Dr. Bettina Mielke / Professor Dr. Christian Wolff

Möglichkeiten und Perspektiven der Korpuslinguistik für die Analyse von Rechtstexten

Recht und Sprache bilden einen untrennbaren Zusammenhang, wobei nicht die gesprochene Sprache, sondern die schriftliche Sprache und damit Texte im Vordergrund stehen. Da juristische Texte mittlerweile nahezu durchgehend in digitaler Form vorliegen, ist es naheliegend, korpus- bzw. computerlinguistische Verfahren zu ihrer Analyse heranzuziehen. Unterschiedliche Ziele, die vom automatisierten Erkennen inhaltlicher Zusammenhänge bis zur Anonymisierung von Urteilen reichen, können dadurch unterstützt werden.

Einführung

Unabhängig von der Frage, ob die Rechtswissenschaft als Geisteswissenschaft anzusehen ist – dies wird nicht einheitlich gesehen – spielt die Auseinandersetzung mit (Rechts-)Texten bei der juristischen Tätigkeit eine ebenso zentrale Rolle wie bei den Geisteswissenschaften. In diesen textzentrierten Wissenschaften sind einerseits *inhalts*bezogene Analysen von Interesse: Welche Konzepte in einem Dokument sind wichtig? Wie ähnlich sind die Inhalte verschiedener Dokumente? Welches Dokument passt am besten zu einer Suchanfrage? Andererseits gibt es zunehmend Untersuchungen, die *stilistische* oder auch *emotionale* Aspekte in Texten herauszuarbeiten versuchen (*sentiment analysis*).
Hinzu kommen Fragestellungen hinsichtlich der Zuordnung von Texten zu bestimmten *Genres* oder der Weiterentwicklung von einzelnen (Text-)Gattungen. Ein typischer Anwendungsfall für stilometrische Analysen ist etwa die Überprüfung oder Feststellung von Autorenschaft: Von welchem Autor wurde ein Text geschrieben? Welche Autoren haben welche Teile eines Textes verfasst?[1] Derartige Analysen sind mittlerweile auch im juristischen Kontext erfolgt.

Ausgangspunkt Digitalisierung

Seit einiger Zeit stehen zunehmend digitale Textressourcen auch im Rechtswesen bereit, z. B. digitale Ausgaben juristischer Fachliteratur, Urteils- und Normdatenbanken, Akten in elektronischer Form, so dass sich gute Voraussetzungen für den Aufbau von Dokumentkollektionen/Korpora als Grundlage texttechnologischer Anwendungen ergeben. Mit der Verfügbarkeit großer digitaler Textkorpora aus einigen Zehntausenden, Hunderttausenden oder gar Millionen von Dokumenten ist die Hoffnung verbunden, andere Erkenntnisse gewinnen zu können als bei der traditionellen, auf notwendigerweise wenige Dokumente konzentrierten intellektuellen Textinterpretation.
Der amerikanische Literaturwissenschaftler Franco Moretti hat dafür den Begriff des „distant reading“[2] geprägt, sozusagen das Lesen einer großen Menge von Texten durch den Computer im Unterschied zum „close reading“ des Menschen, der eine kleine Textmenge intensiv studiert. Die unten beschriebenen Projekte im deutsch- und englischsprachigen Raum, bei denen jeweils Textkorpora aus dem Rechtswesen untersucht wurden, geben eine Vorstellung von den Möglichkeiten.

Korpuslinguistische Methoden

Die Korpuslinguistik ist ein Teilgebiet der (angewandten) Sprachwissenschaft, die zum Ziel hat, mit Hilfe von Computern (und damit als Teil der Computerlinguistik) Sprach- und Textkorpora standardisiert und repräsentativ aufzubauen und auszuwerten.[3] Derartige Korpora können von Verfahren und Anwendungen der Sprach- und Texttechnologie (engl.

1) *Eder*, Rolling Stylometry, Digital Scholarship in the Humanities, 31 (2016), S. 457–469.
2) *Moretti*, Distant reading. Verso Books, 2013.
3) *Lemnitzer/Zinsmeister*, Korpuslinguistik: Eine Einführung, 2015.

Natural Language Processing, NLP) zu ganz unterschiedlichen Zwecken genutzt werden, beginnend mit der Berechnung einfacher numerischer Parameter (z. B. Worthäufigkeiten, Dokumentlänge, mittlere Satzlänge) bis hin zur Extraktion komplexerer Informationen (Bestimmung der Lesbarkeit, Gewinnung formaler Angaben wie Zitationen, Normangaben oder Querverweisen).

Zu den wichtigsten Arbeitsschritten auf dem Weg von einer Dokumentsammlung zu einem einheitlichen Textkorpus zählen dabei:

- Textaufbereitung (einheitliche Kodierung, Entfernung von Sonder- und Steuerzeichen z. B. der Formatierung, ggf. auch Entfernung nicht-sinntragender Wörter, sog. Stoppwörter),
- Segmentierung der Texte in einzelne Sätze und Wörter,
- (optional) Reduzierung gebeugter Formen auf die Stammform (Lemmatisierung) bzw. einfaches Abschneiden von Flexionsendungen (*stemming*),
- Aufbau eines Index, der Informationen darüber enthält, wie häufig (und ggf. auch wo) welches Wort in welchen Texten der Kollektion vorkommt (invertierte Liste, entspricht in etwa dem Index einer Suchmaschine).

Mit Hilfe der Informationen über die Häufigkeit eines Begriffs in einzelnen Dokumenten (Termfrequenz, TF) bzw. der Anzahl der Dokumente, in denen ein Begriff vorkommt (Dokumentfrequenz, DF), lassen sich Metriken wie das Maß Termfrequenz x inverse Dokumentfrequenz (IDF) berechnen (TFxIDF), das als Standardmetrik dafür gilt, wie gut ein Begriff ein bestimmtes Dokument beschreibt.

In je mehr verschiedenen Dokumenten ein Term vorkommt, desto höher ist seine Dokumentfrequenz und desto weniger ist der Term geeignet, ein spezifisches einzelnes Dokument zu beschreiben (in einer Sammlung von Aufsätzen zum Datenschutz kommt der Begriff „Datenschutz" in allen Dokumenten vor und eignet sich daher nicht als differenzierender Begriff). Die TFxIDF-Metrik kann z. B. genutzt werden, um die für ein Dokument inhaltlich wichtigsten Begriffe automatisch zu extrahieren.

Bei Vorliegen von Zeitangaben (Publikationsdatum, Urteilsdatum, Inkrafttreten eines Gesetzes) werden Längsschnittanalysen möglich, durch die Entwicklungen über einen längeren Zeitraum hinweg verfolgt werden können. Es bieten sich auch zusätzliche Visualisierungsformate an; mittlerweile in vielen Kontexten gebräuchlich sind beispielsweise „Wortwolken", bei denen aus den häufigsten Worten eines Textes oder eines ganzen Textkorpus eine textbasierte Darstellung erzeugt wird, die einen schnellen Überblick zu den Inhalten eines Textes erlaubt.

Das nachfolgende Beispiel einer solchen Wortwolke stellt die ca. 100 häufigsten Begriffe einer online verfügbaren Textausgabe des BGB dar, wobei Stoppwörter und Ziffern aus dem Text entfernt wurden. Die Wortwolke wurde mit der online frei verfügbaren Textanalyseumgebung *Voyant Tools*[4] erzeugt.

Die Wortwolke zum BGB

© Dr. Bettina Mielke, Prof. Dr. Christian Wolff

Anwendungsbeispiele der Sprach- und Texttechnologie

Es gibt zahlreiche Anwendungsbeispiele für Sprach- und Texttechnologie, wobei alle auch im Rechtswesen eine Rolle spielen können:

- automatische Übersetzung,
- Informationsextraktion und Textanalyse, z. B. die Erkennung von Eigennamen (etwa zur Anonymisierung von Gerichtsentscheidungen) oder von Produkt-, Orts- oder Firmennamen sowie von Ereignissen sowie Datums- und Zeitangaben; weitergehend auch die Erkennung komplexerer Zusammenhänge in Texten,
- Unterstützung beim Aufbau von Wissensstrukturen (Wissensrepräsentation; vom Text zur rechnergestützten Wissensbasis),
- Information Retrieval, also die Suche in großen Dokumentkollektionen (Suchmaschinen), dazu gehört u. a. die automatische Verschlagwortung und Indexierung,
- Stilometrie, z. B. die automatische Erkennung bzw. Zuweisung der Autorenschaft,
- Sentiment-Analyse, die Analyse von Stimmungen und Emotionen in Texten,
- natürlichsprachliche Frage-Antwortsysteme, die – anders als Information Retrieval-Systeme – zu gestellten Fragen (Suchbegriffen) nicht nur passende relevante Treffer liefern, sondern direkt Antworten erzeugen (mit Hilfe von Verfahren der Textgenerierung),
- natürlichsprachliche Dialogsysteme (sprachgesteuerte Systeme, derzeit u. a. in Form von Chatbots populär),
- Textklassifikation, die automatische Zuordnung eines Textes zu einem Ordnungssystem, zu inhaltlichen Kategorien oder zu einer Gattung, Genre,
- automatische Textzusammenfassung (*abstracting*),
- Textgenerierung, also die Erzeugung von natürlichsprachlichem Text, z. B. aus internen Wissensstrukturen des Rechners.

4) https://voyant-tools.org/.

Teil der Forschung zur Künstlichen Intelligenz (KI)

Als Teilgebiet der Forschung zur Künstlichen Intelligenz haben Computerlinguistik bzw. Sprach- und Texttechnologie auch deren wesentliche Entwicklungsphasen durchlaufen: Nach einer Hochphase wissensbasierter Systeme, in der im Bereich der Texttechnologie beispielsweise explizite Grammatiken als umfangreiche Regelsysteme und digitale Lexika als Wissensspeicher zum Einsatz kamen, sind derzeit vor allem statistische Verfahren und trainierte Modelle auf der Basis künstlicher neuronaler Netze (KNN) im Einsatz. Grundlage derartiger Verfahren ist der Aufbau umfangreicher Textkorpora, die dem Training der Modelle dienen. Das derzeit größte trainierte Sprachmodell, das von der kalifornischen Firma OpenAI 2020 unter dem Namen GTP-3 entwickelt wurde, weist die erstaunliche Zahl von 175 Milliarden Parametern auf.[5] Es wurde u. a. eingesetzt, um (englischen) juristischen Fachtext automatisch in allgemeinverständliche Sprache zu übersetzen.

Projekte im deutschsprachigen Bereich

Derzeit gibt es eine ganze Reihe von unterschiedlichen Arbeitsgruppen im deutschsprachigen Raum, die an der Analyse von Rechtskorpora arbeiten. Nachfolgend sind einige dieser Projekte aufgeführt:

Berlin: Analyse von Entscheidungen des Bundesverfassungsgerichts
Das an der Humboldt-Universität zu Berlin angesiedelte Projekt „Leibniz Linguistic Research into Constitutional Law" (L. L. Con.)[6] hat zum Ziel, ein umfassend annotiertes Korpus mit Entscheidungen des Bundesverfassungsgerichts zu erstellen. Annotationen zu Wortarten und juristischen Kategorien sollen eine maschinelle Auswertung des Entscheidungsaufbaus, der wesentlichen Informationen aus dem Rubrum, der Norm- und Literaturzitate sowie von wiederkehrenden Argumentationsmustern ermöglichen.
Im Rahmen dieses Projekts ging man u. a. der Frage nach, welche Grundrechtsverletzungen in erfolgreichen Verfassungsbeschwerden besonders häufig festgestellt werden, der Verstoß welcher Grundrechte besonders häufig gerügt wird und wie sich diese Werte zueinander verhalten.[7]

© Cifotart – stock.adobe.com

Auch im Rechtswesen sind Datenbanken auf dem Vormarsch mit enorm großen Beständen von Dokumenten aus Gesetzgebung, Rechtsprechung und Literatur.

Düsseldorf/Göteborg: Analyse der Verwaltungsgerichtsentscheidungen 2020/21 zu Corona
Ein deutsch-schwedisches Autorenteam[8] untersucht die deutsche Verwaltungsgerichtsjudikatur hinsichtlich der Entscheidungen zu Corona und arbeitet evidenzbasiert mit Text Mining-Verfahren aus ca. 5.000 Urteilen heraus, ob bei verschiedenen Fallgruppen (u. a. Bildung, Maskenpflicht, Gastronomie, Versammlungen) die Entscheidungen zugunsten der Freiheit oder der Gesundheit getroffen wird.

Erlangen: Korpus mit Entscheidungen des Bundesgerichtshofs
Eine Arbeitsgruppe an der Friedrich-Alexander-Universität Erlangen-Nürnberg um den Juristen *Axel Adrian* und die Computerlinguistin *Stephanie Evert* analysiert ein umfangreiches Korpus mit Entscheidungen des Bundesgerichtshofs.[9]

Heidelberg: Aufbau eines Juristischen Referenzkorpus (JuReKo)
An der Heidelberger Akademie der Wissenschaften wurde ein juristisches Referenzkorpus (JuReKo) des deutschsprachigen Rechts aufgebaut, das Entscheidungstexte, juristische Aufsatzliteratur und Normtexte von 1980 bis 2015 enthält.[10] Es ist vergleichbar mit dem Deutschen Referenzkorpus am Institut für Deutsche Sprache in Mannheim (DeReKo, https://www1.ids-mannheim.de/kl/projekte/korpora.html), das vornehmlich nicht-fachsprachliche Texte (vor allem Pressetexte) in unterschiedlicher Zusammensetzung enthält.
Für ihre jeweiligen Domänen handelt es sich um die weltweit größten Sammlungen linguistisch aufbereiteter Sprachdaten des Deutschen. Untersucht wird anhand der beiden Referenzkorpora der Gebrauch von verschiedenen Begriffen, etwa des Adjektivs *geschäftsmäßig,* dessen Verwendung sich in der juristischen Fachsprache und im allgemeinen Sprachgebrauch deutlich unterscheidet.[11] Es werden dazu die relative Häufigkeit des Ausdrucks in den beiden Korpora über verschiedene Zeiträume herangezogen und die jeweiligen Belegstellen qualitativ analysiert.[12]
Eine andere Studie untersucht arbeitsgerichtliche Entscheidungen mit insgesamt 22,22 Millionen fortlaufenden Wortformen, um das semantische Feld zu *arbeitnehm* einschließlich diachroner Tendenzen in der Entwicklung des Arbeitnehmerbegriffs durch zwei unterschiedliche Teilkorpora (1990–1999 vs.

5) *Dale*, GPT-3: What's it good for? Natural Language Engineering 27 (2021), 113–118.
6) Vgl. https://www.lehrstuhl-moellers.de/llcon.
7) *Wendel*, Welche Grundrechte führen zum Erfolg? Eine quantitative, korpusgestützte Untersuchung anhand von Entscheidungen des Bundesverfassungsgerichts, JZ 2020, 668–679.
8) *Kruse/Langner,* Covid-19 vor Gericht: Eine quantitative Auswertung der verwaltungsgerichtlichen Judikatur, NJW 2021, 3707–3712.
9) Eine Übersicht zur Untersuchungsmethodik und zu ersten Ergebnissen gibt ein von den Autoren veröffentlichter YouTube-Beitrag (https://www.youtube.com/watch?v=a_5U0orSEVs). Vgl. auch den Beitrag von *Adrian* in dieser Ausgabe, S. 9 ff.
10) Vgl. *Vogel/Bäumer/Deus/Rüdiger/Tripps*, Die Bedeutung des Adjektivs geschäftsmäßig im juristischen Fach- und massenmedialen Gemeinsprachgebrauch. Eine rechtslinguistische Korpusstudie als Beispiel für computergestützte Bedeutungsanalyse im Recht, in: LeGes 30 (2019), 3, S. 1–20 (4); https://cal2.eu/core-projects-and-associated-projects/jureko-juristisches-referenzkorpus.
11) *Vogel/Bäumer/Deus/Rüdiger/Tripps* (Fn. 10), S. 18.
12) *Vogel/Bäumer/Deus/Rüdiger/Tripps* (Fn. 10), S. 7 ff.

2000–2012) zu erschließen und den Begriff mit seinem Vorkommen in der Allgemeinsprache zu kontrastieren.[13]

Regensburg: Vergleichende Studien zur österreichischen und deutschen Rechtssprache

Weitere Untersuchungen (Universität Regensburg) befassen sich unter Einsatz korpuslinguistischer Verfahren mit Varietäten der deutschen Rechtssprache durch eine Gegenüberstellung des österreichischen Allgemeinen Bürgerlichen Gesetzbuchs (ABGB) und des deutschen Bürgerlichen Gesetzbuchs (BGB)[14] sowie mit Unterschieden und Gemeinsamkeiten der deutschen und österreichischen Fachsprache Recht im Bereich von Gerichtsentscheidungen.[15]

Ein Ergebnis ist, dass österreichische Gerichtsentscheidungen schwerer verständlich sind als die deutschen Judikate. So weisen die österreichischen Entscheidungen längere Sätze auf und werden nach gängigen Textverständlichkeitsmetriken als schwerer lesbar eingestuft. Dies deckt sich mit Erkenntnissen aus exemplarischen Untersuchungen, wonach die österreichische Rechts- und Verwaltungssprache komplizierter als die deutsche Rechtssprache ist.[16] 2018 und 2019 folgten Untersuchungen zu EuGH-Entscheidungen mittels korpuslinguistischer Methoden.[17]

Schweiz: Analyse zum Wandel des Staatsbegriffs

Für ein Schweizer Textkorpus wurde der Wandel des Staatsbegriffes und -verständnisses unter Verwendung unterschiedlicher Analyseverfahren untersucht.[18]

US-amerikanische Untersuchungen zur Arbeit des Supreme Court

In den Vereinigten Staaten wurden einige Studien vorgelegt, die besonders den Supreme Court, seine Richter und Urteile in den Blick nehmen:

Eine in diesem Sinne traditionelle Korpus-Studie stellt Mouritsen vor, der den Umgang des Supreme Court mit Lexika und lexikalischer Bedeutung untersucht und dabei vor allem auf Fragen der Begriffsinterpretation abstellt.[19]

Den Aspekt der Autorenschaft bei Supreme Court-Urteilen nimmt eine weitere Studie in den Blick, die mit Mitteln der Stilometrie nachzeichnet, wie variabel der Schreibstil unterschiedlicher Richter und ihrer *law clerks* ist.[20]

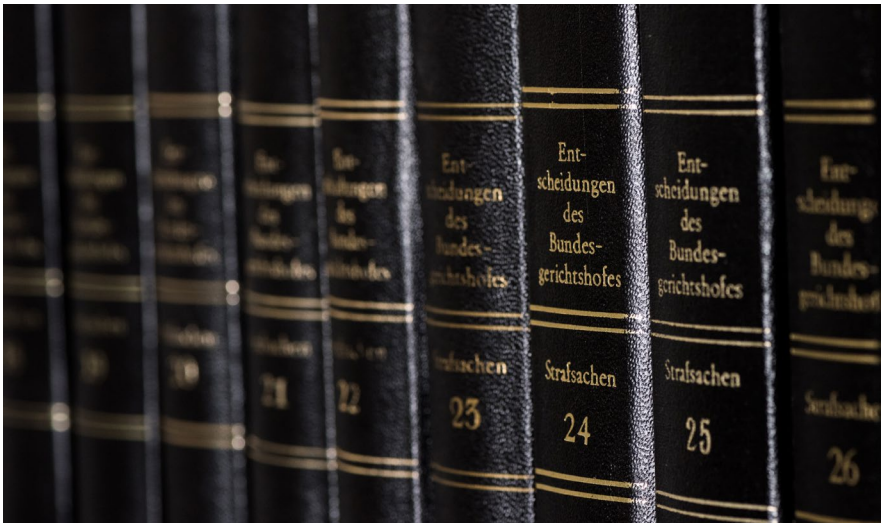
Andere untersuchen ebenfalls stilistische Aspekte des Schreibstils am Supreme Court, wobei es ihnen gelingt, anhand stilistischer Unterschiede in den Texten eine gewandelte Aufgabenverteilung am Supreme Court empirisch nachzuzeichnen.[21]

Eine umfangreiche quantitative Einzelstudie nimmt Sprache und Stil eines einzelnen Supreme Court-Richters, Neil Gorsuch, in den Blick[22]. In einer Längsschnittstudie wurden über mehr als 50 Jahre hinweg die Charakteristika von Supreme Court-Urteilen als spezifisches juristisches Genre untersucht, wobei auch der Aspekt der leichten bzw. schweren Lesbarkeit von Texten eine Rolle spielt.[23]

Ausblick

Trotz der vielfältigen Projekte dürfte die Entwicklung texttechnologischer bzw. korpuslinguistischer Verfahren im Rechtswesen noch am Anfang stehen. Weitergehende Anwendungen wie *predictive*

Die Entscheidungssammlung des BGH: Für die Erstellung eines Korpus müssen die Texte in einem einheitlichen Format vorliegen.



13) *Vogel/Pötters/Christensen*, Richterrecht der Arbeit – empirisch untersucht. Möglichkeiten und Grenzen computergestützter Textanalyse am Beispiel des Arbeitnehmerbegriffs, 2015, S. 93 f., 98 ff., 135.
14) *Mielke/Wolff*, Österreichisch-deutsche Rechtssprache kontrastiv: Eine corpuslinguistische Analyse. In: Schweighofer/Kummer/Hötzendorfer (Hrsg.), Abstraktion und Applikation. Abstraction and Application. Tagungsband des 16. Internationalen Rechtsinformatik Symposions IRIS 2013. Proceedings of the 16th International Legal Informatics Symposium, Österreichische Computer Gesellschaft, Wien 2013, S. 377–384.
15) *Mielke/Wolff*, Österreichische und Deutsche Gerichtsentscheidungen im Sprachvergleich. In: Schweighofer/Kummer/Hötzendorfer/Borges (Hrsg.), Netzwerke. Networks. Tagungsband des 19. Internationalen Rechtsinformatik Symposions IRIS 2016. Proceedings of the 19th International Legal Informatics Symposium, Österreichische Computer Gesellschaft & Erich Schweighofer, Wien 2016, S. 129–138.
16) *Wiesinger*, Das österreichische Deutsch in Gegenwart und Geschichte, 2. durchgesehene und erweiterte Auflage, Wien 2008, S. 109 ff.
17) *Berteloot/Mielke/Wolff*, Deutsches, österreichisches, europäisches Deutsch? Deutschsprachige Fassungen von Urteilen des europäischen Gerichtshofs im Vergleich. In: Schweighofer/Kummer/Saarenpää/Schafer (Hrsg.), Datenschutz/LegalTech. Data Protection/LegalTech. Tagungsband des 21. Internationalen Rechtsinformatik Symposions IRIS 2018. Proceedings of the 21st International Legal Informatics Symposium IRIS 2018, Editions Weblaw, Bern 2018, S. 319–324; *Auer/Berteloot/Mielke/Schikora/Schmidt/Wolff*, Stilometrie in der Rechtslinguistik. Nutzung korpuslinguistischer Verfahren für die Analyse deutschsprachiger Urteile. In: Schweighofer/Kummer/Saarenpää (Hrsg.), Internet of Things. Tagungsband des 22. Internationalen Rechtsinformatik Symposions IRIS 2019. Proceedings of the 22nd International Legal Informatics Symposium IRIS 2019, Editions Weblaw, Bern 2019, S. 375–384.
18) *Abegg/Bubenhofer*, Empirische Linguistik im Recht: Am Beispiel des Wandels des Staatsverständnisses im Sicherheitsrecht, öffentlichen Wirtschaftsrecht und Sozialrecht der Schweiz, Ancilla Iuris 2016, S. 1–41.
19) *Mouritsen*, The Dictionary is not a Fortress: Definitional Fallacies and a Corpus-Based Aproach to Plain Meaning, Brigham Young University Law Review 2010, S. 1915–1980.
20) *Rosenthal/Yoon*, Judicial ghostwriting: authorship on the Supreme Court, Cornell L. Rev. 96 (2010), 1307.
21) *Carlson, Livermore, Rockmore*, A Quantitative Analysis of Writing Style on the US Supreme Court, Wash. UL Rev., 93 (2015), S. 1461.
22) *Varsava*, Elements of Judicial Style: A Quantitative Guide to Neil Gorsuch's Opinion Writing, NYUL Rev. Online, 93 (2018), S. 75.
23) *Livermore/Riddell/Rockmore*, The Supreme Court and the Judicial Genre, Ariz. L. Rev., 59 (2017), S. 837.

analytics-Verfahren im Rechtswesen z. B. zur Prognose von Gerichtsurteilen werden derzeit erprobt; sie setzen ebenfalls aufbereitete Korpora mit juristischen Fachtexten und darauf aufbauende korpuslinguistische Analysen voraus. Die Verfügbarkeit einer größeren Anzahl anonymisierter Urteile könnte dies befördern.

ZU DEN AUTOREN

Dr. Bettina Mielke ist Leiterin der Abteilung Rechtsreferendariat und Staatsexamen am OLG Nürnberg. Sie hat einen Master of Arts (M. A.) in Linguistischer Informationswissenschaft und Germanistik und beschäftigt sich seit langem mit Rechtsinformatik und LegalTech.

Prof. Dr. Christian Wolff ist Inhaber des Lehrstuhls für Medieninformatik an der Fakultät für Informatik und Data Science der Universität Regensburg. Er arbeitet seit langem zu Text Mining und angewandter Sprachtechnologie.

Dr. Bettina Mielke,
Vors. Richterin am Oberlandesgericht Nürnberg
bettina.mielke@olg-n.bayern.de

Prof. Dr. Christian Wolff,
Universität Regensburg
christian.wolff@ur.de