KORPUSLINGUISTIK IN DER RECHTSWISSENSCHAFT. EINE WEBBASIERTE ANALYSEPLATTFORM FÜR EUGH-ENTSCHEIDUNGEN

Bettina Mielke[1] / Christian Wolff[2]

[1] Vorsitzende Richterin am Oberlandesgericht Nürnberg, Lehrbeauftragte an der Universität Regensburg, Fürther Str. 110, 90429 Nürnberg, DE, bettina.mielke@olg-n.bayern.de

[2] Professor, Institut für Information und Medien, Sprache und Kultur, Lehrstuhl für Medieninformatik, Universität Regensburg, 93040 Regensburg, DE, christian.wolff@ur.de, http://mi.ur.de

Schlagworte: ***Natural Language Processing, Named Entity Extraction, Sentiment Analysis, Document Similarity, Korpuslinguistik, Rechtskorpora, Legal Corpus Linguistics, Deep Learning, Machine Learning***

Abstract: *Dieser Beitrag befasst sich mit korpuslinguistischen Untersuchungen im Bereich der Rechtswissenschaft und gibt einen Überblick zu aktuellen Studien. Anschließend stellen wir eine Plattform zur Analyse von Entscheidungen des Gerichtshofs der Europäischen Union (EuGH) vor. Ziel ist es dabei, Komponenten der automatischen Sprachverarbeitung (Natural Language Processing, NLP) am Beispiel von EuGH-Entscheidungen in unterschiedlichen Sprachen evaluieren zu können. Dazu wird eine Verarbeitungs-Pipeline umgesetzt, die auf unterschiedliche Subkorpora der Entscheidungen angewandt werden kann und unterschiedliche NLP-Werkzeuge in einer Arbeitsumgebung zusammenzuführt.*

1. Einführung

Derzeit ist (nicht nur) im deutschsprachigen Raum zu beobachten, dass unterschiedliche Arbeitsgruppen an der Analyse von Rechtskorpora arbeiten. Die nachfolgende Liste ist sicher unvollständig, illustriert aber aktuelle Entwicklungen:

– In Erlangen hat eine Arbeitsgruppe um den Juristen Axel Adrian und den Computerlinguisten Stefan Evert ein umfangreiches Korpus mit Entscheidungen des deutschen Bundesgerichtshofs (BGH) analysiert. Eine Übersicht zur Untersuchungsmethodik und zu ersten Ergebnissen gibt ein von den Autoren veröffentlichter YouTube-Beitrag (https://www.youtube.com/watch?v=a_5U0orSEVs).

– Am Lehrstuhl des Berliner Staatsrechtlers Christoph Möllers ist das Projekt «Leibniz Linguistic Research into Constitutional Law» angesiedelt (L.L.Con., vgl. https://www.lehrstuhl-moellers.de/llcon), das zum Ziel hat, ein umfassend annotiertes Korpus mit Entscheidungen des deutschen Bundesverfassungsgerichts (BVerfG) zu erstellen. Annotationen zu Wortarten und juristischen Kategorien sollen eine maschinelle Auswertung des Entscheidungsaufbaus, der wesentlichen Informationen aus dem Rubrum, der Norm- und Literaturzitate sowie von wiederkehrenden Argumentationsmustern ermöglichen, wobei auch Untersuchungsansätze, die keine Annotationen erfordern, zum Einsatz kommen sollen (vgl. https://www.lehrstuhl-moellers.de/llcon).

– Unter Leitung von Hanjo Hamann und Friedemann Vogel wurde an der Heidelberger Akademie der Wissenschaften ein juristisches Referenzkorpus (JuReKo) des deutschsprachigen Rechts aufgebaut, das Entscheidungstexte, juristische Aufsatzliteratur und Normtexte von 1980 bis 2015 enthält (vgl. Vogel et al. 2019, S. 4; https://cal2.eu/core-projects-and-associated-projects/jureko-juristisches-referenzkorpus).

– Unsere Arbeitsgruppe hat sich mit dem Einsatz von Text Mining sowie mit den deutschen rechtssprachlichen Varietäten befasst (Mielke & Wolff 2004, Mielke & Wolff 2013, Mielke & Wolff 2016, Sippl et al. 2016, Berteloot et al. 2018, Auer et al. 2019).

In diesem Beitrag werden wir zunächst aktuelle Ansätze für die Nutzung computer- und korpuslinguistischer Verfahren bei der Analyse von Rechtstexten, insbesondere von Urteilen vorstellen. Im Anschluss diskutieren wir die wichtigsten Designziele und die technische Umsetzung für eine Analyseworkbench und erläutern, welche NLP-Komponenten eingesetzt und welche Analysen und Metriken unterstützt werden.

Die hier vorgestellte Workbench wurde von einer studentischen Arbeitsgruppe (Thomas Fischer, Philipp Hartl, Andreas Hilzenthaler, Lukas Jackermeier) in einem Projektseminar im Masterstudiengang Medieninformatik im Sommersemester 2020 realisiert. Thomas Schmidt, M. Sc., wissenschaftlicher Mitarbeiter am Lehrstuhl für Medieninformatik, hat hinsichtlich der Auswahl der Analysekomponenten und der Methodenwahl beratend mitgewirkt. Ihnen gilt der Dank der Autoren.

2. Korpuslinguistische Arbeiten im Rechtswesen

In den digitalen Geisteswissenschaften (*digital humanities*) sind die Computerlinguistik und die Korpuslinguistik eine wesentliche Triebkraft der Forschung[1], da über lange Zeit hinweg Text das wichtigste digital verfügbare Medium gewesen ist und Text Mining-Analysen bereits seit geraumer Zeit durchgeführt werden konnten. In den letzten Jahren sind neben *inhalts*bezogenen Untersuchungen (Welche Konzepte in einem Dokument sind wichtig? Wie ähnlich sind die Inhalte verschiedener Dokumente? Welches Dokument passt am besten zu einer Suchanfrage?), die auch die klassischen Fragestellungen des Information Retrieval abdecken, zunehmend Studien populär geworden, die versuchen, *stilistische* Aspekte herauszuarbeiten oder *Wertungen* und *emotionale* Aspekte (*sentiment*) in Texten zu erkennen. Hinzu kommen Fragen der Zuordnung von Texten zu *Genres* oder die Untersuchung von Aspekten der Weiterentwicklung einzelner (Text-) Genres. Ein typischer Anwendungsfall für stilometrische Analysen (Eder 2015) ist etwa die Überprüfung oder Feststellung von Autorenschaft (Wurde ein Text tatsächlich vom bekannten Autor geschrieben? Welche Autoren haben welche Teile eines Textes geschrieben?). Derartige Untersuchungen sind mittlerweile auch im juristischen Kontext erfolgt.

2.1. Inhaltliche Fragestellungen

In den Vereinigten Staaten liegen einige Studien vor, die besonders den Supreme Court, seine Richter*innen und Urteile in den Blick nehmen: Eine in diesem Sinne traditionelle Korpus-Studie stellt Mouritsen 2010 vor, der den Umgang des Supreme Court mit Lexika und lexikalischer Bedeutung untersucht. Den Aspekt der Autorenschaft bei Supreme Court-Urteilen betrachten Rosenthal & Yoon 2011, die analysieren, wie variabel der Schreibstil unterschiedlicher Richterinnen und ihrer *law clerks* ist. Carlson et al. 2016 untersuchen ebenfalls stilistische Aspekte des Schreibstils am Supreme Court. Ihnen gelingt, anhand stilistischer Unterschiede in den Texten eine gewandelte Aufgabenverteilung am Supreme Court empirisch nachzuzeichnen. Livermore et al. 2017 stellen eine Längsschnittstudie vor, in der sie über mehr als 50 Jahre die Charakteristika von Supreme Court-Urteilen als spezifisches juristisches Genre untersuchen und dabei die in den vergangenen Jahren vielbeachtete Methode des Topic Modeling von Texten einsetzen. Sie gehen dabei auch auf den Aspekt der Lesbarkeit von Texten ein. Eine umfangreiche quantitative Einzelstudie stammt von Varsava 2018, die Sprache und Stil eines einzelnen Supreme Court-Richters, Neil Gorsuch, in den Blick nimmt.

Für den deutschsprachigen Bereich legen Abegg & Bubenhofer 2016 eine Korpusanalyse vor, in der sie den Wandel des Staatsbegriffes und -verständnisses für ein Schweizer Textkorpus untersuchen und dabei

[1] Zum wechselseitigen Verhältnis von Computer- und Korpuslinguistik zu den digitalen Geisteswissenschaften vgl. Piotrowski 2013.

unterschiedliche Analyseverfahren vorstellen. Im Rahmen des Projekts «Leibniz Linguistic Research into Constitutional Law» geht Wendel 2020 mit Hilfe von korpuslinguistischen Methoden der Frage nach, welche Grundrechtsverletzungen in erfolgreichen Verfassungsbeschwerden besonders häufig festgestellt werden, der Verstoß welcher Grundrechte besonders häufig gerügt wird und wie sich diese Werte zueinander verhalten (Wendel 2020, 668, 670).[2] Vogel et al. 2019 untersuchen die Bedeutung des Adjektivs *geschäftsmäßig* in der juristischen Fachsprache und im allgemeinen Sprachgebrauch und stellen in ihrer Studie insoweit einen sehr unterschiedlichen Gebrauch fest (Vogel et al. 2019, S. 18).

Mielke & Wolff 2004 legen exemplarisch dar, inwieweit sich Text Mining-Verfahren für die Analyse großer Textkorpora, wie etwa die Berechnung von Kollokationen und deren Visualisierung, für die juristische Informationserschließung und das juristische Wissensmanagement eignen. Mielke & Wolff 2013 untersuchen konstrastiv das (österreichische) Allgemeine Bürgerliche Gesetzbuch (ABGB) und das (bundesdeutsche) Bürgerliche Gesetzbuch (BGB), weitere Studien zu Unterschieden und Gemeinsamkeiten der deutschen und österreichischen Fachsprache Recht im Bereich von Gerichtsentscheidungen finden sich in Mielke & Wolff 2016. Berteloot et al. 2018 und Auer et al. 2019 betrachten EuGH-Entscheidungen mittels korpus-linguistischer Methoden.

Weitere korpuslinguistische Analysen zu rechtlichen Fragestellungen im deutschsprachigen Raum stammen von Müller & Mastroardi 2014, Vogel et al. 2015 und Vogel et al. 2017.

2.2. Aktuelle Beispiele für die Vorgehensweise bei der Urteilsanalyse

Das von Wendel 2020 verwendete Korpus besteht aus 9.261 Entscheidungen des Bundesverfassungsgerichts, von denen 6.579 von der Internetseite des BVerfG stammen (Zeitraum: 1. 1. 1998 bis 31. 12. 2017) und 2.682. Dokumente aus der amtlichen Sammlung (bis 31. 12. 1997). Die Dokumente werden mit der Open Source Software GATE (***G**eneral **A**rchitecture for **T**ext **E**ngineering*, vgl. https://gate.ac.uk/) annotiert, d.h. relevante Textmuster werden gekennzeichnet, etwa der Tenor oder das Zitat einer Grundrechtsnorm (Wendel 2020, 670). Die wiederholte Auswertung von Stichproben ergibt 130 typische Formulierungen (z.B. «rügt die Beschwerdeführerin»), die als Anhaltspunkt für eine Rüge angesehen werden können. Ein weiterer Begriffskatalog wird für Grundrechtszitate erstellt, da die Grundrechte auch mit Worten umschrieben werden (z.B. «wegen Verstoßes gegen den Gleichheitssatz»). Nach einer von ihnen durchgeführten Fehlerabschätzung kann damit in 86 % der Verfassungsbeschwerden, die eine explizite Rüge enthalten, das gerügte Grundrecht gefunden werden. Mithilfe weiterer Annotationen werden Tabellen zu Metadaten aller Entscheidungen (Dateiname, Aktenzeichen, Spruchkörper, Entscheidungsdatum, beteiligte Richterin etc.) erstellt (Wendel 2020, 670). Die annotierten Texte werden als XML-Dokumente gespeichert, mit in Python geschriebenen Skripten auf relevante Annotationen durchsucht und mithilfe der Programmiersprache für Statistikanwendungen R ausgewertet (Wendel 2020, 670 ff. auch zum Umgang mit den vielfältigen Möglichkeiten von Grundrechtszitaten).

Vogel et al. 2019 vergleichen bei Ihrer Untersuchung zur Bedeutung des Adjektivs *geschäftsmäßig* u.a. das Deutsche Referenzkorpus am Institut für Deutsche Sprache in Mannheim (DeReKo, https://www1.ids-mannheim.de/kl/projekte/korpora.html), das vornehmlich nicht-fachsprachliche Texte (vor allem Pressetexte) in unterschiedlicher Zusammensetzung enthält, und das juristische Referenzkorpus (JuReKo, sh. oben), wobei es sich für ihre jeweilige Domäne um die jeweils weltweit größten Sammlungen linguistisch aufbereiteter Sprachdaten des Deutschen handelt (Vogel et al. 2019, S. 4). Sie ziehen dazu die relative Häufigkeit des Ausdrucks in den beiden Korpora über verschiedene Zeiträume heran und unterziehen die jeweiligen Belegstellen einer qualitativen Analyse (Vogel et al. 2019, S. 7 ff.). Vogel et al. 2015 analysieren 9.085 Texte arbeitsgerichtlicher Entscheidungen mit insgesamt 22,22 Mio. fortlaufenden Wortformen und erschließen das semantische Feld

[2] Korpus-basierte Arbeiten zur Auswertung von Entscheidungen des BVerfG durch den Doktorvater der Autorin Mielke aus dem Jahr 1985 werden als frühe Versuche, statistische Auswertungen durchzuführen, genannt (Schumann 1985, Wendel 2020, 669).

zur Zeichenkette *arbeitnehm* einschließlich diachroner Tendenzen in der Entwicklung des Arbeitnehmerbegriffs durch zwei unterschiedliche Teilkorpora (1990–1999 vs. 2000–2012) und kontrastieren den Begriff mit seinem Vorkommen in allgemeinsprachlichen Korpora (Vogel et al. 2015, S. 93 f., 98 ff., 135).

Bei den Untersuchungen von Mielke & Wolff 2013 und 2016 bzw. Berteloot et al. 2018 steht v.a. die Anwendung von Analysetools wie den *Voyant Tools* (https://voyant-tools.org/) oder Tools zu Verständlichkeitsanalysen im Vordergrund, in Auer et al. 2019 finden sich Stilometrie-Analysen von EuGH-Urteilen mittels gängiger Standard-Methoden in diesem Forschungsfeld.

3. Designziele für die Analyseplattform

Für die hier vorgestellte Analyseplattform standen folgende Designziele im Mittelpunkt: Es sollte eine Arbeitsplattform entstehen, die es ermöglicht, für frei zugängliche juristische Texte aktuelle Verfahren der automatischen Sprachverarbeitung zu erproben. Dabei galt es, neuere Analyseformen wie *sentiment analysis* oder Stilometie ebenso zu berücksichtigen wie die technischen Fortschritte beim Einsatz maschinellen Lernens z.B. für das Training von Sprachmodellen, die auch für die «traditionellen» Kernverfahren der maschinellen Sprachverarbeitung (Information Retrieval und Inhaltsextraktion, Dokumentvergleich, maschinelle Übersetzung) von Bedeutung sind. Der Trend zu Deep Learning-Verfahren mit Hilfe künstlicher neuronaler Netze hat sich mittlerweile in als open source-Software gut verfügbaren NLP-Komponenten niedergeschlagen, sodass diese beiden Ziele (neue Analysetechniken / neue NLP-Verfahren) gut realisierbar erschienen. Die Entwicklung der NLP-Komponenten selbst war dabei kein eigenständiges Ziel, vielmehr ging es darum, aktuelle verfügbare NLP-Komponenten zu identifizieren und einzusetzen.

4. EuGH-Urteile als Datengrundlage

Die Auswahl der Urteile des europäischen Gerichtshofs ergab sich aus der guten technischen Verfügbarkeit (Webplattform mit frei zugänglichen Gerichtsurteilen sowie eine eigene Programmierschnittstelle, über die die Datenabfrage dynamisch eingebunden werden kann), aus den bereits vorliegenden Vorarbeiten (siehe oben Kap. 2 und Berteloot et al. 2018, Auer et al. 2019), der vergleichsweise guten Anreicherung der EuGH-Urteile mit Metadaten sowie aus den zusätzlichen relevanten NLP-Fragegestellungen, die sich für ein derartiges mehrsprachiges und damit teilweises paralleles Korpus ergeben. Insgesamt umfasst das über den Prototyp der Workbench aufgebaute Gesamtkorpus 24 Subkorpora in den unterschiedlichen Amtssprachen mit je ca. 5.500 Urteilen des EuGH (Gesamtumfang der Texte: ca. 1,2 GB).

5. Technische Konzeption und Umsetzung

Die funktionalen Komponenten für die Dokumentakquise und die Serversteuerung wurden mit Hilfe der Programmiersprache Python (Version 3.7, https://www.python.org/downloads/release/python-370/) als Skripte realisiert, wobei für die Datenverwaltung die NOSQL-Datenbank MongoDB (https://www.mongodb.com/) genutzt wird. Zusätzlich kommen der *in-memory-store Redis* (https://redis.io/) für eine effiziente Datenverarbeitung und *Celery* (https://docs.celeryproject.org/en/stable/) als Werkzeug zur Lastbalancierung bei daten- und rechenintensiven Analysen zum Einsatz. Die Entwicklung der Komponenten erfolgt nach Methoden des agilen Projektmanagements (Scrum; *Trello* als Kanban-artiges Projektmanagement-Tool, Nutzung von *Github* als Code-Repository). Für die Realisierung der Benutzerschnittstelle als *progressive web app* dient das Javascript-Framework Node.js. *Progressive web apps* haben sich als Design Pattern für plattformübergreifend nutzbare Browser-basierte Anwendungen etabliert.[3]

[3] Vgl. https://developer.mozilla.org/en-US/docs/Web/Progressive_web_apps/App_structure.

5.1. Software-Architektur der Anwendung

Die Software-Architektur der Anwendung umfasst drei wesentliche Komponenten:

1. Die Prozesskette zur *Erstellung des Korpus*: Über das Application Programming Interface zu EUR-LEX lassen sich bis zu 10.000 Dokumente täglich herunterladen, was für die vorliegende Anwendung ausreichend ist. Voraussetzung für den Datenzugriff sind ein gültiger Account für die Datendienste der EU (https://webgate.ec.europa.eu/cas/login) und eine Registrierung für das EUR-LEX-Datenzugriff-API als *web service* (https://eur-lex.europa.eu/content/help/faq/reuse-contents-eurlex-details.html).
2. Die *Serveranwendung*, die unterschiedliche Komponenten der automatischen Sprachverarbeitung einbindet und auf der Basis von JSON[4] ein einfaches Abfragemodell realisiert, über das Korpus-Sprache, Korpus-Auswahl und eine Auflistung der gewünschten Analyseschritte an den Server geschickt werden kann.
3. Eine *Web-App*, mit deren Hilfe Benutzer flexibel unterschiedliche Komponenten zur Analyse der Dokumente einsetzen und anwenden können. Ein wichtiger Aspekt ist hier die Visualisierung von Analyseergebnissen.

5.2. Korpusaufbau

Die Schnittstellen, über die die Rechtstexte einschließlich ihrer Metadaten aus den öffentlich zugänglichen Datenbanken der europäischen Union extrahiert werden können, setzen JSON als einfaches Beschreibungsformat ein. Mit Hilfe des Steuerskriptes für den Korpusaufbau lassen sich Urteile in allen verfügbaren Sprachen, in einer einzelnen Sprache oder für vom Nutzer ausgewählte Sprachen aus EUR-LEX exportieren. Dabei werden entweder kommaseparierte Textdateien (CSV, comma-separated values) generiert oder die Daten werden direkt in eine MongoDB-Datenbank importiert.

5.3. Prozesskette zur Aufbereitung und Analyse der Urteile

Anders als in früheren Studien kommen hier vortrainierte Sprachmodelle zum Einsatz, wobei für die englische Sprache ein bereits auf das Rechtswesen hin optimiertes Modell verfügbar ist. Die beiden für Deutsch und Englisch eingesetzten Modelle basieren auf *spaCy*, einer Softwarewareumgebung für die automatische Sprachverarbeitung, die darauf abzielt, robuste Analyseergebnisse zu liefern («production usage», https://spacy.io/usage/facts-figures#other-libraries). Im Unterschied zu früheren Frameworks wie NLTK (*natural language toolkit,* https://www.nltk.org/, vgl. Burghardt et al. 2014) sind in *spaCy* auch neuere Verfahren aus dem Bereich *deep learning*, also avanciertere Formen künstlicher neuronaler Netzwerke integriert, z.B. werden *convolutional neural networks* (CNN) für die *named entity extraction* genutzt. Ergänzend kommen die *textaCy*-Python-Bibliothek (DeWilde 2020), die auf *spaCy* aufbaut, sowie die *Stanza*-NLP-Bibliothek zum Einsatz (Qi et al. 2020).

In einer ersten Fallstudie werden Komponenten und Sprachmodelle für Deutsch und Englisch verwendet. Dabei kamen folgende trainierte Sprachmodelle zum Einsatz:

– Für die englische Sprache das *Blackstone*-Modell des Forschungslabors des *Incorporated Council of Law Reporting for England and Wales* (ICLR&D)[5], das bereits für spezifische Aspekte der englischen Rechtssprache trainiert ist (Erkennen von Fall-Namen und Fall-Verweisen, Richternamen, Gesetzesverweisen, allgemein *named entity recognition*).

[4] *JavaScript Object Notation*, ein leichtgewichtiges Datenstrukturformat, das für webbasierte Anwendungen große Verbreitung gefunden hat, vgl. https://www.json.org/json-en.html.

[5] Vgl. https://research.iclr.co.uk/blackstone: «Blackstone is a spaCy model and library for processing long-form, unstructured legal text.», nach eigener Aussage das einzige Sprachmodell für die Verarbeitung englischer Rechtstexte, das nicht in einem kommerziellen Kontext entstanden ist.

– Für die deutsche Sprache wird ein mit Hilfe bekannter deutschsprachiger Korpora trainiertes *spaCy*-Modell (https://spacy.io/models/de#de_core_news_md) eingesetzt, das zwar ebenfalls u.a. *named entity recognition* ermöglicht, aber nicht für Rechtstexte spezialisiert ist.

Die mit diesen Modellen mögliche Sprachverarbeitungs-Prozesskette sieht Komponenten für typische Arbeitsschritte vor wie

– die Vorbereitung der Dokumente (Entfernen von Sonderzeichen wie Interpunktion oder wiederkehrenden Textelementen wie Kolumnentiteln, Entfernen von Stoppwörtern),
– die Normalisierung der Materialien,
– die Einteilung der Texte in einzelne Token,
– das Part-of-Speech-Tagging,
– die Erkennung von Eigennamen (*named entity recognition*) sowie auch linguistische Verarbeitungsschritte wie
  – die Grundformreduktion und Lemmatisierung sowie die
  – Erkennung auffälliger N-Gramme / Kollokationen.

Zusätzlich können Frequenzstatistiken (Worthäufigkeiten, Wortlängen, Satzlängen etc.) erstellt oder Analysen auf bestimmte Frequenzklassen beschränkt werden. Auch Untersuchungen, die sich auf Teilkorpora oder einzelne Dokumente beziehen, sind möglich (Anzahl Typens / Tokens / Sätze im Subkorpus oder Dokument).

5.4. Fortgeschrittene Analyseschritte

Die folgenden Komponenten auf der Basis neuerer Verfahren konnten zusätzlich in den Prototyp eingebunden werden:

1. Schlagwortextraktion: Mit Hilfe des *PositionRank*-Verfahrens (Florescu & Caragea 2017) werden Schlüsselbegriffe und -phrasen aus den Dokumenten extrahiert.
2. Untersuchung der Dokumentähnlichkeit: Erlaubt die paarweise Berechnung einer Dokumentähnlichkeit und baut auf dem *Word Embeddings*-Verfahren auf (Mikolov et al. 2013).
3. Metriken zur Lesbarkeit: Die Lesbarkeit des Textes wird nach dem Flesch-Reading-Ease analog zu Carlson et al. 2015, S. 1481 ff. berechnet, wobei die Funktionen aus der *textaCy*-Bibliothek zum Einsatz kommen.
4. Sentiment-Analyse: Ein normalisiertes Ergebnis einer Sentimentanalyse wird für das betrachtete (Teil-) Korpus berechnet (vgl. Schmidt et al. 2019).

5.5. Benutzerschnittstelle der Analyseanwendung

Über die Benutzerschnittstelle lassen sich Anfragen flexibel zusammenstellen, die verschiedene der soeben aufgeführten Metriken umfassen. Die Web-basierte Analyseoberfläche baut auf aktuellen Javascript-Bibliotheken wie React.js auf und erlaubt einen einfachen Zugriff auf die unterschiedlichen Komponenten, ohne vertieftes technisches Wissen vorauszusetzen. Zur Visualisierung von Ergebnissen kommen neben tabellarischen Darstellungen Visualisierungsformate wie Wortwolken oder Balkendiagramme zum Einsatz (Abb. 1).

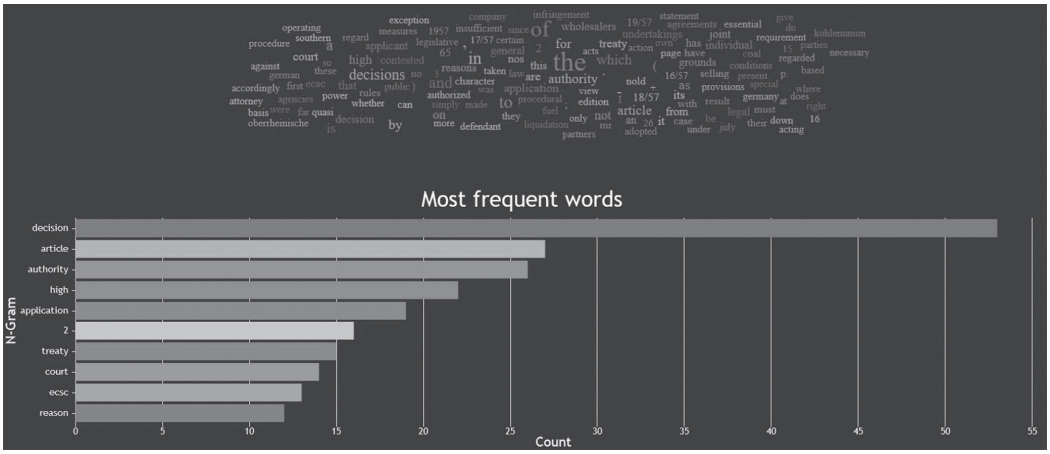

Abb. 1: Visualisierungsformate Wortwolke und Balkendiagramm (Quelle: Fischer et al. 2020, S. 18)

6. Fazit und Ausblick

Mit der Übersicht zum aktuellen Stand des Einsatzes vom NLP-Verfahren im juristischen Bereich und der Realisierung des Prototyps einer arbeitsfähigen Analyseplattform ist ein erstes Etappenziel erreicht. In der nächsten Phase haben wir vor, für das gewählte Korpus Analysefragestellungen zusammenzustellen, Analysen durchzuführen und die Leistungsfähigkeit der Komponenten zu bewerten. Die Benutzerschnittstelle der Plattform wird dabei weiterentwickelt werden, wobei wir dazu das Informationsverhalten typischer Nutzer untersuchen wollen.

Erneut wurde deutlich, dass in der existierenden Toollandschaft die englische Sprache am besten unterstützt wird – der Aufbau vortrainierter Modelle für deutsche Rechtstexte und die deutsche Rechtssprache ist ein Desiderat für die Verbesserung der automatischen Analyse (nicht nur) von Urteilen. Auch die Weiterentwicklung der Analysemethoden selbst ist ein Forschungsziel, beispielsweise im Bereich der bisher vorherrschenden eher formalistischen Lesbarkeitsmetriken.

Literatur

Abegg, Andreas/Bubenhofer, Noah, Empirische Linguistik im Recht: Am Beispiel des Wandels des Staatsverständnisses im Sicherheitsrecht, öffentlichen Wirtschaftsrecht und Sozialrecht der Schweiz, Ancilla Iuris, 2016, S. 1–41.

Auer, Anna-Maria/Berteloot, Pascale/Mielke, Bettina/Schikora, Christine/Schmidt, Thomas/Wolff, Christian, Stilometrie in der Rechtslinguistik. Nutzung korpuslinguistischer Verfahren für die Analyse deutschsprachiger Urteile. In: Erich Schweighofer/Franz Kummer/Ahti Saarenpää (Hrsg.), Internet of Things. Tagungsband des 22. Internationalen Rechtsinformatik Symposions IRIS 2019. Proceedings of the 22nd International Legal Informatics Symposium IRIS 2019, Editions Weblaw, Bern 2019, S. 375–384.

Berteloot, Pascale/Mielke, Bettina/Wolff, Christian, Deutsches, österreichisches, europäisches Deutsch? Deutschsprachige Fassungen von Urteilen des europäischen Gerichtshofs im Vergleich. In: Erich Schweighofer/Franz Kummer/Ahti Saarenpää/Burghardt Schafer (Hrsg.), Datenschutz / LegalTech. Data Protection / LegalTech. Tagungsband des 21. Internationalen Rechtsinformatik Symposions IRIS 2018. Proceedings of the 21st International Legal Informatics Symposium IRIS 2018, Editions Weblaw, Bern 2018, S. 319–324.

Burghardt, Manuel/Pörsch, Julian/Tirlea, Bianca/Wolff, Christian, WebNLP – An Integrated Web-Interface for Python NLTK and Voyant, Proceedings of the 12th edition of the KONVENS conference Vol. 1., 2014, S. 235–240.

Carlson, Keith/Livermore, Michael A/Rockmore, Daniel, A Quantitative Analysis of Writing Style on the US Supreme Court, Wash. UL Rev., 2015, 93, S. 1461–1510.

Dewilde, Burton, Textacy Documentation, Release 0.10.1, August 2020, https://textacy.readthedocs.io/en/latest/index.html.

Eder, Maciej, Rolling Stylometry, Digital Scholarship in the Humanities, 2016, 31, S. 457–469.

Fischer, Thomas/Hartl, Philipp /Hilzenthaler, Andreas/Jackermeier, Lukas, Aufbau und Analyse eines Korpus mit Entscheidungen des europäischen Gerichtshofs. Projektpräsentation, Masterstudiengang Medieninformatik, Universität Regenburg, 2020.

Florescu, Corina/Caragea, Cornelia, Positionrank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, 1105–1115.

Ighreiz, Ali/Möllers, Christoph/Rolfes, Louis/Shadrova, Anna/Tischbirek, Alexander, Karlsruher Kanones: Selbst- und Fremdkanonisierung der Rechtsprechung des Bundesverfassungsgerichts. *Archiv des öffentlichen Rechts* (angenommen).

Livermore, Michael A./Riddell, Allen B./Rockmore, Daniel N., The Supreme Court and the Judicial Genre, Ariz. L. Rev., 2017, 59, S. 837–901.

Mielke, Bettina/Wolff, Christian, Text Mining-Verfahren für die Erschließung juristischer Fachtexte. In: Erich Schweighofer/Doris Liebwald/Günther Kreuzbauer/Thomas Menzel (Hrsg.), Informationstechnik in der juristischen Realität, Aktuelle Fragen der Rechtsinformatik 2004, Verlag Österreich, Wien 2004, S. 269–279.

Mielke, Bettina/Wolff, Christian, Österreichisch-deutsche Rechtssprache kontrastiv: Eine corpuslinguistische Analyse. In: Erich Schweighofer/Franz Kummer/Walter Hötzendorfer (Hrsg.), Abstraktion und Applikation. Abstraction and Application. Tagungsband des 16. Internationalen Rechtsinformatik Symposions IRIS 2013. Proceedings of the 16th International Legal Informatics Symposium, Österreichische Computer Gesellschaft, Wien 2013, S. 377–384.

Mielke, Bettina/Wolff, Christian, Österreichische und Deutsche Gerichtsentscheidungen im Sprachvergleich. In: Erich Schweighofer/Franz Kummer/Walter Hötzendorfer/Georg Borges (Hrsg.), Netzwerke. Networks. Tagungsband des 19. Internationalen Rechtsinformatik Symposions IRIS 2016. Proceedings of the 19th International Legal Informatics Symposium, Österreichische Computer Gesellschaft & Erich Schweighofer, Wien 2016, S. 129–138.

Mikolov, Tomas/Sutskever, Ilya/Chen, Kai/Corrado, Greg S./Dean, Jeff, Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems, 2013, 26, S. 3111–3119.

Mouritsen, Stephen C, The Dictionary is not a Fortress: Definitional Fallacies and a Corpus-Based Aproach to Plain Meaning, Brigham Young University Law Review 2010, S. 1915–1980.

Qi, Peng/Zhang, Yuhao/Zhang, Yuhui/Bolton, Jason/Manning, Christopher D., Stanza: A Python Natural Language Processing Toolkit for many Human Languages.

Piotrowski, Michael, Computerlinguistik und Digital Humanities, DHd – Digital Humanities im deutschsprachigen Raum, 2019, https://dhd-blog.org/?p=2532.

Schmidt, Thomas/Burghardt, Manuel/Wolff, Christian, Herausforderungen für Sentiment Analysis-Verfahren bei literarischen Texten. In: Manuel Burghardt/Claudia Müller-Birn (Hrsg.), Im Spannungsfeld zwischen Tool-Building und Forschung auf Augenhöhe – Informatik und die Digital Humanities. Proceedings Workshop INF-DH-2018, Gesellschaft für Informatik e. V., Bonn 2018.

Schumann, Ekkehard, Die Wahrung des Grundsatzes des rechtlichen Gehörs – Daueraufträg für das BVerfG?, NJW, 1985, S. 1134–1140.

Sippl, Colin/Burghardt, Manuel/Wolff, Christian/Mielke, Bettina, Korpusbasierte Analyse österreichischer Parlamentsreden. In: Erich Schweighofer/Franz Kummer/Walter Hötzendorfer/Georg Borges (Hrsg.), Netzwerke. Networks. Tagungsband des 19. Internationalen Rechtsinformatik Symposions IRIS 2016. Proceedings of the 19th International Legal Informatics Symposium, Österreichische Computer Gesellschaft & Erich Schweighofer, Wien 2016, S. 139–148.

Varsava, Nina, Elements of Judicial Style: A Quantitative Guide to Neil Gorsuch's Opinion Writing, NYUL Rev. Online, 2018, 93, S. 75–112.

Vogel, Friedemann /Bäumer, Benjamin/Deus, Fabian /Rüdiger, Jan Oliver/Tripps, Felix, Die Bedeutung des Adjektivs geschäftsmäßig im juristischen Fach- und massenmedialen Gemeinsprachgebrauch. Eine rechtslinguistische Korpusstudie als Beispiel für computergestützte Bedeutungsanalyse im Recht, in: LeGes 30 (2019), 3, S. 1–20.

Vogel, Friedemann/Hamann, Hanjo/Gauer, Isabelle, Computer-Assisted Legal Linguistics: Corpus Analysis as a New Tool for Legal Studies, Law & Social Inquiry, 2018, 43, S. 1340–1363.

Vogel, Friedemann/Pötters, Stephan/Christensen, Ralph, Richterrrecht der Arbeit – empirisch untersucht. Möglichkeiten und Grenzen computergestützter Textanalyse am Beispiel des Arbeitnehmerbegriffs, Berlin 2015.

Wendel, Luisa, Welche Grundrechte führen zum Erfolg?, Eine quantitative, korpusgestützte Untersuchung anhand von Entscheidungen des Bundesverfassungsgerichts, JZ, 2020, S. 668–679.