Sprache und Medialität des Rechts
Language and Media of Law

Band 1

Recht ist kein Text

Studien zur Sprachlosigkeit im verfassten Rechtsstaat

Herausgegeben von
Friedemann Vogel

Duncker & Humblot · Berlin

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Satz: Konrad Triltsch GmbH, Ochsenfurt
Druck: CPI buchbücher.de gmbh, Birkach
Printed in Germany

ISSN 2512-9236
ISBN 978-3-428-15247-6 (Print)
ISBN 978-3-428-55247-4 (E-Book)
ISBN 978-3-428-85247-5 (Print & E-Book)

Gedruckt auf alterungsbeständigem (säurefreiem) Papier entsprechend ISO 9706 ♾

Internet: http://www.duncker-humblot.de

Juristische Semantik messend verstehen

CAL²Lab – Eine computergestützte Forschungs- und Experimentierplattform als Beitrag zu einer datengestützten Rechtslinguistik

Von *Isabelle Gauer*, Freiburg, *Friedemann Vogel*, Freiburg,
und *Hanjo Hamann*, Bonn

Abstract

The goal of this interdisciplinary project is to develop and test a computer-linguistic platform for assisting legal linguistic research. The platform aims to support analyses of the use of legal language and legal concepts with the help of data-driven methods. It offers semi-automated tools to investigate the structure of legal concepts on multiple levels, while focussing on the contextual (in)determination of legal expressions ("sedimentation of legal dogmatics") from diachronic (concept change) as well as synchronic (differences subject to legal schools, media, text types, legal fiels, etc.) perspectives. The development and application of the platform opens up new possibilities for legal textual work with mass linguistic data annotated and visualized specially for legal linguistic tasks. Therefore, it contributes to the hermeneutic of a statistical understanding of law. Access to this platform will be free of charge.

Schlüsselwörter: Computerlinguistik, Forschungsplattform, juristische Semantik, datengestützte Analyse

I. Erkenntnisinteresse und Ziele

„Rechtsarbeit ist Textarbeit" (Müller, Christensen & Sokolowski 1997): So lautet die grundlegende Einsicht der modernen Rechtslinguistik seit den 1970er Jahren (Felder & Vogel 2017, im Druck). Im Fokus dieser verhältnismäßig jungen Teildisziplin von Sprach- und Rechtswissenschaft steht die Frage, wie sich der juristische Umgang mit Sprache – als Fachsprache – und die damit verbundenen institutionalisierten Vertextungsprozesse modellieren, kritisieren und letztlich optimieren lassen.

Die Bedeutung von sprachlichen Ausdrücken (zum Beispiel in Gesetzen) ist nicht statisch, denn sie ist fortwährenden Deutungskämpfen und Aushandlungsprozessen ausgesetzt: „Recht ist Streit" (Li 2011). Verlässliche Gesetzesinterpretationen werden durch die Möglichkeit vieler sprachlicher Bedeutungen schwierig, im schlimmsten Fall sogar willkürlich – ein fundamentaler Widerspruch zum Anspruch auf Rechtssicherheit des Bürgers. Tatsächlich sind juristische Deutungen in der Praxis nicht beliebig, sondern durch institutionelle Verfahren konventionalisiert. Grundlage

für diese Sedimentierung juristischer Bedeutungen sind vor allem juristische Ausbildung, Methodenlehre und Dogmatik (Savigny 1814; Engisch 1975; Rückert, Seinecke & Rückert-Seinecke 2012; Müller & Christensen 2013). Die rechtslinguistische Forschung der letzten vier Jahrzehnte hat durch eine Vielzahl empirischer Studien zur theoretischen Aufklärung dieser Zusammenhänge beigetragen.

Mit der Digitalisierung der Gesellschaft entstehen neue analytische Möglichkeiten zur quantifizierenden Analyse juristischer Fachsprache (Vogel 2015). Auf Basis tausender Texte lassen sich wiederkehrende Sprachgebrauchsmuster identifizieren und in die qualitative Auslegungspraxis aufnehmen.

Vor diesem Hintergrund entsteht das hier vorgestellte Projekt: Ziel des Vorhabens ist die interdisziplinäre Entwicklung und exemplarische Erprobung einer computergestützten Forschungs- und Experimentierplattform zur datengestützten Analyse juristischer Sprache (kurz: CAL^{2}Lab). Dazu soll ein Instrument für die rechtslinguistische Forschung sowie für die Rechtspraxis bereitgestellt werden, das neue empirische Einsichten in die begriffliche Systematik der Rechtssprache und ihre Geschichte ermöglicht. Damit verbunden ist das Ziel, an konkreten Beispielen die Möglichkeiten und Grenzen statistischer Algorithmen für die juristische Hermeneutik (Methoden der Auslegung) zu erproben. Konkret geht es um den Aufbau der geplanten Forschungsplattform und zugleich die Beantwortung der folgenden drei Fragen:

1. *Statistische Mehrebenen-Kontext-Analyse juristischer Wörter:* Welche objekt- und metasprachlichen Kategorien sind in statistischer Hinsicht für die rechts- und sprachtheoretische Forschung relevant? Ziel ist die Entwicklung und Bereitstellung eines Programms, das dem Anwender die Auswahl eines Ausdrucks (Wort oder Mehrworteinheit) aus einer Sammlung der 200.000 häufigsten Korpus-Ausdrücke erlaubt und anschließend die dazugehörigen Belege als Konkordanzen (d. h. in zeilenweiser Darstellung des Suchausdrucks und der Umgebungswörter) ausgibt. Weiterhin wird eine Vielzahl grundlegender Angaben zur statistischen Verteilung des Ausdrucks auf verschiedenen Ebenen (Autor, Medium, Zeit, Rechtsbereich, Textsorte usw.; vgl. Teil II) hinzugefügt.
2. *Begriffliche Bestimmtheit vs. Unbestimmtheit einzelner juristischer Ausdrücke:* Wie lässt sich der Grad an begrifflicher Sedimentierung – d. h. der Gebrauchsvielfalt bis hin zur völligen Unbestimmtheit – eines juristischen Ausdrucks theoretisch modellieren (als Skala von ‚absoluter Monosemie'/Eindeutigkeit bis ‚absoluter Polysemie'/Vieldeutigkeit) und computerlinguistisch umsetzen? Arbeitsziel ist die Bereitstellung eines Onlinewerkzeugs, das dem Anwender mit einem Klick nicht nur eine Orientierung darüber erlaubt, in welchen Kontexten bzw. Ebenen ein bestimmter juristischer Ausdruck benutzt wird, sondern auch welchen Grad der Bedeutungsfixierung er dabei im juristischen Diskurs aufweist.
3. *Bedeutungsähnlichkeit (partielle Synonymie) in juristischen Wortfeldern:* Wie lassen sich – aufbauend auf einem empirisch ermittelten Bestimmtheitsmaß (siehe Punkt 1) – juristische Wortgebrauchsähnlichkeiten abbilden, kontextspezifische Wort- und Begriffsfelder identifizieren und effektiv visualisieren? Ziel

ist die Bereitstellung eines neuartigen Werkzeugs („induktiver Thesaurus"), das dem Anwender in Abhängigkeit von bestimmten Metadaten (z. B. zu verschiedenen Zeitpunkten oder Zeitschriften) berechnete Gruppen (Cluster) ähnlich gebrauchter, also quasi-synonymer/sinnverwandter Wörter in Form dynamischer Karten (SOM) ausgibt.

Die Entwicklung und anwendungsfreundliche Bereitstellung dieser drei Werkzeuge verspricht eine grundlegende Erweiterung bisheriger Analysemöglichkeiten nicht nur für die juristische und linguistische Theoriebildung, sondern auch für die Praxis in Rechtsprechung (Auslegung mehrdeutiger gesetzlicher Bestimmungen) und Gesetzgebung (Lesarten zukünftiger Normtexte, sog. „Umgekehrte Subsumtion"): Die Tools erlauben dem Anwender einen neuen Blick in Varianz und Fixierungsgrad von Ausdrücken im juristischen Diskurs. Darunter fallen Untersuchungen zum historischen Wandel einzelner juristischer Begriffe, zu sog. „unbestimmten Rechtsbegriffen" (Cattepoel 1979) und der Nachvollzug der Herausbildung dogmatischer Denkschulen.

II. Datengrundlage

Als zentrale Grundlage der Untersuchung dient das Juristische Referenzkorpus des deutschsprachigen Rechts (JuReko der Version 2016/2). Es wurde im Rahmen eines für zwei Jahre durch die Heidelberger Akademie der Wissenschaften finanzierten Projektes entwickelt, das neben dem Korpusaufbau auch die sprach- und rechtstheoretische Fundierung einer datengestützten, interdisziplinären Hermeneutik zum Ziel hat (Vogel & Hamann 2015; Vogel, Hamann & Gauer eingereicht). Das Korpus besteht aus drei verschiedenen Textsorten: Gerichtsentscheidungen (370.000 Texte, ca. 800 Mio. Wortformen), wissenschaftlichen Aufsätzen aus allen einschlägigen juristischen Fachzeitschriften (43.000 Texte, ca. 200 Mio. Wortformen) und allen deutschen Gesetzestexten (6.300 Texte, ca. 15 Mio. Wortformen) aus einem Zeitraum ca. 1981 bis Ende 2015. In der Version 2.0 ist außerdem ein englisches Subkorpus mit 20.000 Texten (ca. 90 Mio. Wortformen) hinzugekommen. JuReko enthält derzeit insgesamt über eine Milliarde Wortformen. Zu sämtlichen Texten stehen automatisch erhobene und manuell kontrollierte Metadaten zur Verfügung, unter anderem zum Autor, Erscheinungsdatum, Medium usw. (siehe Abbildung). Aktuell liegen alle Texte sowohl in TEI-P5-konformem xml-Format als auch in einer mit der Annotationssoftware TreeTagger (Schmid 1994) angereicherten Fassung vor.

III. Methode

Zunächst wird ein Mehrebenenmodell mit forschungsrelevanten Kategorien entwickelt. Dieses Modell bildet die Grundlage zum Aufbau von Kontextprofilen juristischer Ausdrücke, d. h. es erlaubt Portal-Anwendern eine Übersicht, unter welchen zeitlichen oder sprachlichen Umgebungsbedingungen ein bestimmtes Wort oder ein

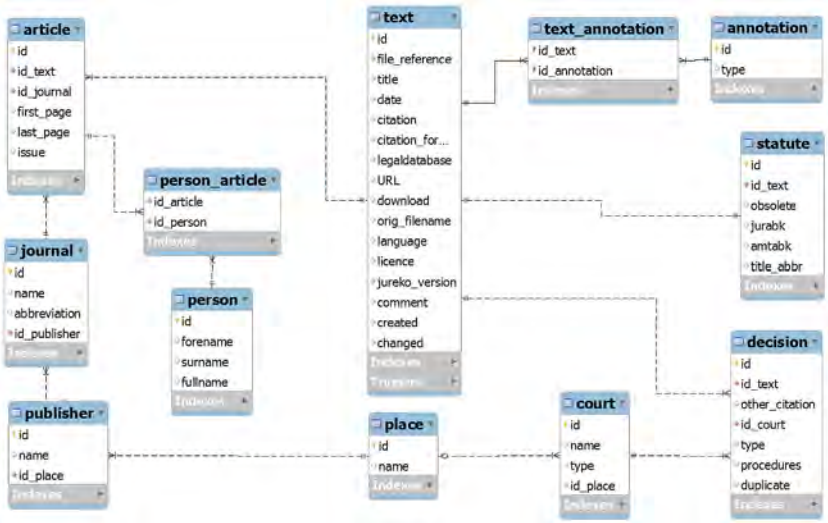

Abbildung: Relationale Datenbank-Struktur des JuReko (2016/2)

Satz typischerweise im Rechtsdiskurs verwendet wird. Folgende vorläufige Kategorien sind zu berücksichtigen:

Oberkategorie	Unterkategorie	Kommentar
N-Gramme	2-Gramme	N-Gramme sind Mehrworteinheiten; Bsp.: Eine Zweiworteinheit und zugleich hinsichtlich ihrer Bedeutung festgefügte juristische Kategorie im Arbeitsrecht zum Ausdruck *Arbeitnehmer* ist *arbeitnehmerähnliche Person.*
	3-Gramme	
	4-Gramme	
	5-Gramme	
Kookkurrenzen Interv.: [-8/+8]	als Lemmata	Kookkurrenzen sind häufig gemeinsam vorkommende Ausdrücke, hier im Abstand von 8 Einheiten links bis 8 Einheiten rechts um den gesuchten Ausdruck. Kookkurrenzen können in ihrer Stammform (Lemma), in der konkreten flektierten Variante (Token) oder als abstrakte Wortart erhoben werden.
	als Token	
	als Wortart (POS)	
Textsorte	Entscheidung	Statistisches Vorkommen des Ausdrucks in Abhängigkeit von unterschiedlichen „Texttypen“; Kategorien werden bei etwaiger Erweiterung der Datenbasis (z. B. auf Kommentartexte) angepasst.
	wiss. Aufsatz	
	Normtext	

Oberkategorie	Unterkategorie	Kommentar
Position im Text	Titelbereich	Die Analyse der Textbereiche, in denen ein gesuchter Ausdruck vorkommt, gibt Aufschluss auf Textstrukturen und sprachliche Handlungsroutinen. Veröffentlichte Gerichtsentscheidungen enthalten etwa folgende Bestandteile: Der Tenor enthält die Urteilsformel; Leitsätze bieten eine kurze Zusammenfassung; Sach- und Streitstand finden sich im Tatbestand; eine Begründung des Urteils wird im Abschnitt Gründe gegeben. Alle nicht untergliederten Texte, z. B. Aufsätze, werden in Quartile und ggf. Fußnoten unterteilt.
	1. Quartil (Text)	
	2. Quartil (Text)	
	3. Quartil (Text)	
	4. Quartil (Text)	
	Fußnoten	
	Tenor	
	Leitsätze	
	Tatbestand	
	Gründe	
Gericht	Instanz	Statistisches Vorkommen des Ausdrucks in Entscheidungen in Abhängigkeit von Instanz, Gerichtstyp (z. B. Oberlandesgericht) und dem Ort. Analysen hierzu geben Aufschluss darüber, welche Begriffe orts-, instanz- oder gerichtstypisch sind.
	Typ	
	Ort	
Medium	Autor	Statistisches Vorkommen des Ausdrucks in juristischen Aufsätzen in Abhängigkeit vom Verfasser und weiteren Daten zur Zeitschrift, in der der Text erschienen ist.
	Zeitschrift	
	Verlag / Name	
	Verlag / Ort	
Zeit (Erscheinen)	insgesamt	Statistisches Vorkommen des Ausdrucks in Abhängigkeit von der Erscheinungszeit der Texte, in denen der Ausdruck gebraucht wird; in Form von Intervallen und einzelnen Jahreszahlen.
	jahresweise	
	5-Jahres-Intervalle	
	10-Jahres-Intervalle	
	vor 1990	
	nach 1990	

Auf Basis einer Auswertung aller in JuReko enthaltenen Texte werden Listen der 200.000 häufigsten Wörter und 2 – 5-Worteinheiten der Rechtssprache generiert. Mithilfe computerlinguistischer Mittel wird zu allen diesen ausgewählten Ausdrücken ein statistisches Kontextprofil nach Maßgabe des Mehrebenenmodells erhoben. Das heißt, es wird zunächst schlicht gezählt, wie oft ein Listenausdruck pro Kategorie belegt ist. Schließlich werden die absoluten Häufigkeitsangaben des Kontextprofils mithilfe statistischer Signifikanztests (LLR, X^2) relativ zum Gesamtkorpus JuReko sowie relativ zum nicht-juristischen-Mediensprachkorpus des Instituts für Deutsche Sprache (IDS, Mannheim) gewichtet. Mit der Signifikanzprüfung lässt sich an-

geben, ob die Häufigkeit eines gewählten Ausdrucks in einer bestimmten Kategorie nicht zufällig, sondern auch im statistischen Sinne „typisch" für die Kategorie ist. Zur anwenderseitigen Ausgabe der fertiggestellten Kontextprofile werden diese schließlich in das Onlineportal integriert und entsprechende Ausgabeformate (tabellarische Ansicht, grafische Ansicht, Konkordanzen) entwickelt. Im Ergebnis lassen sich die jeweiligen Gebrauchsbedingungen eines Ausdrucks sehr genau angeben, also zum Beispiel in welchem Zeitintervall ein Wort in welchen Varianten (Token) wie oft (absolute Frequenz) von welchem Autor in welchem Medium, an welcher Textstelle und mit welchen anderen Ausdrücken gemeinsam (Kookkurrenzen) typischerweise Verwendung findet.

Auf Basis der erstellten Lemma-Kontextprofile ist im nächsten Schritt ein statistisches Maß (Index) zur Bestimmung von relativer Bedeutungsfixierung bzw. Bestimmtheit von Ausdrücken zu entwickeln. Das heißt, es soll ein mathematischer Algorithmus ermittelt werden, der zuverlässig angibt, wie „flexibel" oder auch „unflexibel" ein bestimmter sprachlicher Ausdruck im Recht verwendet werden kann. Bislang existiert ein solches Maß nicht. Es hat den Vorteil, zum Beispiel die Formulierung neuer Gesetzestexte zu erleichtern: es könnte verhindern, dass etablierte (also „unflexible") Ausdrücke missverständlich an anderen Stellen eingesetzt werden. Anschließend wird das Bestimmtheitsmaß für alle Lemmata/Token in Abhängigkeit ausgewählter Kategorien des Mehrebenenmodells berechnet. Die Ergebnisse werden schließlich in die Onlineplattform integriert und mit anwenderfreundlichen Visualisierungen zugänglich gemacht.

Um die relative Bedeutungsähnlichkeit von rechtssprachlichen Ausdrücken zuverlässig anzugeben, wird zunächst ein weiteres Maß entwickelt. Damit ist ein Algorithmus gemeint, der die Kontextprofile in einer mehrdimensionalen Matrix („jeder mit jedem") miteinander vergleicht und den Grad an Gebrauchsähnlichkeit zweier Ausdrücke numerisch wiedergibt. Zur Operationalisierung müssen bestehende Maße (insb. Cosine Similarity, Lee 1999) geprüft und auf die juristischen Daten hin angepasst werden. Anschließend wird die Ähnlichkeit für alle Gebrauchsprofile in Abhängigkeit ausgewählter Kategorien des Mehrebenenmodells berechnet. Die daraus entstehenden mehrdimensionalen Ähnlichkeitsprofile werden anschließend in die Onlineplattform in Form einer SOM-Visualisierung (in Anlehnung an Belica 2008; Keibel & Belica 2007; Vachková & Belica 2009) integriert. Der Portalanwender in gesetzgebenden Verfahren erfährt auf diese Weise, unter welchen Bedingungen ein juristischer Ausdruck in einem Gesetzestext austauschbar (quasi-synonym) ist.

IV. Literatur

Belica, Cyril (2008): Semantische Nähe als Ähnlichkeit von Kookkurrenzprofilen. In: *Korpusinstrumente in Lehre und Forschung / Corpora: strumenti per la didattica e la ricerca / Cor-*

pus Tools in Teaching and Research. Bozen: Alpha beta piccadilly Verlag. URL: http://corpora.ids-mannheim.de/SemProx.pdf [Stand 2016 – 11 – 25].

Cattepoel, Jan (1979): Der unbestimmte Rechtsbegriff als Problem der Rechtssprache. In: *Rechtstheorie. Zeitschrift für Logik und Juristische Methodenlehre, Rechtsinformatik, Kommunikationsforschung, Normen- und Handlungstheorie, Soziologie und Philosophie des Rechts* (10), S. 231 – 246.

Engisch, Karl (1975): Einführung in das juristische Denken. 6. Aufl. Stuttgart: Kohlhammer.

Felder, Ekkehard / Vogel, Friedemann (Hg.) (2017, im Druck): Handbuch Sprache im Recht. Berlin / Boston: De Gruyter. (Handbücher Sprachwissen, 12).

Keibel, Holger / Belica, Cyril (2007): CCDB: A Corpus-Linguistic Research and Development Workbench. In: *Proceedings of the 4th Corpus Linguistics conference*, Birmingham.

Lee, Lillian (1999): Measures of distributional similarity. In: Robert Dale / Ken Church / Lillian Lee (Hg.): *Proceeding. ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, S. 25 – 32.

Li, Jing (2011): „Recht ist Streit". Berlin / New York: De Gruyter.

Müller, Friedrich / Christensen, Ralph (2013): Juristische Methodik: Grundlegung für die Arbeitsmethoden der Rechtspraxis. Aufl. 11. Berlin: Duncker & Humblot.

Müller, Friedrich / Christensen, Ralph / Sokolowski, Michael (1997): Rechtstext und Textarbeit. Berlin: Duncker & Humblot.

Rückert, Joachim / Seinecke, Ralf (Hg.) (2012): Methodik des Zivilrechts – von Savigny bis Teubner. 2. Aufl. Baden-Baden: Nomos.

Savigny, Friedrich Carl von (1814): Vom Beruf unsrer Zeit für Gesetzgebung und Rechtswissenschaft. Heidelberg: Mohr und Zimmer.

Schmid, Helmut (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Vachková, Marie / Belica, Cyril (2009): Self-Organizing Lexical Feature Maps: Semiotic Interpretation and Possible Application in Lexicography. In: IJGLSA 13(2), S. 223 – 260. URL: http://corpora.ids-mannheim.de/IJGLSA.pdf [Stand 2016 – 11 – 25].

Vogel, Friedemann (Hg.) (2015): Zugänge zur Rechtssemantik: Interdisziplinäre Ansätze im Zeitalter der Mediatisierung zwischen Introspektion und Automaten. Berlin / New York: De Gruyter.

Vogel, Friedemann / Hamann, Hanjo (2015): Vom corpus iuris zu den corpora iurum – Konzeption und Erschließung eines juristischen Referenzkorpus (JuReko). In: Heidelberger Akademie der Wissenschaften (Hg.): Jahrbuch der Heidelberger Akademie der Wissenschaften für 2014. Heidelberg: Winter.

Vogel, Friedemann / Hamann, Hanjo / Gauer, Isabelle (eingereicht): Computer Assisted Legal Linguistics: Corpora as a New Instrument for Legal Studies. In: *Law & Social Inquiry.*