Text Mining-Verfahren für die Erschließung juristischer Fachtexte

Bettina Mielke / Christian Wolff+*

**Landgericht Regensburg*
Kumpfmühler Straße 4
D-93066 Regensburg
bettina.mielke@lg-r.bayern.de

+*Universität Regensburg*
Universitätsstr. 31
D-93053 Regensburg
christian.wolff@sprachlit.uni-regensburg.de

Schlagworte: Informationserschließung, Text Mining, juristische Fachtexte, Corpusanalyse, Corpusvergleich, Kollokationen, Visualisierung, Einwortsuche, Terminologieextraktion

Abstract: Der Beitrag stellt Text Mining-Verfahren für die Analyse großer Textcorpora vor und diskutiert, inwiefern sie sich für die juristische Informationserschließung und das juristische Wissensmanagement eignen. Zu den Verfahren gehören die automatische Berechnung aller signifikanten Kollokationen der Begriffe eines Textcorpus, die Visualisierung der wesentlichen Bedeutungszusammenhänge zwischen Begriffen sowie die Extraktion von Fachterminologie durch Vergleich allgemeinsprachlicher und fachlicher Corpora. Am Beispiel einer umfangreichen Urteilssammlung werden exemplarisch Ergebnisse einer Text Mining-Analyse vorgestellt und hinsichtlich ihrer Nutzbarkeit in der juristischen Fachinformation interpretiert. Darauf aufbauend wird untersucht, wie sie sich gezielt für die Präzisierung von Recherchen in juristischen Datenbanken nutzen lassen.

1. Einleitung

Die Menge digital verfügbarer Information hat durch die Verbreitung von Informationssystemen wie dem World Wide Web nicht nur allgemein, sondern insbesondere auch im Bereich der Fachinformation drastisch zugenommen. Dies gilt auch für den juristischen Bereich, wo zwar juristische Informationssysteme seit langem ein bewährtes Arbeitsmittel der Informationserschließung sind[1], die Menge relevanter informationeller Einheiten

[1] Vgl dazu im Überblick für das österreichische Recht: *Liebwald, D.*, Evaluierung juristischer Datenbanken (2003), Verlag Österreich, Wien, insb 74 ff.

einerseits und Erkenntnisse über die Effektivität von Recherchen in solchen Systemen andererseits aber die Frage aufwerfen, inwiefern sich Informationserschließung und Recherche verbessern lassen.[2] Neuere Arbeiten in der Rechtsinformatik greifen daher auch Fragen nach der Entwicklung eines sog *Semantic Web* ua durch Entwicklung juristischer Ontologien auf.[3] In diesem Aufsatz wird untersucht, inwiefern Verfahren des Text Mining eine Hilfestellung für die juristische Informationserschließung im Allgemeinen und die Rechercheunterstützung bei der Suche in juristischen Datenbanken im Besonderen leisten können.[4]

2. Text Mining und Analyse von Textcorpora

Unter Text Mining werden – analog zu den verwandten Begriffsbildungen des Data- oder Web-Mining[5] – Verfahren verstanden, die aus sehr großen und im wesentlichen unstrukturierten Datenmengen relevante Information zu extrahieren in der Lage sind. Für das Text Mining spielen hier insbesondere die Extraktion relevanter Terminologie, die Erkennung bestimmter Begriffstypen wie *named entities* und die Bestimmung relevanter Zusammenhänge zwischen Begriffen eines Textes oder einer Textsammlung eine Rolle. Text Mining-Verfahren arbeiten in der Regel auf der Basis statistischer Algorithmen und können eine wichtige Vorstufe der Wissensmodellierung darstellen, da ihre Ergebnisse zwar eine intellektuelle Wissensaufbereitung nicht ersetzen, sie aber durch Vorgabe von informationellen Strukturen erheblich erleichtern können. Die nachfolgend vorgestellten Daten und Verfahren beziehen sich auf Ergebnisse des Projektes „Deutscher Wortschatz", das seit 1995 an der Abteilung Automatische Sprachverarbeitung (*Prof. Dr. Gerhard Heyer*) des Instituts für Informatik der Universität Leipzig unter Leitung von *PD Dr. Uwe Quasthoff* durchgeführt wird.[6]

---

[2] Vgl *Mielke, B.*, Bewertung juristischer Informationssysteme (2000), Heymanns Verlag, Köln et al, insb 182 f.

[3] Zur Konvergenz von Verfahren des Information Retrieval mit denen der Künstlichen Intelligenz-Forschung vgl *Schweighofer, E.*, The Revolution in Legal Information Retrieval or: The Empire Strikes Back, The Journal of Information, Law and Technology (JILT), 1999 (1), http://elj.warwick.ac.uk/jilt/99-1/schweigh.html (Zugriff Juni 2004).

[4] Alle Beispiele und damit auch die entsprechenden Abkürzungen beziehen sich auf das Recht der Bundesrepublik Deutschland, zB ZPO = Zivilprozessordnung der Bundesrepublik Deutschland.

[5] Vgl *Hearst, M.*, What is Text Mining, Technical Note, University of California at Berkeley, School of Information Management and Systems (SIMS) 2003, http://www.sims.berkeley.edu/~hearst/text-mining.html (Zugriff Juni 2004).

[6] Vgl *Quasthoff, U./Wolff, Ch.*, An Infrastructure for Corpus-Based Monolingual Dictionaries, in: Proc 2nd Int Conference on Language Resources and Evaluation (LREC

2.1. Textcorpora als Grundlage des Text Mining

Voraussetzung des Einsatzes von Text Mining-Verfahren ist der Aufbau einer entsprechenden Textbasis als Analysegrundlage. Im Projekt „Deutscher Wortschatz" ist für das Deutsche ein monolinguales Corpus aus gezielt selektierten allgemeinsprachlichen Quellen entstanden (va Zeitungstexte und ausgewählte Online-Quellen), das laufend ergänzt wird und derzeit einen Umfang von ca 400 Millionen laufenden Wortformen in ca 25 Millionen Sätzen aufweist. In diesem Corpus lassen sich etwa acht Millionen unterschiedliche Vollformen identifizieren.[7]

Neben diesem allgemeinsprachlichen Corpus (Referenzcorpus) existieren zahlreiche Fachcorpora, die fachlich, zeitlich oder institutionell (zB Dokumente einer bestimmten Institution) beschränkt sind. Die auf den Textsammlungen aufsetzenden Text Mining-Verfahren beziehen sich entweder auf das Referenzcorpus und versuchen in diesem Corpus relevante Strukturen zu erkennen, oder sie setzen corpusvergleichende Verfahren ein, um für ein Fachcorpus durch Vergleich mit dem Referenzcorpus Besonderheiten identifizieren zu können. Für die folgenden Betrachtungen ziehen wir ein juristisches Fachcorpus heran, das Urteile der höheren Gerichte der Bundesrepublik Deutschland enthält und bei einem Gesamtumfang von ca 60 Millionen Wortformen aus etwa 3 Millionen Sätzen bei etwa 500.000 unterschiedliche Vollformen besteht.

2.2. Arbeitsschritte bei der Corpusanalyse

Die Analyse der Textcorpora erfolgt als eine Abfolge weitgehend einheitlicher[8] Verfahrensschritte: Ausgehend von der Konvertierung der Ausgangsdokumente in ASCII-Textformat werden Dokumente in Sätze und Sätze in Wörter segmentiert und indexiert, wobei der Erfassung von Häufigkeitsinformation der auftretenden Vollformen besondere Bedeutung zukommt. Die dabei eingesetzte Text Mining-Engine *Concept Composer*[9] berechnet für das gesamte Corpus, dh für alle im Corpus auftretenden Be-

---

2000), Athens, May/June 2000, Vol I, 241-246; *Heyer, G./Läuter, M./Quasthoff, U./Wolff, Ch.*, Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse, in: *Lobin, H. (Hrsg)*, Sprach- und Texttechnologie in digitalen Medien, Proc GLDV-Jahrestagung 2001, Universität Gießen, 71-83.

[7] Das Referenzcorpus ist unter http://wortschatz.uni-leipzig.de, das juristische Fachcorpus unter http://www.texttech.de/urteile online recherchierbar (jeweils Zugriff Juni 2004). Über diese Adressen lassen sich die Daten für die Beispiele unten ermitteln.

[8] Anpassungen sind für Spezifika verschiedener Sprachen und Textsorten erforderlich.

[9] Vgl http://www.texttech.de/produkte.html (Zugriff Juni 2004).

griffe Kollokationsmengen (vgl den folgenden Abschnitt). Externes sprachliches Wissen wie Wortkategorien oder bekannte Mehrwortgruppen kann in die Analyse einbezogen werden.

2.3. Kollokationen und Kollokationsnetze

Unter Kollokationen versteht man das signifikant häufige gemeinsame Auftreten zweier oder mehrerer Begriffe in einem bestimmten Kontext (Textfenster innerhalb eines Dokumentes, Satz, unmittelbare Nachbarschaft). Mit dem oben skizzierten Verfahren werden Satzkollokationen sowie Kollokationen für die unmittelbare Nachbarschaft zweier Begriffe (linke und rechte Nachbarn) ermittelt.[10] Die berechneten Kollokationsmengen für jeden Begriff werden in einer Datenbank abgelegt. Das folgende Beispiel, das der Analysedatenbank des allgemeinsprachlichen Corpus entnommen ist, zeigt die rechten Nachbarschaftskollokationen zu *Hauptstadt*. Die Liste gibt hier nur die stärksten Kollokationen wieder, wobei die Zahlen die jeweilige Stärke des Signifikanzmaßes angeben:

Grosny (2004), Kabul (1099), Kinshasa (682), Kigali (487), Jakarta (417), Bujumbura (372), Mogadischu (370), Manila (334), Duschanbe (332), Sarajewo (327), Algier (319), Skopje (310), Sanaa (308), Colombo (285), Islamabad (275), Dili (273), Nairobi (269), Freetown (266), Monrovia (265), Tirana (262), Harare (254), Seoul (248), Kuala Lumpur (233), Lima (225), Luanda (218), Phnom Penh (194), Addis Abeba (193), Quito (192), Bogota (190), Caracas (180), Kampala (167), Rangun (164), Sarajevo (163), Managua (159), Asmara (157), Minsk (156), Santiago (156), Taipeh (156), Maputo (153), Zagreb (148), Bagdad (145), Port-au-Prince (145), Accra (138), Lusaka (137), Neu-Delhi (137), Machatschkala (134) ...

Bei der Interpretation fällt auf, dass die Kollokationsmenge Namen von Hauptstädten enthält, wobei unter den stärksten Kollokationen vor allem weniger bekannte Städte enthalten sind. Dies läßt sich darauf zurückführen, dass für solche Hauptstädte ihr Status oder ihr Name nicht als bekannt voraus gesetzt werden kann und sie daher häufig mit der Kennzeichnung als Hauptstadt in Texten aufschienen.[11] Eine solche Klasse-Instanz-Beziehung wie für das Konzept *Hauptstadt* und seine Vertreter ist nur ein Beispiel für die vielfältigen Arten semantischer Beziehungen, die sich in Kol-

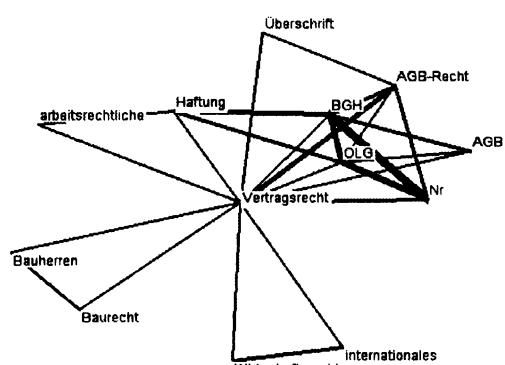
[10] Das Signifikanzmaß entspricht hinsichtlich seiner Ergebnisse in etwa dem *log likelihood*-Verfahren, vgl *Quasthoff, U./Wolff, Ch.*, The Poisson Collocation Measure and its Applications, in: Proc 2nd Int Workshop on Computational Approaches to Collocations, Vienna, July 2002, online: http://www.ai.univie.ac.at/colloc02/PoissonCollocationMeasureQuasthoffWolff_final.pdf (Zugriff Juni 2004).

[11] Vgl den folgenden Beispielsatz zu *Machatschkala*: „Wie der Leiter der Kriminalpolizei in Dagestan, Achmed Suleimanow, am Montag mitteilte, wurde Ismail Dschanchuwatow am Sonntag in der dagestanischen Hauptstadt Machatschkala gefaßt."

lokationsmengen identifizieren, aber bisher nicht ohne weiteres automatisch benennen lassen.

Eine einfache Möglichkeit der Feinanalyse von Kollokationsmengen besteht in der Hinzunahme zusätzlichen Wissens für eine sekundäre Filterung der statistisch errechneten Rohdaten. Für die Filterung kommen unterschiedliche Wissensarten wie Wortkategorien oder der Status von Begriffen zB als Eigennamen in Betracht. Filtert man etwa aus der Kollokationsmenge linker Nachbarn eines Nomens alle Adjektive heraus, so erhält man typische Eigenschaften des Begriffs. Bei Verben als rechte Nachbarschaftskollokationen zu Nomina ergeben sich typische Tätigkeiten.

Eine zusätzliche Nutzungsmöglichkeit für Kollokationen, die unten aus der Perspektive der juristischen Informationserschließung diskutiert wird, ist die Visualisierung von Kollokationsmengen. Dazu werden aus einer Kollokationsmenge diejenigen Kollokationen selektiert, die auch untereinander in einer signifikanten Beziehung stehen, und mit Hilfe eines Layoutalgorithmus in der Ebene positioniert.[12] Das nachfolgende Beispiel zeigt den Kollokationsgraph für den Begriff *Vertragsrecht*, wie er im allgemeinsprachlichen Corpus auftritt. Die Begriffe in jedem Graphen können interaktiv weiterverfolgt werden, dh durch Klick auf ein Konzept im Graph gelangt man zum Kollokationsgraph für den neuen Begriff.



[12] Dabei kommt ein *simulated annealing*-Verfahren zum Einsatz, vgl *Davidson, R./ Harel, D.*, Drawing Graphs Nicely Using Simulated Annealing, ACM Transactions on Graphics, Vol 15, No 4, 1996, 301–331.

3. Juristische Informationserschließung

Im Folgenden wird anhand einiger Beispiele untersucht, ob sich diese Verfahren auch für den juristischen Kontext eignen. Herangezogen wurden dazu drei zivilprozessuale Fragestellungen, die ua den fachlichen Hintergrund für eine Evaluierungsstudie zu juristischen Informationssystemen bildeten.[13]

3.1. Interpretation einzelner Analyseergebnisse

Im einzelnen sollen die Kollokationen zu folgenden zivilprozessualen Themenkomplexen näher analysiert werden: *Wiedereinsetzung in den vorigen Stand*, *Ersatzzustellung* und *Parteiwechsel*.[14]

3.1.1. Wiedereinsetzung

Folgende Satzkollokationen ergeben sich zum Begriff „Wiedereinsetzung", wobei die Werte die Stärke des Signifikanzmaßes angeben (siehe oben):

vorigen (10120), Stand (9082), Versäumung (4186), gewähren (1458), Frist (1229), gewährt (1001), Antrag (977), Berufungsfrist (802), Verschulden (703), beantragt (556), wegen (546), Berufungsbegründungsfrist (544), gegen (453), beantragte (444), 56 (425), Revisionsbegründungsfrist (423), verworfen (399), Berufung (380), einzuhalten (364), unzulässig (343), FGO (334), Fristversäumung (312), 233 (290), versäumte (289), versagt (283), versäumt (278), Einspruchsfrist (267), ZPO (261), Revisionsfrist (254), Schriftsatz (238), verhindert (233), Einlegung (219), Beschwerdefrist (216), 234 (196), Prozeßbevollmächtigten (178), Antragsfrist (176), eingelegt (174), Amts (157), zugleich (143), Oberlandesgericht (140), Klagefrist (140), eingegangenen (136), versäumten (133), Fristversäumnis (127), vorsorglich (113), innerhalb (113), am (109), zweiwöchigen (107), ohne (105), eingegangen (105), verspätet (104), Gewährung (104), 236 (99), Versagung (96), Nachholung (94), Begründung (94), Rechtsmittelfrist (92), Revisionseinlegungsfrist (91), Begründungsfrist (86), rechtzeitig (83), Partei (81), Betracht (81), StPO (79), jemand (76), Wiedereinsetzungsfrist (76) [...]

Die Liste zeigt, dass die Kollokationen im Wesentlichen zu den Begriffen führen, die auch in den einschlägigen Normen der deutschen Zivilprozessordnung (ZPO) vorkommen, obwohl in der Datenbank die Normtexte nicht vorhanden sind. Die meisten sinntragenden Begriffe aus diesen Vorschriften sind zumindest als Wortstamm unter den ersten 20 aufgeführten

[13] Vgl *Mielke*, FN 2; siehe auch *Mielke, B.*, Wie effektiv sind Recherchen in juristischen Informationssystemen? in: *Schweighofer/Menzel/Kreuzbauer (Hrsg)*, Auf dem Weg zur ePerson (2001), 101–112.

[14] Vgl dazu auch die ausgewählten Sachverhalte 2, 4 und 12 in *Mielke*, FN 2, 188 ff.

Satzkollokationen, zudem enthält die Liste Synonyme (wie *Versäumung* oder *versäumt* für *verhindert, Frist einzuhalten*). Es finden sich auch die Zahlen *233* und *234* für die einschlägigen Paragraphenzahlen der ZPO. Die einschlägigen Normen lauten wie folgt:

§ 233 ZPO
War eine Partei ohne ihr Verschulden verhindert, eine Notfrist oder die Frist zur Begründung der Berufung, der Revision, der Nichtzulassungsbeschwerde, der Rechtsbeschwerde oder der Beschwerde nach §§ 621e, 629a Abs. 2 oder die Frist des § 234 Abs. 1 einzuhalten, so ist auf Antrag Wiedereinsetzung in den vorigen Stand zu gewähren.

§ 234 ZPO
(1) Die Wiedereinsetzung muss innerhalb einer zweiwöchigen Frist beantragt werden.
(2) Die Frist beginnt mit dem Tage, an dem das Hindernis behoben ist.
(3) Nach Ablauf eines Jahres, von dem Ende der versäumten Frist an gerechnet, kann die Wiedereinsetzung nicht mehr beantragt werden.

Auch der Leitsatz einer Entscheidung des Bundesgerichtshofs zu diesem Thema, der als einer der Ausgangssachverhalte der Evaluierungsstudie juristischer Informationssysteme diente, lautet: „[...] Dem Prozessgegner ist deshalb auf Antrag Wiedereinsetzung in den vorigen Stand wegen Versäumung der nicht eingehaltenen Frist zur Anfechtung des Urteils (§ 233 ZPO) zu gewähren."[15]

3.1.2. Parteiwechsel

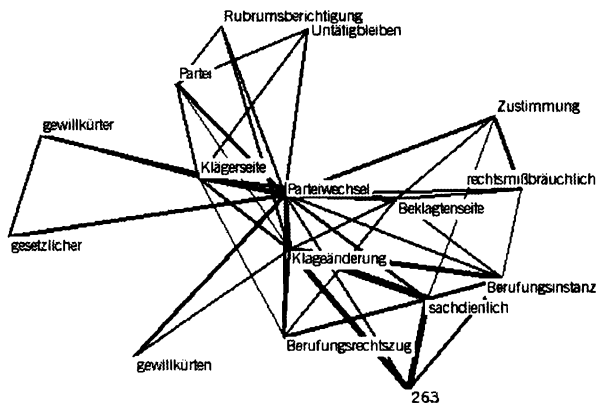
Im Gegensatz zur Wiedereinsetzung ist der gewillkürte Parteiwechsel (im Gegensatz zum gesetzlichen Parteiwechsel) in der deutschen Zivilprozessordnung nicht geregelt. So wird dieses Institut in den verschieden Kommentaren zur Zivilprozessordnung auch an unterschiedlichen Stellen behandelt: Im Kommentar von *Thomas/Putzo* etwa bei der Vorbemerkung zu § 50 ZPO[16], bei *Zöller* bei § 263 ZPO[17] und bei *Stein/Jonas* bei § 264 ZPO[18]. Bei *Thomas/Putzo* ist zum „Meinungsstand zur *gewillkürten* Parteiänderung" (als Oberbegriff zum Parteiwechsel) etwa ausgeführt:

„Das RG sah sie in stRspr [...] als *Klageänderung* an. Daraus folgt, dass sie auch ohne *Zustimmung* des neuen oder alten Beklagten zugelassen wird, wenn sie das Gericht als *sachdienlich (§ 263)* ansieht. [...] Der BGH folgte dem RG [...], gab aber in 21, 285 diesen Standpunkt zunächst auf, unterstellt den *Parteiwechsel* für die *Berufungsinstanz*

[15] Es handelt sich um Sachverhalt 4, vgl *Mielke*, FN 2, 198.
[16] *Putzo*, in: *Thomas/Putzo*, ZPO[25] (2003), Vorbem zu § 50, Rdnr 15.
[17] *Greger*, in: *Zöller*, Zivilprozessordnung[26] (2004), Schmidt, Köln, § 263, Rdnr 3 f, 9, 19ff.
[18] *Schumann, E.*, in: *Stein/Jonas*[21] (1993–1997), Mohr Siebeck, Tübingen, § 264, Rdnr 96 ff.

nicht mehr dem *§ 263,* sondern macht in Analogie zu § 265 Abs 2 S 2 den *Parteiwechsel* von der *Zustimmung* des alten und des neuen Beklagten abhängig, hält sie aber für entbehrlich, wenn sie aus *Rechtsmissbrauch* verweigert wird [...]. Für den ersten Rechtszug hält der BGH die Zustimmung des neuen Beklagten überhaupt für entbehrlich [...]. Diese Rspr des BGH ist inkonsequent und entbehrlich, soweit sie unterschiedslos an der entspr Anwendung des *§ 263* festhält."[19]

Zum zivilprozessualen Konzept *Parteiwechsel* errechnet sich der nachfolgende Kollokationsgraph:



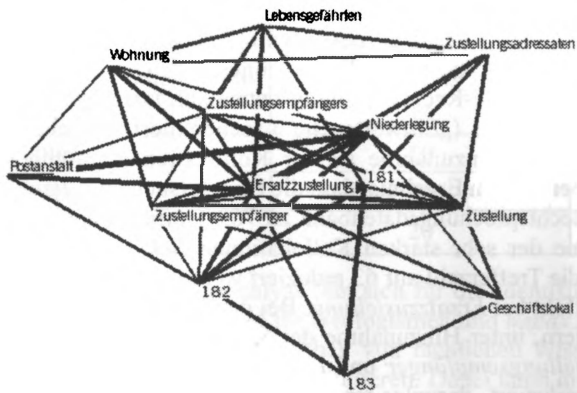
Der Graph zeigt einerseits die grundsätzliche Unterscheidung zwischen *gesetzlichem* und *gewillkürtem* Parteiwechsel auf, andererseits sieht man, dass die Begriffe *Zustimmung* und *rechtsmissbräuchlich* eine Rolle spielen (zur Frage, wann die Zustimmung zum Parteiwechsel erforderlich ist und wann sie wegen Rechtsmissbrauchs entbehrlich ist, insbesondere in der Berufungsinstanz).[20] Außerdem spiegelt er auch eine wichtige Abgrenzungsfrage wider, nämlich zwischen einer *Rubrumsberichtigung* (bei unklarer oder unvollständiger Parteibezeichnung unter Wahrung der Parteiidentität) und einem echten *Parteiwechsel.*

[19] *Putzo*, in: *Thomas/Putzo*, ZPO[25] (2003), Vorbem zu § 50, Rdnr 15; die fett hervorgehobenen Begriffe zeigen Übereinstimmungen mit Konzepten im Kollokationsgraph.

[20] Die in der Evaluierungsstudie zur Bewertung juristischer Informationssysteme als Sachverhalt dienende Entscheidung zu diesem Thema lautet: „Zur missbräuchlichen Verweigerung der Zustimmung zum Parteiwechsel auf der Beklagtenseite im Berufungsrechtszug", vgl Sachverhalt 2 der Evaluierungsstudie, *Mielke*, FN 2, 197.

3.1.3. Ersatzzustellung

Zum Thema *Ersatzzustellung* ergibt sich folgender Kollokationsgraph:



Der Graph spiegelt hier offensichtlich die Rechtslage vor Inkrafttreten des Zustellungsreformgesetzes zum 1. 7. 2002 wider. In der alten Fassung waren die §§ 181 ff ZPO die entscheidenden Normen zu diesem Rechtskomplex. Dabei betraf § 181 ZPO aF die Ersatzzustellung in Wohnung und Haus, wonach die Zustellung in der Wohnung an einen zu der Familie gehörenden erwachsenen Hausgenossen oder an eine in der Familie dienende Person erfolgen kann, wenn die Person, der zugestellt werden soll, in ihrer Wohnung nicht angetroffen wird. Hierzu stellte sich insbesondere die Frage, ob eine Ersatzzustellung an den nichtehelichen Lebensgefährten des Zustellungsempfängers wirksam ist.[21] § 182 ZPO aF befasste sich mit der Zustellung durch Niederlegung, § 183 ZPO aF betraf die Ersatzzustellung im Geschäftslokal. Auch hier zeigt sich wiederum, dass der Graph einen recht guten Überblick zu den wichtigsten Begriffen gibt, die mit dem Thema – zumindest nach der Rechtslage vor dem 1. 7. 2002 – zusammenhängen.[22]

---

[21] Vgl auch Sachverhalt 12 der Evaluierungsstudie, *Mielke*, FN 2, 203.

[22] Durch das Zustellungsreformgesetz ist nunmehr in § 178 ZPO nF geregelt, dass die Ersatzzustellung an einen erwachsenen ständigen Mitbewohner möglich ist, so dass die Frage, ob an einen nichtehelichen Lebensgefährten zugestellt werden kann, keine Rolle mehr spielt.

3.2. Nutzung der Datenanalyse bei der Recherche

Die durch die Kollokationen gewonnenen Begriffe können auch dazu dienen, eine Suchstrategie in Rechtsdatenbanken zu verfeinern. So bietet es sich an, zu einem Ausgangs- oder Zentralbegriff bestimmte Nomina mit einem starken Kollokationsmaß hinzuzunehmen. Um dies zu illustrieren, wurde in der Online-Rechtsprechungsdatenbank *juris*[23] zunächst mit dem Zentralbegriff allein („Einwortsuche") recherchiert und eine Vergleichsrecherche unter Hinzunahme starker Kollokationen durchgeführt. Man kommt dabei zu dem Ergebnis, dass die Einwortsuche mit *Parteiwechsel* in der *juris*-Rechtsprechungsdatenbank zu 769 Treffern[24] führt, während bei Hinzunahme der sehr starken Kollokationen *Beklagtenseite* und *Klageänderung* die Trefferzahl auf 63 reduziert werden kann. Noch stärker zeigt sich der Effekt bei *Ersatzzustellung*. Bei der Einwortsuche kommt man zu 1093 Treffern, unter Hinzunahme der Kollokationen *Niederlegung*, *Wohnung*, *Zustellungsempfänger* und *Lebensgefährte* wird die Recherche auf 7 Treffer reduziert, darunter findet sich auch die Ausgangsentscheidung zum Benutzertest zur Evaluierung juristischer Informationssysteme.[25] Entscheidend ist dabei nicht die offensichtliche Möglichkeit, Ergebnismengen unter Hinzunahme weiterer (beliebiger) Begriffe einschränken zu können, sondern dass dem Nutzer ein stimmiges Begriffsfeld zur Auswahl automatisch vorgelegt werden kann.

3.3. Terminologieextraktion und Ontologieaufbau

Eine weitere Analysemöglichkeit ergibt sich aus dem direkten Vergleich von allgemeinsprachlichem und fachlichem Corpus: Allen Begriffen sind Frequenzklassen zugeordnet, die angeben, um wie viele Zweierpotenzen ein Begriff seltener als der im jeweiligen Corpus häufigste Begriff ist (der Artikel *der*). Damit läßt sich die relative Häufigkeit eines Begriffs im Corpus beschreiben; dieses Maß ermöglicht den direkten Vergleich von Corpora. Es ist anzunehmen, dass fachlich geprägte Begriffe in einem Fachcorpus relativ gesehen häufiger auftreten und damit eine niedrigere Häufigkeitsklasse aufweisen als im allgemeinsprachlichen Corpus. Die

---

[23] Vgl dazu auch *Mielke, B.*, Wie effektiv sind Recherchen in juristischen Datenbanken? in: *Schweighofer/Menzel/Kreuzbauer*, Auf dem Weg zur ePerson (2001), Verlag Österreich, Wien, 101 f sowie ausführlich *Mielke*, FN 2.

[24] Alle Recherchen spiegeln den Stand Juni 2004 wider.

[25] Vgl *Mielke*, FN 2, 203.

nachfolgende Tabelle gibt hierfür einige Beispiele. Daraus ergibt sich, dass zB *Ersatzzustellung* im Fachcorpus um sieben Klassen häufiger ist als im allgemeinsprachlichen Corpus.[26]

Fachbegriff	*Häufigkeits-klasse im Fachcorpus*	*Häufigkeitsklasse im im allgemeinsprachlichen Corpus*	*Differenz*
Parteiwechsel	14	18	4
Ersatzzustellung	13	20	7
Wiedereinsetzung	9	16	7
Klageänderung	11	20	9

Eine solche Frequenzdifferenzanalyse läßt sich für die Identifikation von Fachterminologie (allerdings nicht für Neologismen und *hapax legomena*) nutzen und weitergehend für den Aufbau von fachlichen Wissensnetzen mit der Analyse von Kollokationen kombinieren. Dabei kann die Auswahl relevanter Fachterme und ihrer Kollokationsmengen als Basisinformation für den Aufbau von Ontologien dienen.[27]

4. Fazit

Wir haben versucht, exemplarisch zu zeigen, dass sich Text Mining-Verfahren für das juristische Wissensmanagement und die Recherche in Rechtsdatenbanken einsetzen lassen. Für die Zukunft verbleibt zu untersuchen, wie sie sich auch operativ für Vorhaben der juristischen Informationsaufbereitung (Ontologieaufbau, Erstellung und Erweiterung terminologischer und lexikalischer Datenbestände im Recht) nutzen lassen bzw wie mit ihnen die Recherchemöglichkeiten in juristischen Datenbanken verbessert werden können.

[26] Im fachlichen Corpus bei absoluter Häufigkeit von 212 um den Faktor $2^{13} = 8.192$ seltener als das häufigste Wort, das eine absoluten Häufigkeit von 2.281.996 hat; im allgemeinsprachlichen Corpus bei absoluter Häufigkeit von 8 um den Faktor $2^{20} = 1.048.576$ seltener als das häufigste Wort, das hier eine absolute Häufigkeit von 15.151.724 hat.

[27] Vgl dazu *Böhm, K./Heyer, G./Quasthoff, U./Wolff, Ch.*, Topic Map Generation Using Text Mining, J.UCS – Journal of Universal Computer Science 8(6) (2002), 623–633.