Evidenzbasierte Examensvorbereitung

Zivilrechtliche Lösungsskizzen „auf Distanz“ gelesen

Von Wiss. Mitarbeiter Dr. Dr. **Hanjo Hamann**, J.S.M. (Stanford), Bonn/Berlin*

I. Die schwierige Suche nach dem Examensstoff

Was muss ich für die Klausur wissen, was kann ich auf Lücke lernen? Die Frage ist so alt wie das Klausurwesen, sie treibt jeden Lernenden um und die meisten Lehrenden irgendwann in die Verzweiflung. Es gibt darauf eine einfache und eine richtige Antwort.

Die einfache lautet: Examenskandidaten müssen alles wissen, was das Prüfungsrecht vorschreibt. Sie sollten also § 5 DRiG zusammen mit dem Juristenausbildungsgesetz ihres Landes und der Prüfungsordnung ihrer Universität lesen. Nehmen wir probehalber das nordrhein-westfälische Gesetz über die juristischen Prüfungen (JAG) zur Hand: Nach dessen § 11 Abs. 2 Nr. 1 gehören zum Pflichtfachstoff in der Ersten Juristischen Staatsprüfung die beiden ersten Bücher des BGB, letzteres im „Abschnitt 8 ohne die Titel 2, 11, 15, 18 und 25“. Ausgenommen sind vom Schuldrecht also Teilzeit-Wohnrechte, Auslobung, Einbringung, Leibrente und Vorlegung – ganze 25 der ersten 1059 BGB-Paragraphen. Von den letzten drei Büchern gehören dann zwar „nur“ noch 42 % der Paragraphen zum Pflichtstoff. Das ergibt aber immer noch ein Lernpensum von mindestens 1641 BGB-Paragraphen, die laut Prüfungsrecht im Examen beherrscht werden müssen.[1]

Tatsächlich sind es wohl noch mehr. Denn wie alle Vorschriften werfen auch diejenigen des Prüfungsrechts Auslegungsfragen auf: Wenn § 11 Abs. 2 Nr. 1 JAG NRW im Erbrecht „die Annahme und Ausschlagung der Erbschaft“ zum Pflichtfachstoff erklärt, meint das dann nur § 1943 BGB (Annahme und Ausschlagung der Erbschaft) oder alle 25 zusammengehörigen Paragraphen des Titels über die „Annahme und Ausschlagung der Erbschaft, Fürsorge des Nachlassgerichts“? Wenn im Sachenrecht „aus dem Abschnitt 7 das Recht der Grundschuld“ zum Pflichtfachstoff gehört, ist dann das im selben Abschnitt enthaltene Hypothekenrecht im Umkehrschluss ausgenommen? Oder kommt es durch die Hintertür des § 1192 Abs. 1 BGB – der beide Rechtsinstitute nicht umsonst eng verschränkt – doch wieder in den Pflichtfachstoff mit hinein? Zudem nennt das Prüfungsrecht neben den BGB-Vorschriften auch noch Nebengesetze und Rechtsgebiete, die zumeist „im Überblick“ beherrscht werden müssen (in NRW wären das ProdHaftG, StVG, EGBGB, HGB, GmbHG, Zivilverfahrens- und Arbeitsrecht), wobei wiederum unklar ist, *wieviel* Überblick verlangt wird.

Ignorieren wir diese Auslegungsfragen, dann bleibt von der einfachen Antwort: Ganz oder zum Teil auf Lücke kann gelernt werden, was nicht souverän beherrscht werden muss – also im Zivilrecht alles bis auf 1641 BGB-Paragraphen. Selbst dieses Pensum entspricht jedoch über 40.000 Seiten im Staudinger.[2] *Niemand* beherrscht so viel Stoff souverän.

Die *richtige* Antwort muss deshalb lauten: Examenskandidaten können gar nicht anders, als durchweg jeden Paragraphen teilweise „auf Lücke“ zu lernen. Für den Examensstoff gilt genau das, was *Adam Sagan* kürzlich über die Gesetzgebung sagte: Sie ist kein „System mit Lücken“, sondern ein „System von Lücken“.[3] Kluge Examensvorbereitung nach dem *Dauner-Liebschen* Prinzip „Work smarter not harder“ vermeidet das Lückenlernen also nicht,[4] sondern setzt gezielt und geschickt die richtigen Lücken.

Hierin liegt aber die Krux: Wie lassen sich Lücken geschickt wählen, wie setze ich die richtigen Schwerpunkte für meine beschränkten Ressourcen? Dazu gibt der vorliegende Text einen Impuls. Er schlägt eine Methode vor, um das Lernpensum in der Examensvorbereitung „evidenzbasiert“, also anhand von Daten, zu gewichten – angeregt durch die im juristischen Schrifttum bisher noch fast unbekannten „digitalen Geisteswissenschaften“ (digital humanities). Digitale Methodenkompetenz gehört zunehmend zum Kernbestand juristischer Fähigkeiten[5] – warum nicht in der eigenen Examensvorbereitung damit anfangen?

Der eilige Leser, dessen Examen vielleicht vor der Tür steht, mag direkt zum Hauptteil (Abschnitt III.4) springen, dessen Hauptergebnisse in *Abb. 2* (S. 511), *Abb. 4* (S. 514) und *Abb. 5* (S. 516) visualisiert sind. Wer sich dagegen auch für Methodenfragen interessiert und Anregungen für eine „digitale Rechtsdidaktik“ sucht, wie *Kersten* sie vor einigen Jahren in einer (lesenswerten) theoretischen Vorarbeit vorschlug,[6] wird eher in Abschnitt II. zu Grundlagen des distant reading und in Abschnitt IV. zu den Grenzen der vorgeschlagenen Methodik fündig. Letztere sollten freilich auch eilige Studierende nicht ganz außer Acht lassen.

---

* Der *Verf.* ist Wiss. Mitarbeiter (Senior Research Fellow) am Max-Planck-Institut für Gemeinschaftsgüter in Bonn und externer Habilitand an der Freien Universität (FU) Berlin. Er dankt *Alexander Morell*, *Julian Nyarko*, *Jan-Erik Schirmer*, *Laura Schmitt* und den Teilnehmern des Lehrstuhlkolloquiums von *Andreas Engert* für konstruktive Rückmeldungen, sowie *Britta Padberg*, *Marc Schalenberg* und den Mitarbeitern des Zentrums für interdisziplinäre Forschung (ZiF) für den inspirierenden Aufenthalt in Bielefeld, der diesen Text am weitesten vorangebracht hat.

[1] Die Examensvorbereitung wird selbstredend nicht nach Paragraphen strukturiert, aber letztlich besteht das Examen eben aus Auslegung, Zusammenspiel und Lückenfüllung einzelner Vorschriften, deshalb schadet es nicht, über den Lernaufwand einmal in Paragraphen nachzudenken.

[2] Grob geschätzt: 1641 sind 67 % der 2440 BGB-Paragraphen; der Staudinger umfasst laut Verlag „über 70.000 Seiten in 109 Bänden“ (www.staudinger-bgb.de); 67 % davon wären 47.000 (alle Zahlen Stand 17.2.2020).

[3] *Sagan*, zit. in Hamann, JZ 2020, 84 (85).

[4] So wohl aber *Sanders/Dauner-Lieb*, JuS 2013, 380 (382).

[5] Dazu beispielsweise aus Hamburg und Harvard: *v. Fallois*, Programmieren statt Paragraphen, Bucerius Law School Website (9.3.2020); *Gowder*, Teaching Data Science for Lawyers pp., Library Innovation Lab Blog (9.7.2019).

[6] *Kersten*, JuS 2015, 481.

II. Lesen auf Distanz – „*Distant Reading*“

40.000 Staudinger-Seiten sind gewaltig viel Lesestoff. Zum Glück stehen juristische Studierende mit einer solchen Herkulesaufgabe nicht allein: Der italienische Literaturtheoretiker *Franco Moretti* beispielsweise bemerkte zu Beginn des Jahrtausends, dass inzwischen zwar große Teile der Weltliteratur mit wenigen Mausklicks öffentlich verfügbar sind, dass Literaturwissenschaftler aber trotzdem nur einen „minimalen Bruchteil“ ihres Fachgebiets überblicken. Selbst eine auf den englischen Roman im 19. Jahrhundert spezialisierte Forscherin *müsse* eigentlich einen Kanon von mindestens 200 Werken beherrschen.[7] 200 Romane von je etwa 200 Seiten, das wären: 40.000 Seiten. Selbst das genügte *Moretti* allerdings nicht: Auch der englische Roman im 19. Jahrhundert bestehe schließlich nicht nur aus dem Kanon (der kaum ein Prozent der Literatur ausmache), sondern aus “the novels that were actually published: twenty thousand, thirty, more, no one really knows”.[8] Verglichen damit haben es juristische Examinanden dank des Prüfungsrechts noch leicht.

Wie also gehen andere Text- und Geisteswissenschaften mit der Überforderung durch solche Textmengen um? Darauf schlug *Moretti* damals eine Antwort vor, die seit Jahren die Methodik und Grundlagen vieler Geisteswissenschaften erschüttert: Er stellte der traditionellen Textlektüre, die er als „close reading“ bezeichnete, eine neue Art des Lesens gegenüber, für die er im Jahr 2000 den Begriff „distant reading“ einführte.[9] Damit meinte er den Versuch, eine große Menge von Texten ohne vertiefte Lektüre zu erschließen, zu systematisieren und vorzustrukturieren – also einen

> “process of deliberate reduction and abstraction […] where distance is […] not an obstacle, but a specific *form of knowledge*: fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures.”[10]

Nichtlesen als Erkenntnisgewinn also. Als Möglichkeit, Strukturen zu erkennen, wo vorher nur Worte waren. Als ein in die digitale Welt übertragenes Abrücken von den Bäumen, um wieder den Wald zu sehen. Durch die Extraktion quantitativer Daten aus einem großen Textkorpus, so *Morettis* Versprechen, lasse sich etwas über die Textgesamtheit lernen – gewissermaßen aus der Vogelperspektive –, das uns bei traditioneller, „enger“ Lektüre entgangen wäre.

Es war nicht der erste und jedenfalls nicht der einzige Versuch, Texte mittels digitaler Werkzeuge zu verdichten („verdaten“). *Morettis* Begriffsschöpfung „distant reading“ ist eingängig und einprägsam, fand aber erst vor kurzem Einzug in die US-amerikanische Rechtsliteratur,[11] und die deutsche Rechtswissenschaft.[12] Ähnliche Ansätze konkurrieren (in unterschiedlichen Graden von Automatisierung) unter den Bezeichnungen „Big Data Legal Scholarship“,[13] „Law and Corpus Linguistics“,[14] „Computer Assisted Legal Linguistics“,[15] „Law as Data“[16] und vielen weiteren. Hier soll es uns aber nicht um Begriffsprioritäten und methodische Hegemonien gehen, sondern um die grundlegende Idee, Texte auf Kennzahlen einzudampfen, um den „Sinn für größere Zusammenhänge zu schärfen“, wie *Moretti* formulierte.[17]

Diese Grundidee ist beileibe nicht jeder juristischen Frage angemessen und kann hergebrachte Methoden nicht ersetzen – das gilt in der Rechts- genau wie in der Literaturwissenschaft.[18] Es gibt aber Anwendungsfälle, in denen ein solches Vorgehen sinnvoll sein kann – in der Rechtswissenschaft[19] ebenso wie in der juristischen Ausbildung.

III. Distanzlesen für die Examensvorbereitung

Wie jedes Werkzeug lässt sich auch Distanzlesen ganz unterschiedlich einsetzen. So kann es in der Literaturwissenschaft thematische Gemeinsamkeiten von Texten erkennen helfen, die Lesern vielleicht entgangen wären (etwa durch sog. topic modelling). Die folgende Studie soll Grundprinzipien des computergestützten Distanzlesens auf juristische Texte anwenden und dessen Potential für die Rechtsdidaktik ausloten.

1. Auswahl des Textmaterials

Auf den ersten Blick erscheint es vielleicht reizvoll, das Distanzlesen direkt an den schon erwähnten 40.000 Textseiten des „Staudinger“ oder an Lehrbuchtexten zu erproben. Ungeachtet der Frage, wie sinnvoll das für die Examensvorbereitung wäre, scheitert dies zumeist an den beteiligten Verlagen: Obwohl viele Kommentar- und Lehrwerke digital vorliegen und obwohl § 60d Abs. 1 S. 1 UrhG es gestattet, zu wissenschaftlichen Zwecken „daraus insbesondere durch Normalisierung, Strukturierung und Kategorisierung ein auszuwertendes Korpus zu erstellen“, erschweren Verlage die digitale

[7] *Moretti*, Graphs, Maps, Trees: Abstract Models for Literary History, 2005, S. 3 f.
[8] *Moretti* (Fn. 7), S. 4.
[9] *Moretti* (Fn. 7), S. 1 mit Verw. auf eine Vorarbeit von 2000.
[10] *Moretti* (Fn. 7), S. 1 (*Hervorhebung* im Original).
[11] *Mocsari*, Duke Law J. Online 2018, 41 Fn. 3, (abrufbar unter dlj.law.duke.edu/2018/09/s [28.9.2020]); inzwischen auch *Livermore/Rockmore* (Fn. 16), S. 3–19.
[12] Beiläufig (und skeptisch) etwa *Kersten*, JuS 2015, 481 (486): „Wir müssen insbesondere lernen, mit den digitalen Methoden des ‚non-reading of legal texts‘ [hier Fn. 52 mit Verweis auf distant reading] kritisch umzugehen“.
[13] *Fagan*, Virginia J. of Law & Technology 20 (2016), S. 2.
[14] *Mouritsen*, International J. of Language & Law 6 (2017), S. 67, abrufbar unter doi.org/10.14762/jll.2017.067 (28.9.2020).
[15] *Vogel/Hamann/Gauer*, Law & Social Inquiry 43 (2018), S. 1340, abrufbar unter doi.org/10.1111/lsi.12305 (28.9.2020).
[16] *Livermore/Rockmore* (Hrsg.), Law As Data: Computation, Text, and the Future of Legal Analysis, 2019.
[17] *Moretti* (Fn. 7), S. 1, zitiert oben bei Fn. 10.
[18] Auf einige unvermeidliche methodische Beschränkungen wird unter IV. noch einzugehen sein.
[19] Der interessierte Leser findet vielfältige Anregungen etwa in den Publikationen und laufenden Projekten von *Jens Frankenreiter* (Columbia), *Julian Nyarko* (Stanford) und *Friedemann Vogel* (Siegen).

Textauswertung zunehmend – nicht nur für eigene Texte, sondern sogar für gemeinfreie amtliche Werke (§ 5 UrhG) wie das Bundesgesetzblatt oder die deutsche Rechtsprechung. Darin liegt derzeit die größte Herausforderung für den Einsatz digitaler Werkzeuge im deutschen Recht.[20]

Die evidenzbasierte Examensvorbereitung kann diesem Problem mit einer gewissen Gelassenheit begegnen, denn ihr bestes Rohmaterial ist ohnehin nicht die kommerziell verlegte Rechtsliteratur. Die beste Vorbereitung auf das Examen sind immer noch: Examensklausuren.[21] Diese sind zum Teil frei online abrufbar – so etwa für die Zweite Staatsprüfung dank des Berliner Internet-Klausurenkurses.[22] Für das vorliegende Pilotprojekt ließ sich glücklicherweise ein (noch größerer) Bestand von Originalklausuren für die Erste Staatsprüfung nutzen. Es handelt sich um alle zivilrechtlichen Examensklausuren, die in einem großen Bundesland zwischen 2009 und Mitte 2019 gestellt wurden, samt jeweiliger Lösungsskizzen. Bei je drei Zivilrechtsklausuren pro Examenstermin und zwei Examensterminen pro Jahr waren das insgesamt 63 Klausuren, die digital im pdf-Format vorlagen, mit 1.166 A4-Seiten Lösungsskizzen (also im Schnitt 18 ½ Seiten Lösungsskizze pro Klausur). Studierende könnten dieses konkrete Material wohl allenfalls per IFG-Anfrage erhalten, aber die im Folgenden vorgeführten Methoden sollten sich auch auf andere Textsammlungen übertragen lassen.

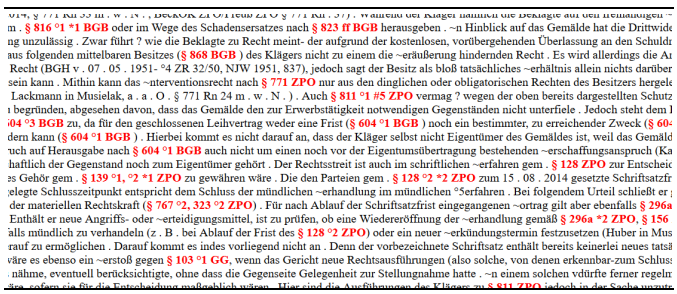
2. Vorbereitung des Textkorpus

Wird in der Computerlinguistik eine größere Anzahl von Texten zur systematischen Auswertung zusammengestellt, spricht man von einem *Korpus* – wie schon einst *Kaiser Justinian*,[23] und nun auch der oben zitierte § 60d UrhG.[24] Computerlinguisten interessieren sich für sprachliche Metastrukturen (Syntaktik, Semantik, etc.), deshalb bereiten sie ihre Korpora im ersten Schritt immer mit Blick auf dieses Erkenntnisinteresse auf. Beispielsweise lassen sie von einem Computerprogramm zunächst alle Wortarten im Text identifizieren (sog. „part-of-speech tagging“, PoS). Für die Examensvorbereitung würde uns die Aufbereitung der in einer Klausurlösung verwendeten Wortarten natürlich wenig helfen. Stattdessen greifen wir auf eine andere Form der Textauszeichnung zurück, die juristischen Gutachten stets innewohnt:

Wo immer eine juristische Frage auftaucht, steht im Gutachten ein Paragraphenzeichen; wo immer juristische Zusammenhänge erörtert werden, häufen sich die Paragraphenzeichen. Zugleich werden Fragen und Zusammenhänge durch die zitierte Paragraphennummer eindeutig inhaltlich bezeichnet. Mit einem Computerprogramm sollten sich also jene Stellen auffinden lassen, die durch ein Paragraphenzeichen markiert sind, und jene Paragraphennummern extrahieren lassen, die den Inhalt des Problems näher benennen.[25]

Dazu wurden die Lösungsskizzen in mehreren Umwandlungsschritten in ein auswertbares Paragraphenkorpus verwandelt: Zunächst wurden die 1.166 Textseiten aus dem pdf-Format in unformatierten Fließtext konvertiert. Daraus wurden im nächsten Schritt alle Paragraphenzitate computergestützt extrahiert.[26] Dabei stellten sich teils unvorhergesehene Schwierigkeiten – etwa dass in einer Lösungsskizze Paragraphen durchweg ohne Gesetzesangabe zitiert wurden oder dass etliche Lösungsskizzen Kommentarstellen oder Lehrbücher zitierten, deren Kapitelzählung ebenfalls durch Paragraphen erfolgt. Der Extraktions-Algorithmus wurde wiederholt überarbeitet, bis in der Kontrollansicht keine systematischen Fehler mehr erkennbar waren:

Abb. 1: Kontrollansicht für die Paragraphen-Extraktion



Gänzlich ausschließen lassen sich Fehler in der Paragraphenextraktion nicht – schon weil zum Teil das Datenmaterial nicht fehlerfrei war.[27] Nach stichprobenartiger Durchsicht dürften allerdings kaum mehr als 2–3 % der Paragraphenzitate dem Extraktionsalgorithmus entgangen sein (falsch-negativ), ebenso wie von den erkannten Paragraphenzitaten kaum mehr als 2–3 % zu Unrecht (falsch-positiv) erfasst worden sein dürften.

3. Beschreibung des Datenmaterials

Nach dem beschriebenen, computergestützten Verfahren, konnten aus dem Textmaterial zuletzt 11.993 Paragraphenzitate extrahiert werden, also im Schnitt 190 Paragraphenzitate

[20] Nicht umsonst stammen drei der vier oben zitierten Methodeninnovationen (Fn. 13–16) aus den USA.

[21] So für seine „Anforderungsanalyse“ bereits *Kuhn*, JuS 2011, 1066 (1071 f.).

[22] Abrufbar unter berlin.de/gerichte/kammergericht/karriere/rechtsreferendariat/vorbereitungsdienst/zusatzangebote#INTNTKLKU (1.10.2020).

[23] Anspielung darauf bei *Vogel/Hamann*, Jahrbuch 2014 der Heidelberger Akademie der Wissenschaften, 2015, S. 275.

[24] § 60d UrhG ist die erste Vorschrift des deutschen Bundesrechts, die mit „Korpus“ kein handgefertigtes Werkstück meint (so etwa § 1 Abs. 2 Nr. 22 HoblMstrV).

[25] Dabei geht es also (wie schon oben in Fn. 1 erläutert) nicht darum, einzelne Paragraphen zu finden und in der Examensvorbereitung gezielt zu pauken, sondern darum, mittels Paragraphen und ihres Zusammenspiels ein Gesamtbild der didaktischen Problemstrukturen zu erhalten, das sich auch bei vertiefter Lektüre von über 1.000 Seiten Lösungsskizzen nicht ohne Weiteres erschließen würde.

[26] Durch sog. reguläre Ausdrücke, ein Standardwerkzeug der meisten Programmiersprachen. (Für das vorliegende Projekt kam Anaconda/Python 3.6.4 zum Einsatz.)

[27] Dazu nur ein Beispiel: In einer Lösungsskizze war § 128 BGB zitiert, wo dem Sinn nach (den ein Computer natürlich nicht versteht) nur § 128 HGB gemeint gewesen sein kann.

pro Lösungsskizze, oder ein Paragraphenzitat auf 230–250 Zeichen Text. (Das entspricht der Länge des vorigen Satzes.) Da viele Paragraphen mehrfach zitiert werden, sagt die Anzahl der Paragraphenzitate natürlich nichts über die Anzahl der zitierten Paragraphen aus. Aggregiert man alle Mehrfachzitationen, so ergibt sich vielmehr, dass über alle Lösungsskizzen hinweg lediglich 1.051 verschiedene Paragraphen zitiert worden waren, davon über die Hälfte (52 %) nur in je einer einzigen Lösungsskizze, und von diesen noch fast drei Viertel (73 %) weniger als drei Mal. Damit verbleiben also nur 655 Paragraphen, die mindestens drei Mal oder in mindestens zwei verschiedenen Lösungsskizzen auftauchten.

Dieser Wert beträgt nur knapp zwei Fünftel der Paragraphenzahl (1.641), die allein im BGB zum prüfungsrechtlichen Pflichtstoff gehören. Das evidenzbasierte Vorgehen ermöglicht also schon eine erhebliche Eingrenzung, die sich durch weitere Auswertungen nun noch schärfer konturieren lässt.

4. Datenauswertung: Drei Beispiele

Für die Datenauswertung wollen wir Werkzeuge, die in den digitalen Geisteswissenschaften etabliert sind, im Kontext unserer Examenslösungen erproben. Dazu nutzen wir beispielhaft drei aufeinander aufbauende Werkzeuge, deren zunehmende Komplexität auch den damit möglichen Erkenntnisgewinn erhöht. Da diese Werkzeuge für andere Einsatzzwecke entwickelt wurden (worauf noch einzugehen ist), kann es hier nur um die Übernahme ihrer jeweiligen Grundprinzipien gehen. In dieser angepassten Version bilden die vorliegenden Auswertungen aber hoffentlich recht anschauliche Illustrationen des Vorgehens zur Frequenzanalyse (a), Kookkurrenzanalyse (b) und Textvektorisierung (c).

a) Frequenzanalyse

Als *Frequenzanalyse* könnte man die Auswertung bezeichnen, wie häufig welche Paragraphenzitate vorkommen.[28] Zunächst wollen wir fragen, welche der oben erwähnten 655 Paragraphen am häufigsten zitiert wurden. Diese Frage lässt sich durch statistische Tabellierung rasch beantworten: Die folgende Tabelle führt die 21 häufigsten Paragraphen (d.h. alle, die insgesamt über achtzig Mal erwähnt wurden) in allen untersuchten Klausurlösungen auf.

Sie belegt zunächst (wenig überraschend), dass die vier Grundnormen des allgemeinen Schuld-, Delikts-, Kauf- und Kondiktionsrechts mit Abstand am häufigsten (je über 150 Mal) zitiert wurden. Auch dass 19 der 21 meistzitierten Vorschriften aus dem BGB stammen, passt zu den prüfungsrechtlichen Vorgaben, die die Beherrschung von Nebengesetzen nur „im Überblick“ erwarten. Umso überraschender dann allerdings die hohen Rangplätze (Zeilen 8 und 14) zweier HGB-Normen. Auch § 179 BGB (Haftung des Vertreters ohne Vertretungsmacht) gehört unter den zwanzig meistzitierten Paragraphen wohl zu den Überraschungskandidaten, zumal seine Häufigkeit dem für seine inflationäre Verwendung berüchtigten § 242 BGB kaum nachsteht. Nur 15 der 21 meistzitierten Paragraphen (nebst neun im weiteren Verfolgerfeld[29]) wurden öfter als in jeder vierten Klausur relevant.

Tab. 1: Häufigste 21 Paragraphen in Examenslösungen

		Zitate		Lösungen		Zitate /
§	**Gesetz**	**abs.**	**rel.**	**abs.**	**rel.**	**Lösung**
280	BGB	478	4,0 %	38	60,3 %	12,6
823	BGB	279	2,3 %	36	57,1 %	7,8
433	BGB	178	1,5 %	31	49,2 %	5,7
812	BGB	163	1,4 %	31	49,2 %	5,3
346	BGB	139	1,2 %	17	27,0 %	8,2
323	BGB	135	1,1 %	21	33,3 %	6,4
241	BGB	129	1,1 %	27	42,9 %	4,8
128	HGB	124	1,0 %	13	20,6 %	9,5
929	BGB	122	1,0 %	22	34,9 %	5,5
275	BGB	122	1,0 %	24	38,1 %	5,1
437	BGB	114	1,0 %	17	27,0 %	6,7
816	BGB	103	0,9 %	10	15,9 %	10,3
281	BGB	101	0,8 %	16	25,4 %	6,3
161	HGB	100	0,8 %	9	14,3 %	11,1
311	BGB	99	0,8 %	25	39,7 %	4,0
985	BGB	97	0,8 %	20	31,7 %	4,8
818	BGB	96	0,8 %	10	15,9 %	9,6
242	BGB	90	0,8 %	32	50,8 %	2,8
249	BGB	82	0,7 %	28	44,4 %	2,9
179	BGB	82	0,7 %	7	11,1 %	11,7
932	BGB	82	0,7 %	15	23,8 %	5,5

Anm.: abs. („absolut“) steht für die gezählte Häufigkeit, rel. („relativ“) für denselben Wert in % der jeweiligen Grundgesamtheit (= 11.993 Paragraphenzitate in 63 Lösungsskizzen).

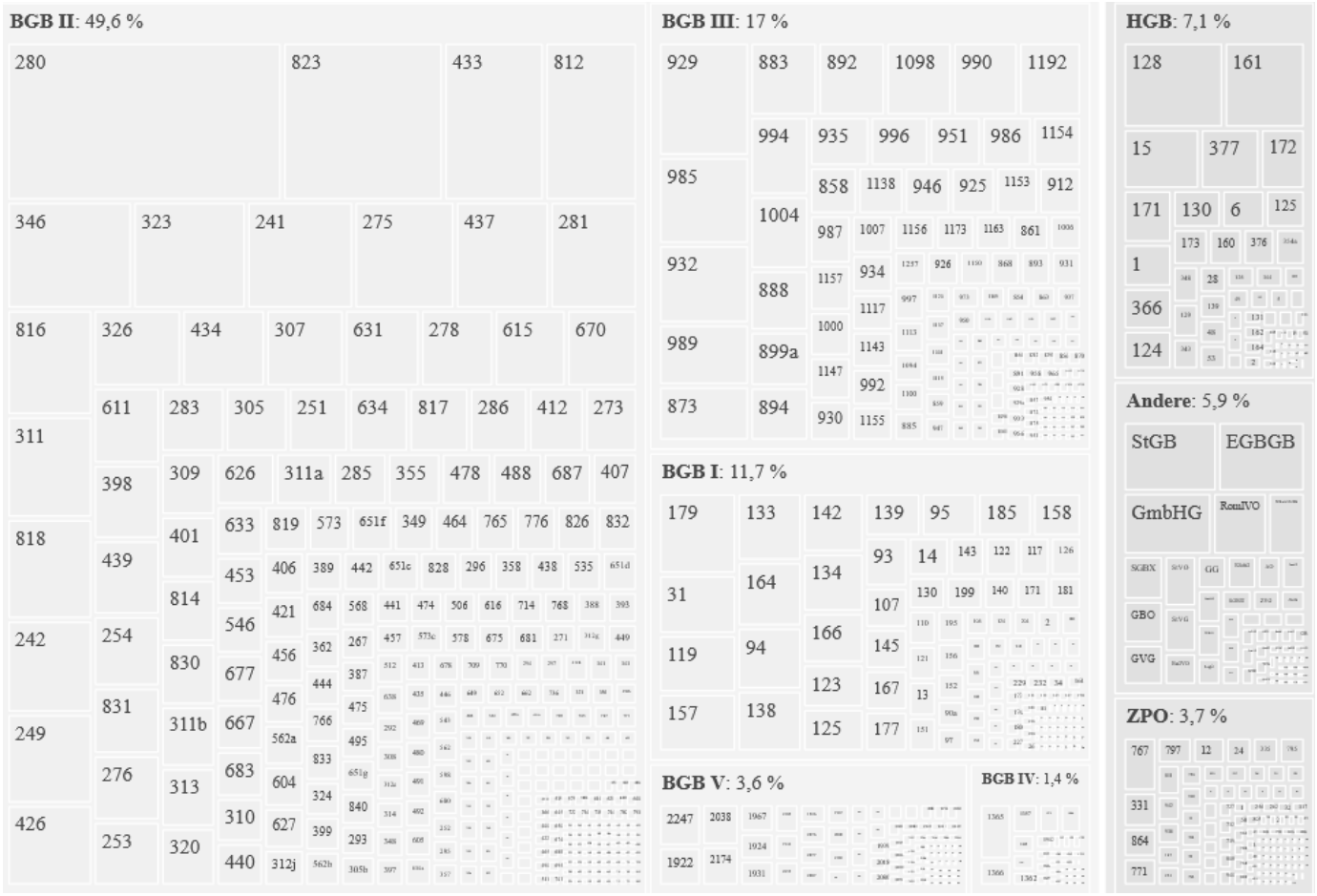
Schon der Beginn dieser Tabelle lässt erahnen, wie solche Daten die Schwerpunktsetzung in der Examensvorbereitung erleichtern könnten: Taucht § 433 BGB doppelt so häufig auf (1,5 %) wie § 932 BGB (0,7 %), ließe sich der Vorbereitungsaufwand entsprechend skalieren. Nun wäre es etwas umständlich, die Liste von Rang 22 bis Rang 655 fortzusetzen und Zeile für Zeile in Zeitaufwand umzurechnen. (Wer sich daran probieren möchte: Den Rest der Tabelle stelle ich online zur Verfügung.[30]) Deshalb empfiehlt sich ein anderes Darstellungsformat aller in den Lösungsskizzen zitierten Paragraphen: Das folgende Kacheldiagramm übersetzt die Häufigkeit jedes Paragraphen in Flächeneinheiten und stellt diese als verschachtelte Rechtecke dar.

[28] Man spricht auch von „beschreibender“ (deskriptiver) Statistik, dazu ausf. *Hamann*, Evidenzbasierte Jurisprudenz, 2014, S. 74 ff.

[29] §§ 278, 157, 133, 164, 254, 398, 276, 166 und 158 BGB.

[30] Abrufbar unter hanjo.1hamann.de/research/zjs2020-tab1.csv (1.10.2020).

Abb. 2: Gewichtsanteile aller zitierten Paragraphen über alle Examenslösungen hinweg



Anm.: Kacheldiagramm (tree map) erstellt mit dem treemap-Befehl des Python-Moduls Plotly Express.

Links oben sehen wir abermals die fünf häufigsten Paragraphen aus der vorigen Liste, die alle zu Buch 2 (Schuldrecht) des BGB gehören. Daneben zeigt das Kacheldiagramm aber auch, dass Buch 2 des BGB generell fast die Hälfte der zitierten Paragraphen stellt – und deshalb in der Examensvorbereitung (wie üblich) den größten Aufwand rechtfertigt. Zudem zeigt das Diagramm, dass nur ein Sechstel (16,7 %) der zitierten Paragraphen aus Nebengesetzen stammen (in grau), und von diesen wiederum die meisten aus ZPO, HGB und StGB. Entlegenere Gesetze wie das StVG und die GBO sind im Diagramm kaum noch sichtbar.

Damit bestätigt unsere Auswertung also den für einzelne Examensklausuren schon zuvor festgestellten Befund: Detailwissen zu exotischen Paragraphen ist „an keiner Stelle erforderlich […] – eine beruhigende Erkenntnis für jeden Examenskandidaten."[31] Wie das obige Kacheldiagramm illustriert, wird im Zivilrechtsexamen vor allem Grundlagen- und Zusammenhangswissen abgeprüft – nicht die von manchen Studierenden gefürchteten Exoten.[32]

b) Kookkurrenzanalyse

Die eben aufgestellte These über das *Zusammenhangswissen* lässt sich leider allein durch Frequenzanalyse gar nicht erhärten. Die absolute Häufigkeit einzelner Paragraphen sagt schließlich nichts über deren Zusammenhänge. Dafür benötigen wir den Zitationskontext, etwa von § 280 BGB. Diese Norm dürfte schon aufgrund ihrer Verweisungsstruktur praktisch nie allein auftauchen. Erst ihr Kontext verleiht dem Normzitat seine spezifische Bedeutung, die für die Examensvorbereitung relevant ist. Ähnliches gilt für die bisher unausgesprochen gebliebene Aggregation von Paragraphenteilen: Die obige Beschreibung lässt noch nicht erkennen, wie oft in § 823 BGB der erste Absatz (absolute Rechtsgüter) und wie oft der zweite Absatz (Schutzgesetze) angesprochen war. Kurzum: Wir benötigen zusätzliches *Kontextwissen*.

Zur Untersuchung sprachlicher Gebrauchskontexte behelfen sich digitale Geisteswissenschaften mit sog. Kookkurrenzen bzw. n-grams. Das sind wiederkehrende Verbindungen aus mehreren („n") sprachlichen Einheiten („tokens"), die zusammen auftreten („ko-okkurieren"), also festgefügte Mehrworteinheiten und formelhafte Sprachmuster bilden. Diese prägen gerade die Rechtssprache auffällig stark, wie Sprachwissenschaftler bei der quantitativen Erforschung der Rechtssprache (Rechtskorpuslinguistik) feststellen:

[31] *Kuhn*, JuS 2011, 1066 (1071 f.).

[32] Medizinstudierende sprechen hier übrigens von „Kolibris". Ich danke *Alexandra Hartmann* für diesen Hinweis.

> „Rekurrente Sprachmuster [können …] *Hinweise auf Sedimente juristischer Dogmatik* sein. Die Rechtssprache ist, so scheint es, voll von derartigen dogmatischen Verhärtungen, die […] die juristische Arbeit orientieren. Man denke etwa an […] feststehende Phrasen wie *menschenwürdiges Existenzminimum* (mit Blick auf Art. 1 Abs. 1 GG), *Recht auf informationelle Selbstbestimmung* oder *Grundrecht auf Gewährleistung der Vertraulichkeit und Integrität informationstechnischer Systeme* (beide mit Blick auf Art. 2 Abs. 1 GG) […]."[33]

Wo die rechtslinguistische Forschung also dogmatische Sedimente identifiziert, indem sie wiederkehrende Wortverbindungen computergestützt aufspürt, lässt sich das zwanglos auf die Kontextualisierung von Paragraphenzitaten übertragen: Was in der natürlichen Sprache als 3-gram erscheint (d.h. als festgefügte Verbindung aus drei Spracheinheiten) mit den Bestandteilen „culpa", „in" und „contrahendo" (oder auch „Verschulden", „bei" und „Vertragsschluss"), ist als Normzitat ebenso ein 3-gram (d.h. eine festgefügte Dreier-Paragraphenkette) aus den Normen § 280 Abs. 1 BGB, § 241 Abs. 2 BGB und § 311 Abs. 2 BGB. Die Forschungslogik der digitalen Geisteswissenschaften passt also zur hiesigen Frage, wenn wir *Paragraphenketten*[34] als gewissermaßen „sedimentierte" Einheiten mehrerer in einem bestimmten Kontext zusammengehöriger Paragraphen verstehen. Wer solches Sediment findet, lernt etwas über den Kontext, ohne den Text genau lesen zu müssen – so die Annahme der linguistischen Kookkurrenzanalyse. Stimmt das auch für Paragraphenzitate?

Schauen wir uns deren Kookkurrenzen (d.h. Paragraphenketten) näher an, so stellen wir zunächst fest, dass Quasi-„Mehrworteinheiten" knapp ein Fünftel aller Paragraphenzitate ausmachen. 79,6 % der Paragraphenzitate in unserem Korpus bestehen aus einem Paragraphen, 14,5 % aus zweien, 4,4 % aus dreien, die verbleibenden 1,5 % aus mindestens vier Paragraphen. Die längste Kette verband sieben Paragraphenteile: §§ 280 Abs. 1, 3, 281 Abs. 1, 433, 434 Abs. 1 S. 2 Nr. 2, 437 Nr. 3, 453 Abs. 1 BGB. Das Zusammenspiel dieser Normen sollte also beherrscht werden, um Ketten dieser Länge verstehen und in der Klausursituation nötigenfalls nachbilden zu können.

Für effizientes Lernen kommt es allerdings wiederum weniger darauf an, die längsten Paragraphenketten zu beherrschen als die üblichsten. Fragen wir deshalb nach den am häufigsten anzutreffenden Paragraphenketten. 18 solcher Ketten waren in je mindestens fünf Examenslösungen zu finden – mit vier Ausnahmen alles Zweierketten, wie die folgende Tabelle zeigt.

Tab. 2: Häufigste 18 Paragraphenketten in Examenslösungen

Paragraphenkette	Lösungen	abs.	rel.
§§ 133, 157 BGB	27	49	2,6 %
§§ 280 I, 241 II, 311 II BGB	14	41	2,2 %
§§ 989, 990 BGB	11	41	2,2 %
§§ 929 S. 1, 932 BGB	10	38	2,0 %
§§ 23 Nr. 1, 71 I GVG	10	10	0,5 %
§§ 280 I, 241 II BGB	9	23	1,2 %
§§ 873 I, 925 I BGB	8	18	0,9 %
§§ 670, 677, 683 S. 1 BGB	8	14	0,7 %
§§ 280 I, III, 283 BGB	7	16	0,8 %
§§ 929 S. 1, 930 BGB	7	12	0,6 %
§§ 280 I, 437 Nr. 3 BGB	7	10	0,5 %
§§ 161 II, 128 HGB	5	29	1,5 %
§§ 280 I, 280 III, 281 BGB	5	18	0,9 %
§§ 437 Nr. 1, 439 I BGB	5	12	0,6 %
§§ 823 II BGB, 263 StGB	5	9	0,5 %
§§ 134, 138 BGB	5	8	0,4 %
§§ 195, 199 BGB	5	7	0,4 %
§§ 12, 13 ZPO	5	6	0,3 %

Anm.: abs. („absolut") steht für die gezählte Häufigkeit, rel. („relativ") für denselben Wert in % der Grundgesamtheit (= 1.905 verwendete Paragraphenketten).

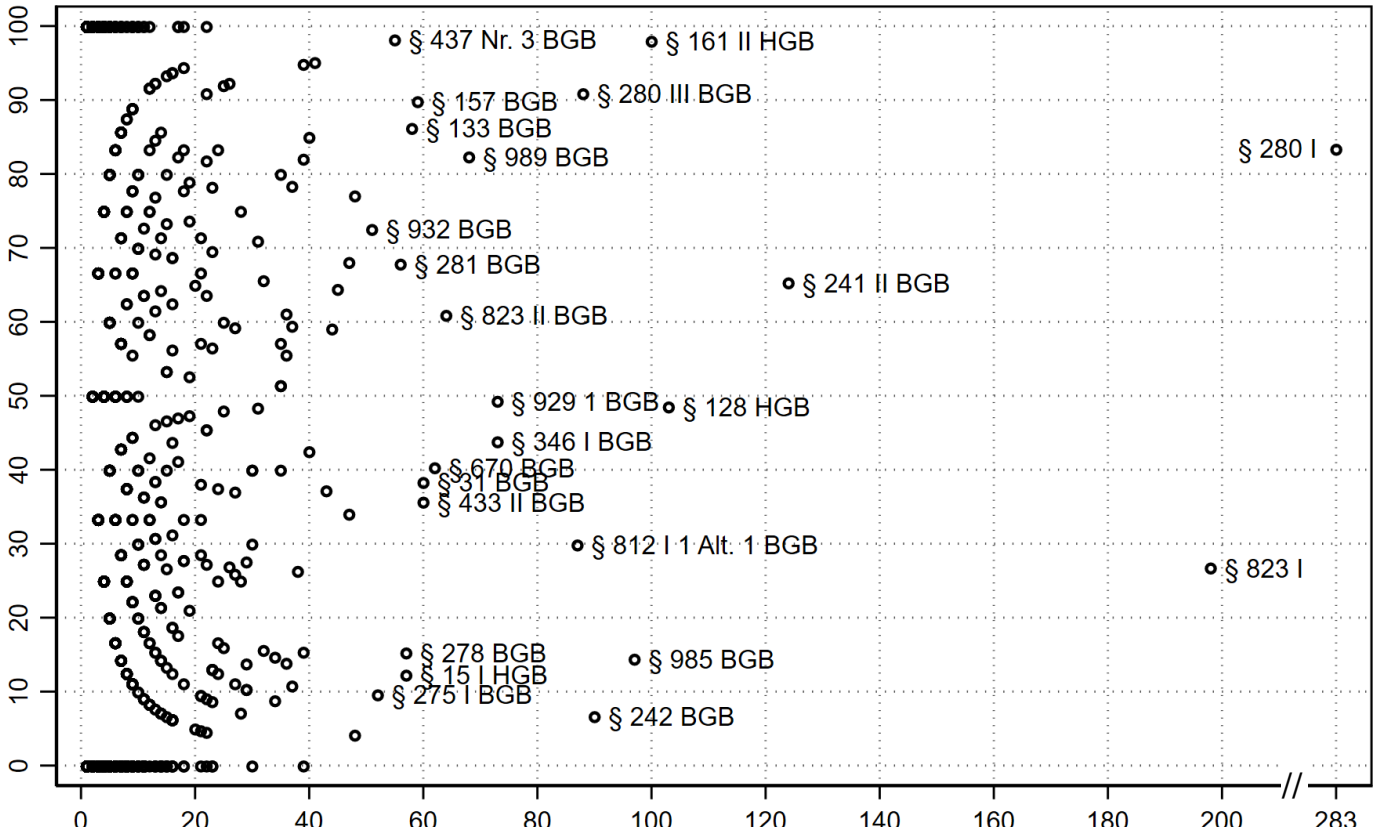
Wenn Paragraphen wie die in der Tabelle genannten sich häufig mit anderen verbinden, wie oft trifft man sie dann überhaupt „allein" und wie oft „in Gesellschaft" an? Dafür können wir in Anlehnung an den obigen (rechtslinguistischen) Sprachgebrauch eine Metrik entwickeln: den *Sedimentierungsgrad* eines Paragraphen. Kommt ein Paragraph überhaupt nur in Paragraphenketten vor, so beträgt sein Sedimentierungsgrad 100 (Prozent); beteiligt er sich nie an Ketten, so liegt der Sedimentierungsgrad bei null. Die Grafik auf der folgenden Seite (Abb. 3) trägt den Sedimentierungsgrad aller 2338 zitierten Paragraphenteile gegen ihre absolute Häufigkeit ab. Sie zeigt beispielsweise, dass § 280 Abs. 1 BGB nicht nur der meistzitierte, sondern auch einer der am stärksten (nämlich in gut vier von fünf Fällen) sedimentierten Paragraphen ist, noch übertroffen etwa von § 437 Nr. 3 BGB, der fast nur in Ketten vorkommt.

Unter der Grafik findet sich sodann eine detaillierte Auflistung (Tab. 3), die im Tabellenkopf die fünf meistzitierten Paragraphenteile aufführt (alle, die mindestens 100 Mal auftraten) und darunter in absteigender Reihenfolge ihre häufigsten Kookkurrenzpartner, d.h. jene Paragraphen, mit denen sie besonders oft gemeinsame Ketten bilden. Dieses Darstellungsformat erlaubt es, für die Examensvorbereitung die kontextuelle Einbettung jedes Paragraphen nachzuvollziehen und eine Norm wie § 823 Abs. 1 BGB anhand ihrer üblichen Gebrauchskontexte (bspw. der Organhaftung nach § 31 BGB oder dem Schadensrecht der §§ 249 ff. BGB) zu verstehen.

[33] *Vogel/Christensen/Pötters*, Richterrecht der Arbeit, 2015, S. 90 (im Abschnitt „Korpuslinguistik – eine kurze Einführung für Rechtswissenschaftler" – *Hervorhebungen* im Original).

[34] Als Paragraphenkette wurde jede nicht durch Worte unterbrochene Aneinanderreihung von Paragraphen behandelt, egal ob verbunden durch Kommata, „und" oder „i.V.m.".

Abb. 3: Häufigkeit und Sedimentierungsgrad aller 1.980 zitierten Paragraphenteile



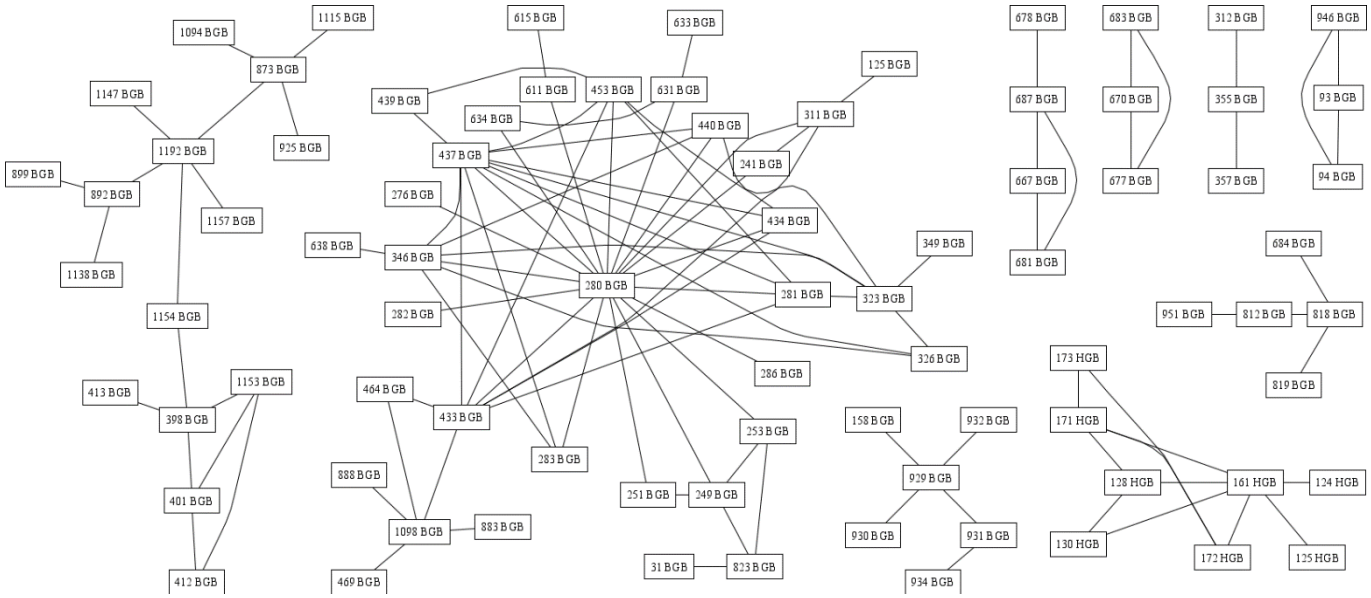
Anm.: Die horizontale Achse trägt absolute Häufigkeit (Anzahl) ab, die vertikale den Sedimentierungsgrad (in %). Jeder Punkt ist ein Paragraphenteil, benannte traten mindestens 50 Mal auf. Erstellt mit dem twoway-Befehl von Stata 16.0.

Tab. 3: Zehn häufigste Kookkurrenzen der fünf häufigsten Paragraphenteile.

	§ 280 I BGB	**§ 241 II BGB**	**§ 823 I BGB**	**§ 161 II BGB**	**§ 128 HGB**
1	§ 241 II BGB	§ 280 I BGB	§ 31 BGB	§ 128 HGB	§ 161 II HGB
2	§§ 280 III, 283 BGB	§§ 280 I, 311 II BGB	§§ 249 I, 251 I BGB	§ 124 HGB	§§ 161 II, 171, 172 HGB
3	§§ 241 II, 311 II BGB	§§ 280 I, 311 II Nr. 1 BGB	§§ 249 II, 253 II BGB	§ 125 I HGB	§§ 130, 161 II, 171, 172, 173 HGB
4	§ 437 Nr. 3 BGB	§ 280 BGB	§ 398 BGB	§§ 128, 171, 172 HGB	§§ 110, 161 II HGB
5	§§ 280 III, 281 BGB	§§ 280 I, 311 III BGB	§ 249 I BGB	§§ 128, 130, 171, 172, 173 HGB	§§ 130, 161 II HGB
6	§§ 241 II, 311 II Nr. 1 BGB	§§ 253 II, 280 I BGB	§ 253 II BGB	§§ 15 I, 160 I, 171 I, 172 I HGB, 433 II BGB	§§ 161 II HGB, 426 II BGB
7	§§ 280 II, 286 BGB	§ 311 II BGB	§§ 1007 I, II, 861 BGB	§§ 110, 128 HGB	§ 488 BGB
8	§§ 241 II, 311 III BGB	§§ 280 I, 311 BGB	§§ 249, 253 II BGB	§§ 124 HGB, 311 III 1 BGB	§§ 130, 161 II HGB
9	§§ 241 II, 253 II BGB	§§ 280 I, 535 I BGB	§§ 280, 611 BGB	§§ 124 HGB, 812 I 1 Alt. 1 BGB	§ 129 HGB
10	§ 433 BGB	§§ 280 I, III, 282 BGB	§§ 989, 990 BGB	§§ 124 HGB, 831 BGB	§ 130 HGB

Anm.: Die erste Spalte benennt den Rangplatz (1 = am häufigsten).

Abb. 4: Netzwerk der Kookkurrenzen von wiederholt in Lösungsskizzen relevanten Paragraphen



Anm.: Netzwerkdarstellung erstellt mit yEd 3.20 (tree balloon layout). Abgebildet sind Teilnetze ab drei Kookurrenzpartnern.

Beide Darstellungen auf der vorigen Seite haben freilich eine Schwäche: Sowohl die Elemente einer Paragraphenkette als auch deren Zitationsreihenfolge sind oft „Geschmacksfragen". Deshalb erprobt Abb. 4 ein weiteres Darstellungsformat, um die Einbettung von Paragraphen in Gebrauchskontexte zu visualisieren: Als Netzwerkdarstellung zeigt sie alle Paragraphen („Knoten"), die mindestens drei Mal oder in mindestens zwei verschiedenen Lösungsskizzen auftraten, und verbindet diejenigen, die dabei Paragraphenketten bildeten. Dargestellt sind alle Teilnetze mit mindestens drei Knoten; sie sind zum Teil unverbunden, weil im konkreten Korpus keine Kookkurrenzen zwischen manchen Paragraphen bestanden. In einem größeren Textkorpus könnten sich durchaus neue Verknüpfungen („Kanten") ergeben, was unter Umständen aber die Übersichtlichkeit beeinträchtigen kann.[35]

In der Netzwerkdarstellung gruppieren sich viele bekannte Kookkurrenzen in sog. Clustern: Rechts unten beispielsweise ein Handelsrechtscluster, links daneben ein Mobiliarsachenrechtscluster, usw. Das größte und am stärksten integrierte Teilnetz hingegen clustert um § 280 Abs. 1 BGB herum, der nicht umsonst als „zentrale Anspruchsgrundlage" des Leistungsstörungsrechts gilt.[36] In der Netzwerkdarstellung bekommt diese müde Metapher eine real sichtbare Gestalt. Examenskandidaten sollten jede der Querverbindungen verstehen, die von § 280 BGB ausgehen und sich in die Verästelungen seines Netzwerks fortsetzen.

c) Textvektorisierung

Dass in der Netzwerkdarstellung Cluster um bestimmte Paragraphen entstehen, darf niemanden überraschen: Diese Paragraphen wurden vom Autor der Klausurlösung bewusst zusammen zitiert und in einer Kette aufgeführt. Die Kookkurrenzanalyse macht also nur sicht- und überschaubar, was ein verständiger Leser des Textes mit etwas Mühe auch von Hand rekonstruieren könnte. Für die dritte Auswertung widmen wir uns deshalb einer Frage, die sich manuell kaum noch mit vertretbarem Aufwand beantworten ließe und deshalb das volle Potential des „Distanzlesens" erkennen lässt: Welche Vorschriften haben *außerhalb* von Paragraphenketten das größte Näheverhältnis zueinander?

Dafür bauen wir auf einer Methodik auf, die in den digitalen Geisteswissenschaften seit gut zehn Jahren besondere Aufmerksamkeit erfährt, aber im Prinzip nur die eben vorgeführte Kookkurrenzanalyse konsequent weiterentwickelt und stärker automatisiert: *Textvektorisierung*. Das bedeutet, nicht nur die gezielte gemeinsame Verwendung von Worten (hier Paragraphen) zu analysieren, sondern gewissermaßen mit dem Blick eines außerirdischen Anthropologen die Beziehung *jedes* Wortes (bzw. jedes Paragraphen) zu *jedem anderen* im selben Text auszumessen. Das Korpus wird dazu in eine Matrix verwandelt, die für jedes Wort sein Vorkommen an jeder Position des Textes als 0 oder 1 kodiert. Dann können Texte oder Textteile als Vektoren („word vector") für mathematische Berechnungen genutzt werden, zum Beispiel in Einbettungsmodellen (sog. word embeddings). Letztere kommen etwa in den heute zunehmend diskutierten und für zahlreiche Anwendungen (nicht zuletzt dem Suchalgorithmus von Google, oder für IBM Watson) genutzten Technologien zur Verarbeitung natürlicher Sprache („natural language processing") zum Einsatz.

[35] Weil in großen Netzwerken oft jeder mit jedem verbunden ist, gerät deren Visualisierung leicht zum „Haarknäuel", sog. „network hairball".

[36] *Plate*, Das gesamte examensrelevante Zivilrecht, 6. Aufl. 2016, S. 639; *Lorenz*, in Beck'scher Online-Kommentar zum BGB, Stand: 1.2.2020, § 280 Rn. 1.

Bilden wir ein einfaches Beispiel, um das Vorgehen Schritt für Schritt zu illustrieren. Nehmen wir zunächst einen Satz, wie er genau so in einer Klausurlösung stehen könnte:

> „Der Eigentumserwerb nach §§ 873, 925 BGB war nicht rechtsgrundlos i.S.v. § 812 Abs. 1 Satz 1 Var. 1 BGB, wenn der Käufer darauf einen Anspruch aus § 433 I 1 BGB hatte."

Herkömmliche Textvektorisierung würde zunächst die nicht sinntragenden Worte wie „der" oder „einen" (sog. stop words) sowie Zahlen und Sonderzeichen tilgen, um den Rechenaufwand auf Inhaltsworte wie „Eigentumserwerb" und „rechtsgrundlos" zu beschränken. Dass dadurch die syntaktische Einbettung dieser Worte verloren geht und aus dem bewusst strukturierten Satz gewissermaßen ein „Sack" von Worten wird (wörtl. „bag of words"), ist dabei gewollt: Gerade dieser Vereinfachungsschritt ermöglicht den „spezifischen Erkenntnisgewinn", den *Moretti* eingangs anpries.

Das weitere Vorgehen ähnelt der Kookkurrenzanalyse: Je häufiger zwei Worte im selben „Sack" (Satz/Absatz/etc.) landen, desto mehr haben sie miteinander zu tun. Aber auch wenn zwei Worte nie im selben Sack landen, dafür aber stets gemeinsam mit denselben anderen Worten, lässt sich eine Relation zwischen ihnen ermitteln (z.B. eine synonymische). Der Fantasie sind keine Grenzen gesetzt: Nehmen wir aus dem Sack für das Wort „Richter" alles heraus, was auch im Sack „Gericht" steckt, und fügen dann alles hinzu, was im Sack „Kanzlei" enthalten ist. Welchem Sack ähnelt das Resultat am meisten? Vielleicht demjenigen für das Wort „Anwalt"? oder „Partner"? oder „Sozius"? Was als mathematisches Hütchenspiel beginnt, endet nicht selten in scheinbar intelligenter Analogiebildung, die von derjenigen eines denkenden Menschen nur schwer zu unterscheiden ist – daher spricht man bei solchen statistischen Auswertungen oft von „künstlicher Intelligenz" („artificial intelligence").

Wir wollen uns hier mit bescheideneren Zielen begnügen und anhand der Paragraphenzitate nur das Grundprinzip demonstrieren. Dafür bedarf es freilich einer entscheidenden Abwandlung: Statt Zahlen und Sonderzeichen zu tilgen, müssen wir diese vielmehr behalten und den Rest entfernen – jene Worte zwischen den Paragraphen, die für unsere Auswertung bloß Füllmaterial darstellen. Auch dieses Füllmaterial birgt jedoch eine wertvolle Information, die wir nicht verlieren sollten: Je mehr Füllmaterial zwischen zwei Paragraphen liegt, desto weniger haben sie wahrscheinlich miteinander zu tun.

Deshalb gehen wir in folgenden Schritten vor: Zunächst vereinheitlichen wir im obigen Beispielsatz die Zitierweisen („Abs. 1" und „I" werden beide zu „°1", „§§" wird zu „§", etc.). Das geht in einem so kurzen Satz wie dem oben zitierten noch von Hand, bei längeren Texten erlauben sog. reguläre Ausdrücke eine Automatisierung dieses Schrittes. Zu Illustrationszwecken zählen wir die Zeichen jeder Zeile auch gleich fortlaufend durch:

```
Der Eigentumserwerb nach § 873, 925 BGB
123456789012345678901234567890123456789
war nicht rechtsgrundlos i.S.v. § 812 °1
123456789012345678901234567890123456789
 *1 |1 BGB, wenn der Käufer darauf einen
123456789012345678901234567890123456789
 Anspruch aus § 433 °1 *1 BGB hatte.
123456789012345678901234567890123456
```

Dieser Text lässt sich nun mithilfe der durchnummerierten Zeichenzählung in Vektoren konvertieren, die jedes enthaltene Paragraphenzitat an jeder Position des Textes mit 0 (für abwesend) oder 1 (für anwesend) kodieren. Da der Text vier Paragraphen und 156 Zeichen umfasst, besteht die Matrix aus vier Vektoren (Spalten) und 156 Zeilen, von denen hier nur zehn wiedergegeben werden sollen:

Position	§ 873	§ 925	§ 812	§ 433
…				
24	0	0	0	0
25	0	0	0	0
26	1	0	0	0
27	0	0	0	0
28	0	0	0	0
29	0	0	0	0
30	0	0	0	0
31	0	0	0	0
32	0	0	0	0
33	0	1	0	0
34	0	0	0	0
…				

Wir sehen sofort das Problem an dieser Darstellung: Fast die komplette Matrix besteht aus Nullen (sog. sparse matrix) und die Datenmenge explodiert auf das Vierfache, weil jeder Paragraph an jeder Textstelle nur einmal auftauchen kann und ansonsten durchweg Nullen aufweist. Je mehr Paragraphen im Text zitiert sind (und je öfter dieselben), desto größer die Aufblähung, deshalb besteht die größte technische Herausforderung für die Textvektorisierung darin, diese Matrix so zu komprimieren, dass die Datenmenge sich noch für Berechnungen eignet und zugleich möglichst wenig Information verloren geht („dimensionality reduction"). Hier liegen die großen technischen Herausforderungen des Verfahrens, die leicht eine informatische Doktorarbeit füllen. Wir wollen uns damit nicht näher befassen, sondern die naheliegende Form der Komprimierung verwenden: Wir extrahieren jedes Paragraphenzitat mit seiner jeweiligen Startposition im Text, die über die Menge des Füllmaterials Auskunft gibt:

§	Normteil	Gesetz	Startposition
873		BGB	26
925		BGB	33
812	°1 *1 |1	BGB	73
433	°1 *1	BGB	135

Mit dieser komprimierten Vektorisierung können wir nun rechnen: Wie weit liegen je zwei Paragraphen im Text auseinander? Vier Paragraphen ergeben fünf paarweise Distanzen:

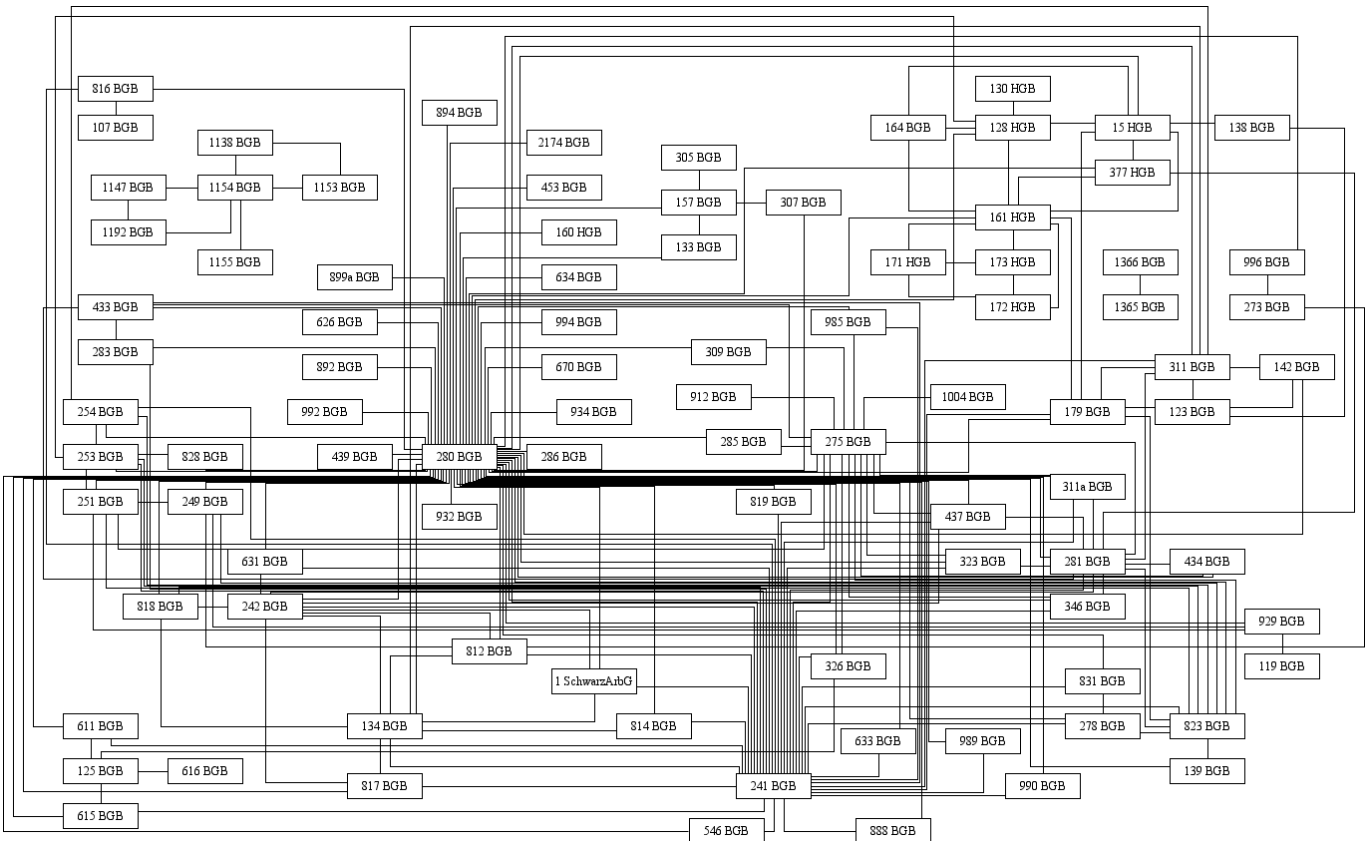
	433	812	873	925
433	0	-62	-109	-102
812	62	0	-47	-40
873	109	47	0	7
925	102	40	-7	0

Die Diagonale können wir ignorieren, da ein Zitat zu sich selbst immer eine Distanz von null aufweist. Auch die rechte obere Hälfte ist ausgegraut, weil sie nur die linke untere Hälfte spiegelt. Diese nicht ausgegrauten Zahlen hingegen sagen uns, wie „nah“ sich zwei Paragraphen stehen, auch wenn sie gar nicht in einer Paragraphenkette (Kookkurenz) auftauchen. Vorzeichen ignorieren wir, weil sie nur von der Sortierung der Matrix abhängen. In unserem Beispiel stehen sich § 873 BGB und § 925 BGB also mit einem Abstand 7 am nächsten. Man könnte ihren Abstand sogar als null definieren – sie also als ein und dasselbe Normzitat behandeln – weil es ohnehin meist vom Zufall abhängt, welche Vorschrift einer Kette vorn und welche hinten zitiert wird. (Deutlicheres Beispiel: Warum sollten § 280 BGB und § 437 BGB im Normzitat „§§ 280 Abs. 1, 433 Abs. 1, 434 Abs. 1, 437 Nr. 3 BGB“ einen größeren Abstand haben als im Normzitat „§§ 433 Abs. 1, 434 Abs. 1, 437 Nr. 3, 280 Abs. 1 BGB“?) Jedenfalls zeigen die paarweisen Distanzen, dass § 812 BGB dem § 433 BGB um 47 Einheiten nähersteht als § 873 BGB. Das war in diesem Fall zwar Zufall, weil der Satz auch andersherum hätte formuliert sein können. Aber genau deshalb wollen wir systematisch alle Paragraphen in allen Lösungsskizzen untersuchen, um strukturelle Näheverhältnisse zu entdecken, die nicht durch Zufälligkeiten einzelner Lösungsskizzen beeinflusst sind – darin liegt just der Sinn des Distanzlesens.

Lassen wir den Computer also das eben für einen einzelnen Satz Vorgeführte für jede Klausurlösung in unserem Korpus wiederholen. Das dauert selbst auf einem fünf Jahre alten Einsteigerlaptop nur etwa fünf Minuten. Um anschließend die systematischen Nähebeziehungen der Paragraphen darzustellen, beschränken wir uns auf solche, die pro Lösungsskizze mindestens fünf Mal und in wenigstens zwei Lösungsskizzen oder drei Mal gemeinsam vorkamen (ohne Paragraphenketten, denen wir uns bereits gewidmet haben). Die folgende Grafik zeigt die 200 Paragraphenpaare (91 Einzelparagraphen) mit der geringsten durchschnittlichen Textdistanz, also gewissermaßen der engsten wechselseitigen „Verschaltung“.

Wer diesen Schaltplan eine Weile studiert, mag aufschlussreiche Verknüpfungen etwa zwischen § 241 BGB und ganz unterschiedlichen Schuldrechtsnormen, sowie rechts oben die wohl wichtigsten HGB-Klausurnormen entdecken – oder links oben den Beleg für die eingangs vermutete Pflichtstoffrelevanz des Hypothekenrechts vermittels § 1192 BGB.

Abb. 5: Schaltplan der 91 am nächsten zueinander genannten Paragraphen in allen Lösungsskizzen



Anm.: Verbindungen stehen für die 200 geringsten mittleren Textdistanzen; erstellt mit yEd 3.20 (orthogonal compact layout).

Obwohl der Schaltplan auf der vorigen Seite (*Abb. 5*) zahlreiche plausible Normbeziehungen veranschaulicht, ist er frei von jeder juristischen Wertung oder dogmatischen Theorie entstanden – allein aus gemessenen Textabständen. Anders gewendet: Auch ein Forscher aus einer fernen Galaxie hätte durch „Distanz"-Lesen diesen Schaltplan erstellen können, obwohl er von unserem Recht überhaupt nichts und von unserer Schriftkultur nur so viel versteht, dass Texte aus Zeichen bestehen und Zeichen nach einem „§"-Symbol irgendwie besonders sind. Dieses Gedankenexperiment verdeutlicht das analytische Potential datengestützter Textauswertung, das momentan unter dem etwas großspurigen Schlagwort „Legal Tech" die Frage aufwirft, welchen Sinn und (möglicherweise ganz neuen?) Zweck die Juristenausbildung noch haben kann. Aber das steht auf einem anderen Blatt.[37]

Für die Zwecke des vorliegenden Pilotversuchs mag es Examenskandidaten helfen, den obigen Schaltplan zu studieren und sich zu fragen, ob sie das Zusammenspiel aller „verschalteten" Paragraphen verstehen und in einer Klausur intelligent darstellen könnten. Mehr sollte es für vier Punkte eigentlich nicht brauchen. Weniger aber auch nicht.

IV. Kritik und Ausblick

Die berichteten Auswertungen sind primär ein Pilotversuch, wie Distanzlesen im Recht funktionieren könnte. Wer darauf seine Examensvorbereitung stützen möchte, muss sich dringend auch deren Beschränkungen bewusst machen:

1. Lösungsskizzen mögen zwar für die „Anforderungsanalyse" hilfreich sein,[38] sind aber nicht autoritativ. Der Klausurersteller schlägt eine Lösung (von vielen denkbaren) vor, und schon die Korrektoren mögen andere Systemverständnisse vertreten. Mehr als dass das in der Lösungsskizze zugrundegelegte Normensystem „vertretbar" ist, darf man also nicht erwarten. Es gibt keine Richtigkeitsgewähr.

2. Lösungsskizzen zielen auf ein bestimmtes Publikum – und das sind nicht Studierende. Sie richten sich an Korrektoren, sind deshalb als „Lösungshinweise" überschrieben und ausweislich der stets vorweggeschickten Anmerkung „nicht als Musterlösung zu verstehen". Diese Vorbemerkung endet traditionell mit den Worten: „Von den Kandidatinnen und Kandidaten kann eine Darstellung in der Tiefe der Lösungshinweise nicht erwartet werden." Deshalb mögen auch einige der darin zitierten Paragraphen und ihre Querverbindungen, auf denen unsere Auswertungen beruhten, nicht zum erwarteten Klausurwissen gehören. Allerdings konnten wir oben ohnehin nur die jeweils prominentesten Paragraphen und Zusammenhänge vorstellen (Paragraphen, die mindestens drei Mal oder in mindestens zwei Klausurlösungen relevant wurden, etc.) und würden uns auf der sicheren Seite irren, wenn wir den Inhalt von Lösungsskizzen als bare Münze für die Examensanforderungen nähmen.

3. Eine folgenreiche Beschränkung ergibt sich daraus, dass Lösungsskizzen kein einheitliches Format haben, sondern von Klausurerstellern mit unterschiedlichen Sorgfaltsmaßstäben und Textverarbeitungsfähigkeiten in Word getippt und dann vom Justizprüfungsamt im pdf-Format ausgegeben werden. Manche enthalten Gliederungen und/oder Zwischenüberschriften, die gewisse Paragraphen als Prüfungspunkte hervorheben und damit überrepräsentieren (weil derselbe Paragraph in Gliederung, Überschrift *und* Text zitiert wird, also drei Mal so häufig wie in anderen Lösungsskizzen, die ihn vielleicht nur im Text erwähnt hätten). Manche Lösungsskizzen enthalten Fußnoten und/oder Literaturnachweise mit Paragraphenzitaten, die sich nicht vollständig automatisiert entfernen lassen. Zum Teil berufen sich Klausurersteller auch auf die Anforderungen der Prüfungsordnung, weshalb die JAPrO als eines der wiederholt zitierten Gesetze erfasst wird, obwohl sie in studentischen Klausuren natürlich nichts zu suchen hat. In Einzelfällen missachten Klausurersteller sogar die anerkannten Zitierregeln und führen Vorschriften beispielsweise konsequent ohne Angabe des Gesetzes an – was die spätere Verarbeitung natürlich enorm erschwert. Solange also Prüfungsämter keine formal und inhaltlich einheitliche Gestaltung von Lösungsskizzen sicherstellen (idealerweise in strukturierten Formaten wie XML), werden Techniken des Distanzlesens zwangsläufig fehleranfällig bleiben – wenngleich zu hoffen ist, dass dieselben Fehler wenigstens nicht systematisch alle Lösungsskizzen durchziehen.

4. Schließlich stellt sich auch die Frage, was frühere Klausurlösungen über künftige Klausuren aussagen. Das vorliegend ausgewertete Material deckt eben nur die vergangenen zehn Jahre ab, während Klausurersteller bekanntlich gern aktuelle „Aufhänger" und neue Rechtsfragen suchen. Da Lösungsskizzen erst einige Zeit nach Abschluss der Examensklausuren freigegeben werden, wird das Material zur Zeit seiner Auswertung immer schon veraltet sein. Zugleich gibt es freilich keinen Automatismus, dass vergangener Prüfungsstoff nicht erneut abgeprüft werden könnte, und es gilt das oben zum Struktur- und Zusammenhangswissen Gesagte: Detailwissen exotischer Normen wird auch in Zukunft nicht erforderlich sein. Das belegt auch der folgende empirische Test: Hätten wir nur zwei Drittel (also die 42 ältesten) unserer Lösungsskizzen ausgewertet, so hätte jede weitere Skizze im Schnitt weniger als neun neue Paragraphen eingeführt und in gut zwei Drittel der Fälle (15 von 21) das Lernpensum um weniger als 1 % erhöht. Anders gewendet waren 98,2 % der in den neuesten drei Lösungsskizzen enthaltenen Paragraphen bereits in den vorangegangenen 60 aufgetaucht. Statt einer Kristallkugel genügt für künftige Klausuren also ein solider Überblick über schon mal Dagewesenes.

Letztlich kann Distanzlesen natürlich weder die gründliche Lektüre juristischer Methodenliteratur noch die Übung an konkreten Fällen ersetzen. Wohl aber vorstrukturieren. Damit sind jene Studierenden im Vorteil, die sich frühzeitig Programmierkenntnisse aneignen und ihren Horizont über die klassische Textexegese der letzten zweitausend Jahre hinaus erweitern. Die so erworbenen Fähigkeiten werden ihnen auch in allen juristischen Berufen bald von Nutzen sein.

[37] Sehr zugänglich etwa der Essay von *Dyevre*, The Future of Legal Theory and the Law School of the Future, 2015.
[38] *Kuhn*, JuS 2011, 1066 (1071 f.), vgl. oben Fn. 21.