger zu 2) -

Prozessbevollmächtigte zu 1), 2):

Rechtsanwälte Hanke & Kollegen , Pilatusweg 32, 90491 Nürnberg

gegen

Hellmut Johann , Seeuferstraße 35e, 85132 Schernfeld

- Beklagter zu 1) -

Thea Johann , Seeuferstraße 35e, 85132 Schernfeld

- Beklagte zu 2) -

Realistisch Abkürzen Schwärzen

Nur Fehlstellen anzeigen

Automatische Mehrfachaktualisierung

Amtsgericht Erlangen	>	Amtsgericht Ingolstadt	
Formales • Gericht			
<input type="text" value="Amtsgericht Ingolstadt"/> <input checked="" type="checkbox"/>			
11 C 122/20	>	13 C 510/20	
Formales • Aktenzeichen			
Mozartstraße 23, 91052 Erlangen	>	Haakonweg 61, 85053 Ingolstadt	
Adresse • Ortsangabe			
09131/782-01	>	52838 / 528 - 74	
Formales • Aktenzeichen			
09131/782-105	>	42557 / 137 - 487	
Formales • Aktenzeichen			
23.7.2020	>	20. 8. 2020	
Datum • Prozessgeschichte			
Patrick Müller	>	Xaver Heuer	
Natürliche Person • Name			
Schillerstraße 24, 91054 Erlangen	>	Tuplenstraße 54, 85053	

Demonstrator

AnGer 2023

corpora.linguistik.uni-erlangen.de/leak/

Beyond legal documents

Performance of AnGer 2023 without domain adaptation



Hilfe

Risiko ?
Wie gefährlich die Veröffentlichung der Textstelle wäre.

Niedrig Mittel Hoch

Mindestkonfidenz 0% ?
Je höher der Wert, desto sicherer ist die automatische Erkennung.

Tags ? Nur verwendete Tags anzeigen
Sie können ganze Kategorien ("Tags") hier ausschließen.

- Sonstiges • Sonstiges
- Natürliche Person • Identifizierendes Merkmal
- Juristische Person • Name j
- Adresse • Ortsangabe a
- Adresse • Identifizierendes Merkmal
- Datum • Sachverhalt s

Das ist die Motivation hinter Hersheys neuer Partnerschaft mit der amerikanischen Behörde für Entwicklungszusammenarbeit (USAID) und ECOM, unserem größten Kakaolieferanten in Ghana.

Letztes Jahr riefen wir ein kleines Pilotprogramm ins Leben, um Kleinbauern dabei zu helfen, ihre Kakaoproduktion zu steigern, die durch den Kakaoanbau bedingte Abholzung zu beenden und die Widerstandskraft der Kakaobäume zu verbessern.

Die Initiative richtet ihr Augenmerk insbesondere auf zwei Herausforderungen, mit denen jeder Anbaubetrieb in Westafrika konfrontiert ist: Grundbesitz und Finanzierung.

Laut Angaben der Lands Commission in Ghana haben weniger als 2 Prozent der 800.000 Kakaobauern des Landes einen rechtlichen Anspruch auf den von ihnen bewirtschafteten Grund und Boden.

Vielmehr erhalten sie durch informelle Absprachen mit lokalen Anführern oder Grundbesitzern Zugang zu landwirtschaftlichen Flächen.

Traditionellerweise ermöglichten diese mündlichen Vereinbarungen den Bauern die Rodung von Wäldern und die landwirtschaftliche Nutzung der entstandenen Flächen.

Wenn die Kakaobäume allerdings nach etwa 30 Jahren – oder im Falle von Krankheiten früher – keinen Ertrag mehr abwerfen, müssen die Bauern vom ursprünglichen Grundbesitzer eine Genehmigung zur Wiederbepflanzung einholen.

In Zeiten einer historisch hohen Nachfrage nach Grund und Boden weigern sich lokale Anführer und Landbesitzer zunehmend, dem Wunsch der Bauern nach Wiederbepflanzung zu entsprechen.

Den Bauern bleiben zwei Möglichkeiten, von denen keine wünschenswert ist: entweder Urwälder roden und von vorne beginnen oder sich gänzlich aus dem Geschäft verabschieden.

Die Partnerschaft mit USAID und ECOM zielt darauf ab, dieses Problem durch die Beseitigung mancher Hürden für die Wiederbepflanzung zu lösen.

ECOM entwickelte ein innovatives Finanzierungsmodell, das den Bauern hilft, alte oder kranke Bäume zu entfernen und sie durch widerstandsfähige und ertragreichere Hybridsorten zu

German online news

Hilfe

Risiko ?
Wie gefährlich die Veröffentlichung der Textstelle wäre.

Niedrig Mittel Hoch

Mindestkonfidenz 0% ?
Je höher der Wert, desto sicherer ist die automatische Erkennung.

Tags ? Nur verwendete Tags anzeigen
Sie können ganze Kategorien ("Tags") hier ausschließen.

- Sonstiges • Sonstiges
- Formales • Aktenzeichen k
- Natürliche Person • Name n
- Natürliche Person • Identifizierendes Merkmal
- Juristische Person • Name j
- Adresse • Ortsangabe a
- Datum • Prozessgeschichte d
- Datum • Sachverhalt s

Bisher galt es als böse Verschwörungstheorie und „Corona-Ketzerei“, wenn jemand behauptete, wir hätten es bei Corona mit einer „Test-Pandemie“ zu tun – und ohne die ausgiebigen Tests wäre die Situation nie so eskaliert und es gäbe keine Corona-Maßnahmen, wie wir sie kennen.

Wer diese Position vertritt, muss sich auf Ausgrenzung, Diffamierung und Hass einstellen.

Und das, obwohl sie kein Geringeres als das österreichische Gegenstück zum deutschen RKI-Chef Lothar Wieler vertritt: „Ohne PCR-Tests wäre die Pandemie niemandem aufgefallen“, sagte Professor Dr. Franz Allerberger, der Leiter der österreichischen Agentur für Gesundheit und Ernährungssicherheit (AGES).

Ich habe hier bereits im Juni darüber berichtet.

In den großen Medien wurde die Aussage aus berufenem Munde allerdings so gut wie totgeschwiegen.

Wie so vieles, was nicht in das offizielle Corona-Narrativ passt.

Und nun das!

Kein Geringeres als Jens Spahn, seines Zeichens Gesundheitsminister und einer der obersten Betreiber der Corona-Panik, sagt kaum verlausuliert nichts anderes als der „österreichische Wieler“.

Und wieder hören die meisten weg, und man tut so, als sei die Aussage gar nicht gefallen.

Dabei erfolgte sie vor einem großen Publikum – im „Ersten“.

In der Talkshow „Hart aber fair“ meinte da der Minister am 30. August: „Wenn wir geimpfte Menschen auch genauso testen, wie ungeimpfte, dann hört diese Pandemie nie auf.“

Damit zerstört der Minister regelrecht seinen eigenen Corona-Kurs – mit einem einzigen Satz. Man muss ihn einfach nur logisch zu Ende denken.

Das große Glück des Ministers ist aber, dass dies offenbar zumindest in den großen Medien offenbar kaum jemand tut.“

German Telegram

Thanks for listening!

<https://www.linguistik.phil.fau.de/projects/leak-anger/>