

Rechtsvergleich mit Hilfe Künstlicher Intelligenz (KI)

Posted on 26. Januar 2026 by Klaus F. Röhl

Mustererkennung ist eine herausragende Fähigkeit fortgeschrittener KI. Den »strukturellen« Vergleich beherrscht KI jedoch bisher kaum. Daher gilt es zu überlegen, was man von einem Ähnlichkeitsvergleich innerhalb eines Rechtssystems und über mehrere Rechtssysteme hinweg erwarten kann. KI bedeutet sicher nicht das Ende der Rechtsvergleichung, wird das Feld aber doch grundsätzlich verändern.[\[1\]](#) Was man erwarten kann, hängt wiederum von den verfügbaren Methoden[\[2\]](#) ab.

Ich will den Rechtsvergleich im Auge behalten, aber mich zuvor informieren, was KI insoweit leistet. Diese Fortsetzung dient deshalb zunächst der Selbstverständigung. Der Kenner mag sie überschlagen.

Natural Language Processing (NLP), also die direkte Verarbeitung natürlicher Sprache, ist heute KI-Standard.[\[3\]](#) Es gibt viele juristische Texte – Gesetze, Urteile, Literatur – die maschinenlesbar zur Verfügung stehen und die man mit Hilfe von KI befragen könnte. Doch kann man den Ähnlichkeitsvergleich juristischer Texte bisher anscheinend nicht vollständig Verfahren automatisierter Inhaltsanalyse überlassen, die selbstständig nach Mustern suchen. Juristischen Datenbanken gestatten die Informationsgewinnung nur mit Hilfe einer Stichwortsuche. Auf diese Weise kann man wohl einschlägige Texte finden, aber keine Muster entdecken.[\[4\]](#) Juristische Texte haben Eigenschaften, die die automatische Extraktion von Strukturen erschweren. Juristisches Denken stützt sich stark auf implizites Wissen, das derzeit schwer zu extrahieren ist. Rechtsnormen sind voll von unausgesprochenen Verweisungen. Sie sind sowohl auf Satzebene als auch auf der Wortebene interpretationsfähig und interpretationsbedürftig.[\[5\]](#) Das Verständnis verlangt die Berücksichtigung eines (sprachlichen) Kontextes, der über die üblicherweise von den LLM berücksichtigten Textabschnitte und auch über den einzelnen Satz hinausgeht. Daher steckt die Automatisierung der semantischen Suche in juristischen Texten noch immer in den Anfängen. Aber die haben es in sich.

Um sich den Möglichkeiten des automatisierten Ähnlichkeitsvergleichs zu nähern, gilt es, zunächst Vergleichsobjekte und Vergleichsmengen näher zu bestimmen. Die

Bestimmung der Vergleichsmengen ist die einfachere Aufgabe.

Vergleichsmenge ist eine Menge juristischer Texte, die größer ist, als dass man sie ohne Hilfsmittel übersehen könnte. Es ist allerdings ausgeschlossen, alle juristischen Texte zu berücksichtigen. Je nach Fragestellung muss man eine Auswahl treffen. Findet sich das Vergleichsobjekt V_1 häufiger in Normen, die vor oder nach dem Jahr 2000 erlassen wurden? Findet sich das Vergleichsobjekt V_2 nicht nur in Entscheidungen des Bundesverfassungsgerichts, sondern auch in solchen des US-Supreme Courts usw.? Die Auswahl kann sich auf eine der drei großen Gruppen – Normtexte, Entscheidungen, Literatur – beschränken. Sie kann nach der Entstehungszeit der Texte begrenzt werden. Sie kann sich auf bestimmte Sach- und Fachgebiete konzentrieren. Für die Rechtsvergleichung wird man Texte aus zwei oder mehreren Rechtssystemen einbeziehen. So ergeben sich vielfache Möglichkeiten für Schnittmengen.

Eine technische Begrenzung der Vergleichsmenge folgt aus der Verfügbarkeit der Daten. Man kann davon ausgehen, dass die juristischen Texte mindestens der letzten Jahrzehnte digital entworfen und gespeichert wurden. Aber das bedeutet längst nicht, dass sie für eine automatisierte Auswertung durch Dritte verfügbar sind. Viele Texte werden überhaupt nicht digital publiziert.^[6] Viele stehen hinter einer Bezahlschranke. Wo juristische Texte in digitaler Form öffentlich zugänglich sind, wie es heute bei obergerichtlichen Entscheidungen in der Regel der Fall ist, müssen sie meist einzeln abgerufen werden. So genannte Bulk-Downloads sind kaum möglich. Diese Einschränkungen treffen allerdings vor allem die KI-Industrie, die solche Texte gerne zum Training domänenspezifischer LLM verwenden möchte. Ihr fehlt die technische Schnittstelle (API), die den direkten Zugriff auf größere Datenmengen erlaubt.^[7] Für den wissenschaftlichen Bedarf gelingt es meist, kleinere Teilmengen zusammenzustellen. Hier liegt das Problem eher darin, die Texte für das maschinelle Lernen vorzustrukturieren,

Schwieriger als die Bereitstellung einer Textmenge zur automatisierten Auswertung ist die Bestimmung von Vergleichsobjekten, jedenfalls dann, wenn diese nicht trivial sein sollen. Das ist ja der Grund, warum die Rechtsvergleichung weitgehend auf den Ähnlichkeitsvergleich verzichtet. Trivial erscheint zunächst, was eine Standard-Methode der KI reproduziert, die als Named Entity Recognition (NER) geläufig ist. Für dieses Verfahren wird eine Reihe »benannter Entitäten« codiert, nach denen das Programm in den Texten suchen soll. Zu den Entitäten gehören standardmäßig Orts- und Zeitangaben, Personen und Organisationen.

Interessanter wird es, wenn dem Programm domänenspezifische Kategorien vorgegeben werden (Legal Entity Recognition = LER).[\[8\]](#) Juristische Dokumente enthalten – neben den üblichen Eigennamen (Personen, Orte, Organisationen) – zahlreiche fachspezifische Entitäten. Für die Suche in juristischen Texten etwa

Rechtssubjekte: Parteien, Richter, Anwälte,

Normen: Gesetze, Paragraphen, Absätze, Artikel, Verweisungen,

Verträge: Vertragsparteien, Klauseln, Fristen, Zahlungsbedingungen,

Gerichtliche Entscheidungen: Entscheidungsart, Instanz, Datum, Aktenzeichen,

Ort und Jurisdiktionseinheiten: Land, Bundesland, Rechtskreis.

Aber so richtig spannend sind solche Fragen noch nicht. Sie kratzen immer noch an der Oberfläche. Interessant ist immerhin, dass zwischen Vergleichsobjekt und Vergleichsmenge eine Relation zu bestehen scheint derart, dass bei sehr großen Vergleichsmengen auch oberflächliche Ähnlichkeiten Bedeutung gewinnen. Das zeigt das von der Politikwissenschaft beschriebene Phänomen der Isomorphie der Institutionen.

Damit die Ähnlichkeitssuche in begrenzten Textmengen interessant wird, müssen weitere Entitäten benannt werden, die trotz ihrer Äußerlichkeit zu tieferen Einsichten verhelfen können. Einige solcher Vergleichsobjekte drängen sich auf. Drei will ich hier anführen.

Friedemann Vogel hat darauf aufmerksam gemacht, dass der Anteil »mehr oder weniger feststehender Mehrworteinheiten« in Gerichtsentscheidungen deutlich häufiger ist als in Presse- oder sprachwissenschaftlichen Texten.[\[9\]](#) Ein zweites Muster bilden die nach dem Vorbild von Code Smells so genannten Law Smells, nämlich suboptimale Gestaltungen von Rechtssätzen. Das dritte Muster, dessen Relevanz auf der Hand zu liegen scheint, sind »Fälle«. Dieses Muster ist interessant, weil vielleicht über Fälle die Verbindung zu (funktionalen) Problemen hergestellt werden kann.

Zunächst zu den Mehrworteinheiten. Sie erscheinen in unterschiedlicher Zusammensetzung in verschiedenen Vergleichsmengen, die aus der Grundmenge der Rechtstexte jeweils einen *consideration set* bilden. Ganz grob lassen sich die Vergleichsmengen für die Suche nach Mehrworteinheiten in Rechtsnormen,

Gebrauchstexte wie Verträge und Geschäftsbedingungen sowie dogmatische Texte im weitesten Sinne einteilen. Die letztere Menge soll alle Text einschließen, die irgendwie die Verbindung zwischen Rechtsnormen und Entscheidungen herstellen, also etwa Urteilsbegründungen, Kommentare und das dogmatische Schrifttum im engeren Sinne. In Normtexten kann man etwa nach gleichlautenden Begriffen und Definitionen suchen, in Gebrauchstexten nach Vertragsklauseln und in dogmatischen Texten nach Formeln, die Prinzipien aufrufen (z.B. »Integrität und Vertraulichkeit informationstechnischer Systeme«[\[10\]](#), BVerfG) oder als Argument dienen (z. B. Verhältnismäßigkeit).

Die inzwischen schon »klassische« Datenverarbeitung findet solche Mehrworteinheiten allerdings nur, wenn sie mit genau vorgegebenem Wortlaut gesucht werden. Das ist bloßes Retrieval mit Hilfe der Booleschen Operatoren. Von fortgeschrittener KI erwartet man eine semantische Mustersuche. Das heißt, sie sollte die Vergleichsobjekte auch herausfinden, wenn sie nicht nach dem Wortlaut, aber sinngemäß ähnlich sind. Ein Verhältnismäßigkeitsargument müsste also auch erkannt werden, wenn von Proportionalität die Rede ist oder gar nur davon gesprochen wird, dass im konkreten Fall der Zweck das Mittel nicht rechtfertigen kann. Um diese Leistung zu erbringen, können die großen Sprachmodelle mit annotierten Textcorpora trainiert werden. Über solches Training und seinen Erfolg berichten *Kilian Lüders, Stohlmann* u. a..[\[11\]](#) Um besser zu verstehen, worum es geht, sind zunächst einige Grundprinzipien der KI in Erinnerung zu rufen.

Die verbreitete Vorstellung »vollautomatischen Lernens« trifft in der Praxis nicht zu. Die KI-Industrie beschäftigt in erheblichem Umfang Menschen für Datenannotation und -aufbereitung. Dieses Segment ist ein struktureller Bestandteil moderner KI-Entwicklung, insbesondere bei Deep Learning. Deep-Learning-Modelle benötigen für viele Aufgaben gelabelte Trainingsdaten. Bilder mit müssen Objektmarkierungen versehen werden, Texte mit Kategorien, Entitäten und und weiteren Kennzeichnungen. Diese Label können, zumal für Rechtstexte, nicht zuverlässig automatisch erzeugt werden. Die KI-Industrie beschäftigt, unsichtbar für Endnutzer, in Billiglohnländern Hunderttausende Menschen weltweit. Für domänenspezifische Annotation braucht man jedoch Experten. In der Folge sind die juristischen Textcorpora für das Training von KI knapp. In der Regel werden Texte erst für konkrete Forschungen entsprechend aufbereitet, und oft werden sie anschließend nicht für die Weiterverwendung zu Verfügung gestellt.

Wie gesagt: Natural Language Processing (NLP), also die direkte Verarbeitung natürlicher Sprache, ist heute KI-Standard. Der Jurist als Informatik-Laie denkt hier

an die großen, teilweise auch öffentlich zugänglichen Systeme der generativen KI wie ChatGPT, Gemini, Claude oder Perplexity. Die Stärke dieser Systeme ist ihre Fähigkeit zum Umgang mit natürlicher Sprache. GPT steht für einen generativen vortrainierten Transformer (*generative pre-trained transformer*). Ein Transformer ist eine von Google entwickelte Deep-Learning-Architektur. Sie »versteht« Sprache, indem sie sich nicht nur separate Wörter merkt, sondern die Wahrscheinlichkeit von Wortfolgen berücksichtigt. Dadurch wird die Bedeutung eines Wortes in seinem Kontext erfasst. Allerdings übernehmen die Programme nicht die Wörter nicht unmittelbar aus Texten, sondern »tokenisieren« sie zuvor, das heißt zerlegen sie in die bedeutungstragenden Bestandteile sowie Vorsilben und Endungen.

BERT-Modelle berechnen die Wahrscheinlichkeiten vorwärts und rückwärts, also »bidirektional«. Für jede Eingabesequenz wird das Attention-Maß eines jeden Wortes (Tokens) zu jedem anderen berechnet. Typische Transformer-Modelle verarbeiten Sequenzen mit einer Länge von 512, 1024 oder 2048 Token. Diese Begrenzung entsteht durch den hohen Rechen- und Speicheraufwand, da für jedes Token sogenannte Query-, Key- und Value-Vektoren berechnet werden, so dass der Aufwand für die Attention-Berechnung quadratisch mit der Sequenzlänge wächst. Wie die Werte aus mehreren Eingabesequenzen und darüber hinaus aus verschiedenen Texten koordiniert und zusammengeführt und schließlich auf Anfragen (Prompts) reproduziert werden, ist dann die Spezialität der einzelnen Programme. Weiterer Programmarbeit bedarf es dann, wenn einzelnen Token mit Hilfe von *named entities* oder Mehrworteinheiten explizit semantischer Gehalt zugewiesen wird. Wichtig ist schließlich, dass die gespeicherten Werte »trainiert« werden können, das heißt, wenn weitere Texte durch den Transformer geschickt werden, verbessert sich die Kontextempfindlichkeit.

Der Jurist als Informatik-Laie hat zunächst die *general-purpose chatbots im Blick*, die großen Allzweck-LLM, die, wie man hört, mit gewaltigen Textmengen trainiert worden sind, und die, wie man inzwischen täglich erfährt, auf Fragen aller Art erstaunliche Antworten generieren. Eine Probe bietet der [Eintrag vom 23. 11. 2025](#). Auch auf juristische Fragen bleiben die Systeme keine Antwort schuldig. Freilich kann man sich auf die Antworten nicht verlassen. Ich habe mehrfach erlebt, dass überholte Rechtsquellen angeführt und Belege erfunden wurden. Aber die genannten Systeme erheben auch gar nicht den Anspruch, Experten zu ersetzen. Näher dran sind domänenspezifische Expertensysteme, von denen längst eine ganze Reihe am Markt ist, z. B. Beck-online, Libra. Diese Systeme kombinieren vortrainierte Sprachmodelle mit domänenspezifischem Training und Human-in-the-

Loop-Mechanismen, etwa durch fachliche Validierung der Ergebnisse oder gezielte Nachannotation problematischer Fälle. Ich habe zu diesen Systemen nur begrenzten Zugang und damit bisher keine Erfahrungen gemacht, über die sich zu berichten lohnt.

NER, LER und das Training mit annotierten Textcorpora helfen, Texte nach vorgegebenen Vergleichsobjekten = Mustern zu durchsuchen, führen aber noch nicht zu selbständigen Mustererkennung. Es handelt sich immer noch um maschinelles Lernen mit explizit definierten Merkmalen (*features*). Der Anspruch der fortgeschrittenen KI-Systeme geht dahin, gedankliche Muster zu erkennen, auch ohne dass diese Muster vorgegeben und die LLM darauf trainiert werden (*Deep Learning*). Es könnte wohl sein, ja es ist sogar zu vermuten, dass in Rechtstexten Ähnlichkeiten vorhanden sind, die bislang noch nicht bewusst wahrgenommen und schon gar nicht Es erscheint daher plausibel – ja sogar naheliegend –, dass in Rechtstexten Ähnlichkeiten und Strukturen vorhanden sind, die bislang nicht bewusst wahrgenommen und schon gar nicht systematisch erfasst worden sind.

Zurück zur Rechtsvergleichung führt noch ein kleiner Umweg über die Internetseite <https://huggingface.co>. HuggingFace ist eine von mehreren Open-Source-Programm-Bibliotheken für künstliche Intelligenz. Sie bietet Open-Source-Implementierungen von Transformer-Modellen für Text-, Bild- und Audiodaten. Es handelt sich um vortrainierte Sprachmodelle, das heißt Programmbausteine mit offenen Schnittstellen (API). Sie sind über die Transformer-Architektur soweit vortrainiert, dass sie natürliche Sprache verarbeiten. Über die Schnittstellen können sie für ein bestimmtes Sach- oder Fachgebiet (eine »Domäne«) oder für bestimmte Aufgaben weiter trainiert werden. Das Training kann also entweder allgemein zu einem besseren Verständnis der Fachsprache oder speziell zur Erkennung spezifischer Muster geführt werden. Für das Training benötigte man annotierte Dateien, das heißt solche, in denen die Sequenzen, denen das Programm besondere Aufmerksamkeit schenken soll, gekennzeichnet (gelabelt) werden. Diese Annotierung ist Hand- und Fleißarbeit, die vorab geleistet wird. Danach bieten viele Programme die Möglichkeit zur Feinarbeit durch *human-in-the-loop*. Das heißt, das Modell erzeugt zunächst Vorschläge, die von menschlichen Experten überprüft, korrigiert und iterativ in das Training zurückgeführt werden.

Gleich mehrere vortrainierte Programme, wie sie auf HuggingFace angeboten werden, wurden auch von den Forschungsarbeiten benutzt, auf die ich in einer weiteren Fortsetzung eingehen will.

[1] *Lutz-Christian Wolff*, Artificial Intelligence ante portas: The End of Comparative Law?, *The Chinese Journal of Comparative Law* 7, 2019, 484–504.

[2] Für einen neueren Überblick *Daniel Fürst/Mennatallah El-Assady/Daniel A. Keim/Maximilian T. Fischer*, Challenges and Opportunities for Visual Analytics in Jurisprudence, Artificial Intelligence and Law 2025.

[3] So schon vor drei Jahren für die Journalismusforschung *Valerie Hase/Daniela Mahl/Mike S. Schäfer*, Der »Computational Turn«: ein »interdisziplinärer Turn«? Ein systematischer Überblick zur Nutzung der automatisierten Inhaltsanalyse in der Journalismusforschung, *M&K* 2022, 60–78.

[4] Durch die Programmierung von Regelausdrücken ([Regex](#)) können auch Fundstellen wie Aktenzeichen oder Normbezeichnungen und Zitate gefunden werden.

[5] *Clement Guitton/Aurelia Tamò-Larrieux/Simon Mayer/Gijs van Dijck*, The Challenge of Open-Texture in Law, *Artificial Intelligence and Law* 33, 2025, 405–435.

[6] *Hanjo Hamann*, Der blinde Fleck der deutschen Rechtswissenschaft – zur digitalen Verfügbarkeit in-

stanzgerichtlicher Rechtsprechung, *JZ* 2021, 656, 658.

[7] Es gibt zwei erstaunliche Privatinitiativen, die sich um solche Zugangsmöglichkeiten bemühen, die Rechtsprechungsdatenbank OpenJur (<https://openjur.de/>) und der Rechtswissenschaftler *Sean Fobbe* mit seiner Internetseite Open Data (<https://seanfobbe.com/de/data/>).

[8] *Elena Leitner/Georg Rehm/Julian Moreno-Schneider*, Fine-Grained Named Entity Recognition in Legal Documents, in: *Maribel Acosta/Philippe Cudré-Mauroux/Maria Maleshkova/Tassilo Pellegrini/Harald Sack/York Sure-Vetter*, *Semantic Systems. The Power of AI and Knowledge Graphs*, 2019, 272–287. Für ein kommerzielles Angebot aus den USA vgl. [The Ultimate Guide to Recognizing Legal Entities with Legal NLP](#).

[9] *Friedemann Vogel*, Calculating Legal Meanings? Drawbacks and Opportunities of Corpus-Assisted Legal Linguistics to Make the Law (more) Explicit, in: *Janet Giltrow/Dieter Stein* (Hg.), *The Pragmatic Turn in Law*, 2017, 287–306 (hier zitiert aus Preprint: https://www.data.friedemann-vogel.de/texte/FVogel_CalculatingLegalMeanings_06032015_preprint.pdf); ders., Der Richter, (k)ein Bot?!, in: *Liane Wörner/Rüdiger Wilhelmi/Jochen Glöckner/Marten Breuer/Svenja Behrendt* (Hg.), *Digitalisierung des Rechts*, 2024, 9–26, S. 12.

[10] Die Formulierung erscheint zwölf Mal in BVerfGE 120, 274.

[11] *Kilian Lüders/Bent Stohmann*, Classifying Proportionality – Identification of a Legal Argument, *Artificial Intelligence and Law* 2025, 1051–1078.

Ähnliche Themen

- [Automatisierte Rechtsvergleichung](#)
- [Vergleichsobjekte und Vergleichsmengen](#)
- [Der Vergleich des Vergleichs als Weg zur Interdisziplinarität](#)
- [Travelling Models VIII: Nun kommt man auch in Frankfurt auf den Trichter.](#)
- [Travelling Models I: Rechtsvergleichung](#)