GENERALISIERUNG VON FORMELHAFTEN TEXTBESTANDTEILEN IN JURISTISCHEN KORPORA: EINSATZ- UND ENTWICKLUNGSPOTENTIAL

Frieda Josi / Christian Wartena / Ulrich Heid

Frieda Josi, Dissertantin, Universität Hildesheim, Institut für Informationswissenschaft und Sprachtechnologie
Lübecker Str. 3, 31141 Hildesheim, Deutschland, DE
Frieda.Josi@hs-hannover.de

Christian Wartena, Professor, Hochschule Hannover, Fachbereich Informationsmanagement
Expo Plaza 12, 30539 Hannover, Deutschland, DE
Christian.Wartena@hs-hannover.de; http://textmining.wp.hs-hannover.de

Ulrich Heid, Professor, Universität Hildesheim, Institut für Informationswissenschaft und Sprachtechnologie
Lübecker Str. 3, 31141 Hildesheim, Deutschland, DE
Heid@uni-hildesheim.de; https://www.uni-hildesheim.de/fb3/institute/iwist/mitglieder/heid

Schlagworte: ***Standardisierung, Rechtsdokumente, Text Mining, Formelhafte Textabschnitte***

Abstract: *Generalisierte Rechtsdokumente, bei denen für die individuellen Ausprägungen eines Vertrages die Positionen im Text bekannt sind, können eingesetzt werden, um erstens das Genehmigungsverfahren von Neuverträgen automatisiert zu unterstützen und zweitens als Vertragsgenerator neue Rechtsdokumente vorausgewählt zur Verfügung zu stellen. In diesem Beitrag wird, mithilfe von bekannten juristischen Texten gezeigt, wie formelhafte Textabschnitte identifiziert und häufige individuelle Ausprägungen klassifiziert werden können, um als Musterabschnitte eingesetzt zu werden. Es werden Einsatzbereiche vorgestellt und vorhandenes Potential für Legal Tech-Anwendungen aufgezeigt.*

1. Einleitung

Im Genehmigungsverfahren von Verträgen kann eine automatisierte Analyse der vorherigen genehmigten Verträge dazu eingesetzt werden, standardisierte oder häufig wiederkehrende Bestandteile der Vertragstexte zu prüfen. Diese müssen anschließend nur bei Abweichungen zur manuellen Sichtung vorgelegt werden. Auch für das Aufsetzen eines Vertragstextes kann eine Sammlung von vorformulierten Textabschnitten aus juristischen Dokumenten genutzt werden, um Rechtssicherheit zu gewährleisten. Diese formelhaften Textabschnitte sind Bestandteil des impliziten Wissens von JuristInnen, müssen jedoch von diesen erst u. a. über eigene Erfahrung mit dem Aufsetzen von Vertragsdokumenten erworben werden. Durch die automatisierte Analyse und Identifizierung dieser formelhaften Textabschnitte kann dieses implizite Wissen explizit gesichert, durchsuchbar und abrufbar gemacht werden.

2. Stand der Forschung und aktuelle Legal Tech-Anwendungen

Das Verfahren, das in diesem Beitrag vorgestellt wird, basiert auf einer Vielzahl von Forschungsansätzen. Zum einen sind es Arbeiten, die sich mit der Segmentierung von Texten beschäftigen. Formelhafte Textabschnitte, die in Dokumenten identifiziert werden müssen, können als Sequenzen von Sätzen definiert werden. Auf dem Gebiet der Sequenzerkennung und beim Vergleich dieser Sequenzen, beispielsweise mit Sequenzen in einer Datenbank, gibt es einige Arbeiten, die z.B. das Problem der längsten gemeinsamen Teilsequenz (Longest

Common Subsequence)[1] und auch das Problem des längsten wiederholten Substrings (Longest Repeating Sequence)[2, 3] lösen. Eine Zielsetzung bei der Sequenzerkennung ist, dass wiederkehrende Teilsequenzen, die im Voraus nicht bekannt sind, identifiziert werden. Es muss eine Entscheidung getroffen werden, was eine geeignete Teilsequenz ist, bevor in den Korpora nach den wiederverwendeten Teilsequenzen gesucht werden kann. Dies kann beispielsweise durch die Modellierung der kürzesten Korpusbeschreibung[4] geschehen, oder durch die Mindestbeschreibungslänge (MDL)[5] für die Wortsegmentierung, oder der Sequenzsegmentierungen,[6] oder durch sequentielles Pattern Mining.[7, 8]

Auch Legal Tech-Unternehmen entwickeln Anwendungen und Lösungen, die unterstützend im Genehmigungsverfahren und für die Vertragsgenerierung eingesetzt werden. Das Unternehmen *Glanos GmbH* ist Entwickler der Software *DataSphere*[9]. Die Möglichkeiten, die diese Software bieten soll, sind u. a.: die Bestandsanalyse von vorhandenen Verträgen; eine automatische Extraktion von Fristen, Terminen und Auslaufzeiten in Leasingverträgen; Hinweis auf fehlende Klauseln und Vorschläge für die Formulierung dieser Klauseln. Für die Umsetzung setzen die Entwickler hauptsächlich auf eine kontinuierliche Erweiterung von Regeln und nutzen Textkorpora mit juristischen Fachbegriffen für eine Schlagwort- und Synonymanalyse. Die *IntraFind Software AG* bietet mit *AnalyzeLaw* eine Software für Kanzleien und Rechtsabteilungen in Unternehmen an.[10] Neben dem Zuordnen der Verträge zu den jeweiligen Rechtsgebieten, werden auch weitere Inhalte extrahiert, z.B. Personen, Vertragsparteien und Vertragswert. Des Weiteren identifiziert die Software ungültige Klauseln. Nach eigenen Angaben setzt das Unternehmen dafür Machine Learning-Verfahren und Methoden der künstliche Intelligenz ein.

3. Vorverarbeitung anhand Dokumentformat und juristischen Textphänomenen

Ausgehend vom pdf-basiertem Dokument fallen einige umfangreiche Vorverarbeitungsschritte an, um einen bereinigten Text zur Weiterverarbeitung nutzen zu können. Diese Vorarbeiten haben wir in *Structural Analysis of Contract Renewals*[11] und *Preparing Legal Documents for NLP analysis*[12] detailliert beschrieben. Auf Dokumentebene müssen die Anordnung und die Funktion der Textteile beibehalten werden. Der Aufgabenbereich wird in Abschnitt 3.1 überblicksartig vorgestellt. Für die Bereinigung von Texten, die aus juristischen Dokumenten extrahiert werden, sind zusätzlich zu den Standardverfahren aus der natürlichen Sprachverarbeitung (Computerlinguistik), weitere Bereinigungsschritte notwendig. Diese werden im Abschnitt 3.2 vorgestellt. Wir setzen für unsere Verfahren zwei Korpora ein. Ein Korpus besteht aus Urteilen aus dem Strafrecht

1 Chvátal, V. u. Sanko, D. Longest common subsequences of two random sequences. Journal of Applied Probability 12 (2), 1975, 306–315.

2 Lloyd, A. Suffix Trees for LCS, Implementation, http://www.allisons.org/ll/AlgDS/Tree/Suffix. Zugriff am 2020-06-26, 2008.

3 Merkurev, O. u. Shur, A.M. Searching Long Repeats in Streams. In N. Pisanti & S. P. Pissis (Hrsg.), 30th annual symposium on combinatorial pattern matching (cpm 2019) (Bd. 128). Dagstuhl, Germany: Schloss Dagstuhl-Leibniz-Zentrum für Informatik. 2019.

4 Rissanen, J. Modeling by shortest data description. Automatica, 14 (5), 1978, 465–471.

5 Robinet, V. u. Lemaire, B. MDLChunker: A MDL-based model of word segmentation. Proceedings of the 31st Annual Meeting of the Cognitive Science Society, 2009, 2866–2871.

6 Chan, S., Kao, B., Yip, C. u. Tang, M. Mining emerging substrings. In Eighth international conference on database systems for advanced applications, 2003. (DASFAA 2003).

7 Fournier-Viger, P., Gomariz, A., Campos, M. u. Thomas, R. Fast vertical mining of sequential patterns using co-occurrence information. In: Advances in knowledge discovery and data mining (Bd. 8443, S. 40–52). Springer International Publishing. 2014.

8 Singer, F. u. Lemmerich, P. Analyzing sequential user behavior on the web. In Proceedings of the 25th international conference companion on world wide web. International World Wide Web Conferences Steering Committee. 2016, 1035–1036.

9 Bauer, C. u. Rolletschek, G. DataSphere glanos GmbH. Zugriff am 2020-08-17 auf https://www.glanos.de/datasphere. 2020.

10 KÖGL, F. u. MESSER, B. Legaltech für Juristen und Unternehmensrechtsabteilungen. Zugriff am 2020-08-17 auf https://www.intrafind.de/loesungen/branchen/juristen-und-unternehmensrechtsabteilungen. 2020.

11 Josi, F. u. Wartena, C.: Structural Analysis of Contract Renewals. In: Proceedings of the ACM CIKM 2018 Workshops. Turin, 2018.

12 Josi, F., Wartena, C., u. Heid, U. Preparing Legal Documents for NLP analysis: Improving the Classification of Text Elements by using Page Features, NATP 2022, 8th International Conference on Natural Language Processing, Zürich 2022, Manuskript eingereicht zur Publikation.

(aus den Jahren von 1999 bis Anfang 2020), die vom BGH gesprochen wurden. Ein zweites Korpus besteht aus Verträgen, die von den Stadtverwaltungen Hamburg und Bremen aufgrund von Transparenzstrategien öffentlich zur Verfügung gestellt werden und aus einigen Verträgen zwischen Hochschulen und Servicepartnern, die ebenfalls frei verfügbar sind.[13]

3.1. Textextraktion aus gezeichneten PDF-Dokumenten

Für die Textextraktion klassifizieren wir zuerst das Seitenlayout jeder Dokumentseite. In den Dokumenten, die sich im Korpus befinden, sind das einspaltige, zweispaltigen Seiten und Seiten, die einspaltig sind jedoch links Überschriften in Form von Marginalien haben. Dieser Schritt ist notwendig, damit der Textfluss des Dokumentes für die nachfolgende Analyse erhalten bleibt. Der Klassifikator teilt eine Seite in viele schmalen Spalten auf und lernt aus den (normalisierten) Textmengen pro Spalte, zu welcher Layoutklasse eine Seite gehört.[14]

3.2. Textbereinigung

Wenn das Seitenlayout der Dokumentseiten bekannt ist, werden alle Textelemente Textklassen zugeordnet. Dies geschieht ebenfalls mithilfe eines Klassifikators, der mit Merkmalen, die die Textfunktion beschreiben, trainiert wurde. Die Merkmale, die eingesetzt wurden, sind beispielsweise Information zur Platzierung auf der Seite, die Angabe der Schriftgröße und weitere. Dabei werden die Textklassen *Kopfzeile*, *Überschrift*, *Text*, *Aufzählungszeichen*, *Stempel und Unterschrift*, *Fußzeile* klassifiziert. Durch das Wissen über die Textklasse, kann Wissen über die Funktion der Textelemente erhalten bleiben, z. B. dass das Textelement eine Überschrift abbildet. Durch diesen Vorgang können wir Textelemente aus der Kopf- und Fußzeile, sowie aus den Bereich für Unterschriften und Stempel herausfiltern und erhalten den bereinigten Vertragsinhalt. Da bei der Textextraktion aus PDF-Dokumenten auch viele Worttrennungen und bei der Texterkennung auch viele OCR-Fehler (Optical Character Recognition) auftreten, sind weitere Methoden notwendig, um den Text zu korrigieren. Da die Fachtextsorten *Vertrag* und *Urteil* eine Vielzahl von Abkürzungen einsetzen, sind Standardmethoden mit eigenen Texten weiter trainiert worden, um den Text optimal in einzelne Sätze aufzuteilen.

4. Generalisierung von formelhaften Textabschnitten

Die Generalisierung von formelhaften Textabschnitten in juristischen Texten soll dazu eingesetzt werden, schematische Textbereiche zu definieren, die sehr häufig in Dokumenten verwendet wurden, die jedoch musterhafte Variationen zulassen. Diese Variationen können einzelne Begriffe bzw. Satzteile sein.[15] Durch die Verbindung von formelhaften Abschnitten und den identifizierten variablen Begriffen kann eine umfangreiche Vertragsanalyse durchgeführt werden. Wenn beispielsweise ein formelhafter Abschnitt in einem zu prüfendem Vertrag identifiziert wird und die möglichen variablen Begriffe wurden nicht verwendet, so ist dieser Abschnitt einer manuellen Prüfung vorzulegen. Sollten die variablen Änderungen in den formelhaften Abschnitten jedoch einer bekannten Variante entsprechen z. B. dem Namen des Gerichts, ist dieser Vertragsteil inkl. der individuellen Anpassungen dennoch automatisiert analysierbar.

[13] Die Quelle für die Urteile aus dem Strafrecht ist: https://www.hrr-strafrecht.de/hrr/db. Die Quellen für das Korpus, das aus Verträgen aufgebaut ist, wurden veröffentlicht unter: http://textmining.wp.hs-hannover.de/juver.html. Auf dieser Projektseite werden alle Korpora als Zip-Datei zur Verfügung gestellt.

[14] Eigene Veröffentlichungen, siehe Abschnitt 3.

[15] Diese Satzteile können auch als Routineausdrücke bezeichnet werden. Siehe Płominska, M.: Routine expressions in german legal texts – an attempt at typology. Colloquia Germanica Stetinensia 29, 239–253, 2020.

4.1. Verfahren für die Identifizierung von formelhaften Textabschnitten

Das detaillierte Verfahren für die Identifizierung der formelhaften Textabschnitten wird in [JOSI et al. 2021] beschrieben. Nach der Extraktion und Bereinigung der Texte werden diese in Sätze geteilt und zu ähnlichen Satzclustern gruppiert. Durch die Gruppierung kann jeder Satz in einem Dokument durch das entsprechende Satzcluster repräsentiert werden. Mithilfe des Apriori-Algorithmus[16] können längste wiederkehrende Sequenzen in der Dokumentensammlung ermittelt werden. Diese Satzsequenzen stellen häufig vorkommende formelhafte Textabschnitte dar.[17]

4.2. Beispiele und Ergebnisse

In diesem Abschnitt werden sowohl formelhafte Textabschnitte, als auch individuelle Ausprägungen der Varianten gezeigt. Das Forschungsprojekt ist noch nicht abgeschlossen und die vorgestellten Ergebnisse bilden somit nur den aktuellen Stand des entwickelten Verfahrens ab.

4.2.1. Formelhafte Textabschnitte

4.2.1.1. Formelhafte Textabschnitte im Korpus „Urteile aus dem Strafrecht“

In Abbildung 1 sind die Satzfolgen [25487,5712,907,58], [25487,3911,907,58] und kürzere Teilsequenzen für einen formelhaften Textabschnitt aus dem Korpus „Urteile aus dem Strafrecht“ aufgezeigt. In der Abbildung 2 wird für jedes Satzcluster in dieser Sequenz ein Satzbeispiel geben. Die Zahlen an den Kanten (Pfeilen) geben die Anzahl der in unserem Korpus beobachteten Vorkommen der jeweiligen Sequenzen an.

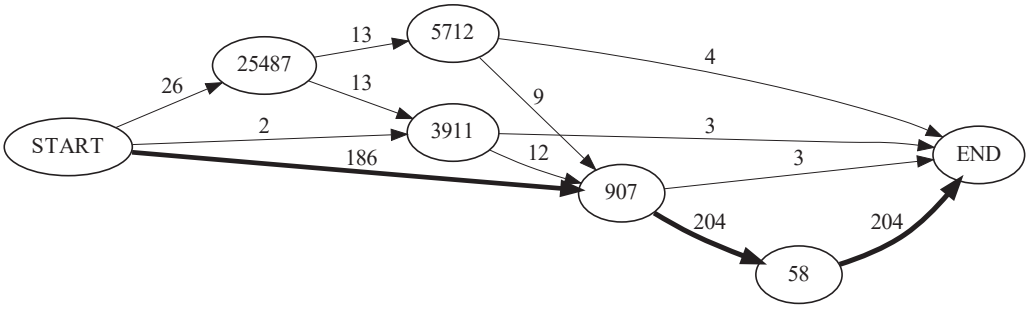

Abbildung 1: Formelhafter Abschnitt aus dem Korpus „Urteile StGB“.

[16] Agrawal, Rakesh u. Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. 1994, 487–499.

[17] Josi, F., Wartena C. u. Heid, U. Representing Standard Text Formulations as Directed Graphs. Springer International Publishing, Cham, 2021, 475–487.

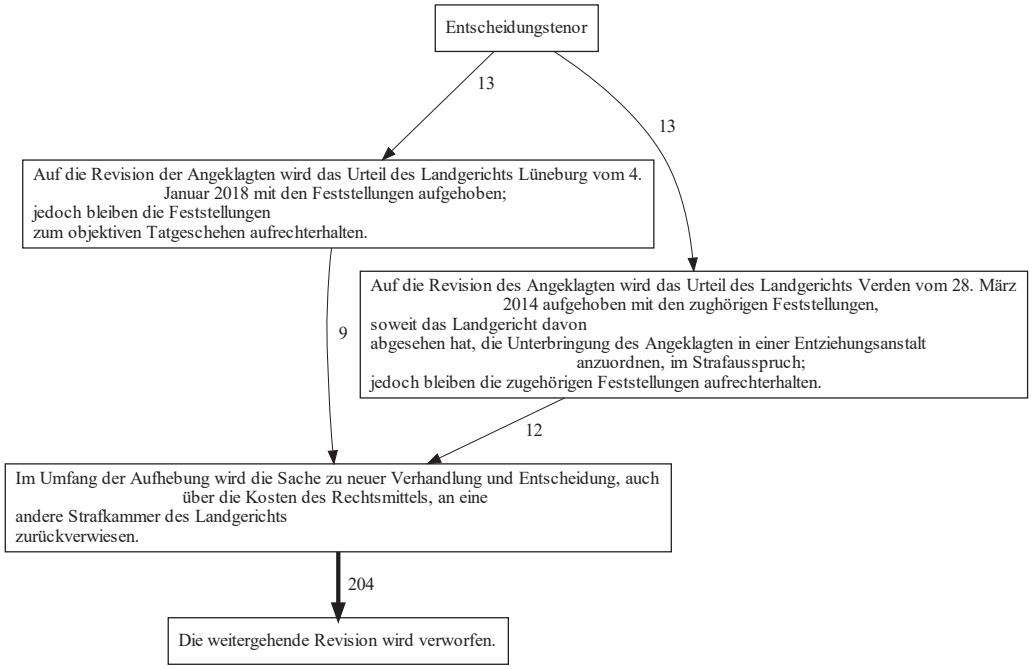

Abbildung 2: Variation in einem formelhaften Textabschnitt, Korpus „Urteile zum StGB“

Das Beispiel aus Abbildung 1 und 2 zeigt den Entscheidungstenor, der formelhaft aus einer Abfolge von bis zu drei Sätzen bestehen kann. Dies ist für 13 Urteile der Fall. In weiteren Urteilen wurden kürzere Teilabschnitte identifiziert.

4.2.1.2. Formelhafte Textabschnitte im Korpus „Verträge aus Stadtverwaltungen und Hochschulen“

In Abbildung 3 ist der formelhafte Abschnitt für die Satzfolgen [623,624,1239], [623,624,3487,11535] und einige kürzere Abfolgen abgebildet. Diese Satzfolgen kommen im Korpus der Verträge vor und zeigen Bestimmungen zum Vereinbarungszeitraum auf. In Abbildung 4 wird jeweils ein Satz aus dem Satzcluster gezeigt. Die längsten möglichen Satzfolgen kommen jeweils neun- und elfmal im Korpus vor.

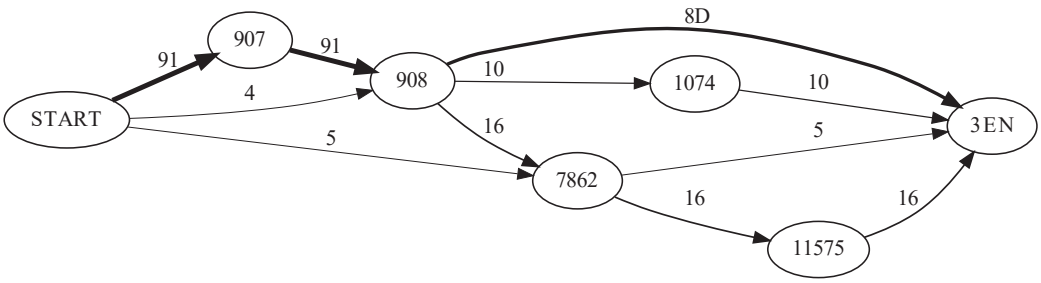

Abbildung 3: Satzfolgen, die einen formelhaften Textabschnitt im Korpus „Verträge aus Stadtverwaltungen und Hochschulen“ abbilden.

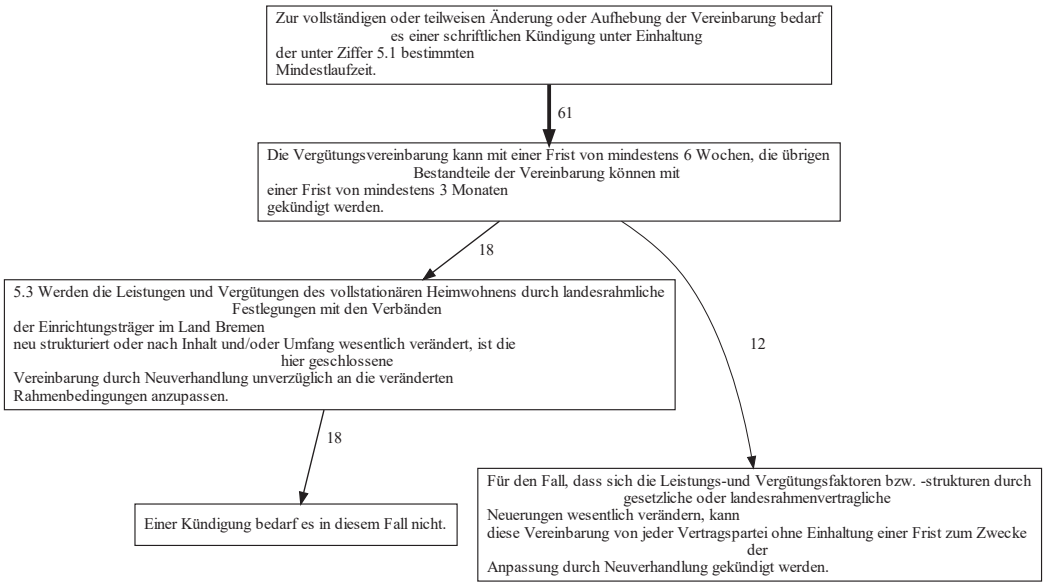

Abbildung 4: Variation in einem formelhaften Textabschnitt, aus dem Korpus „Verträge“.

4.2.2. Musterhafte Variation und individuelle Ausprägung

Wenn die individuellen Satzfolgen eines formelhaften Textabschnittes weiter betrachtet werden, wird erkennbar, dass die individuellen Ausprägungen klassifiziert werden können. Beispielsweise Variationen im Einsatz von Plural- und Singularform von Artikeln, Substantiven und Pronomina; Angaben zu Zeiträumen und Nummerierungen und weitere individuellen Anpassungen. Des Weiteren werden im Korpus „Verträge aus Stadtverwaltungen und Hochschulen“ aber auch Fehler aus der Texterkennung und -extraktion sichtbar. Dies ist ein weitere Besonderheit des vorgestellten Ansatzes, da formelhafte Textabschnitte auch bei fehlerhaften Sätzen, identifiziert werden können.

4.2.2.1. Individuelle Ausprägungen von formelhaften Textabschnitten im Korpus „Urteile aus dem Strafrecht“

Abbildung 5 und 6 zeigen die musterhaften Variationen, die in den beiden formelhaften Textabschnitten im Korpus vorgekommen sind. Die Knoten im Graph sind indexiert.

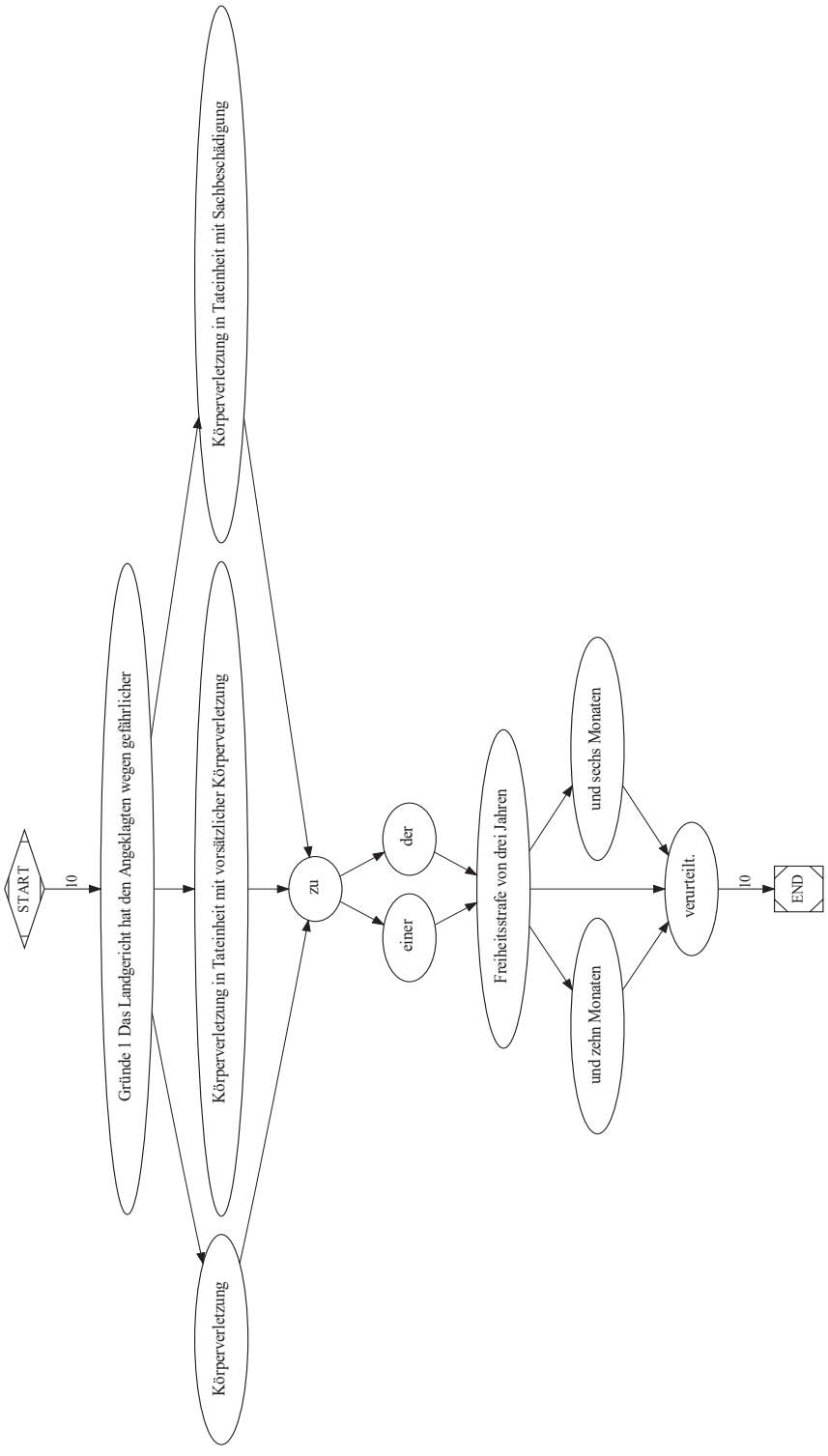

Abbildung 5: Wortgraph mit indiv. Ausprägungen

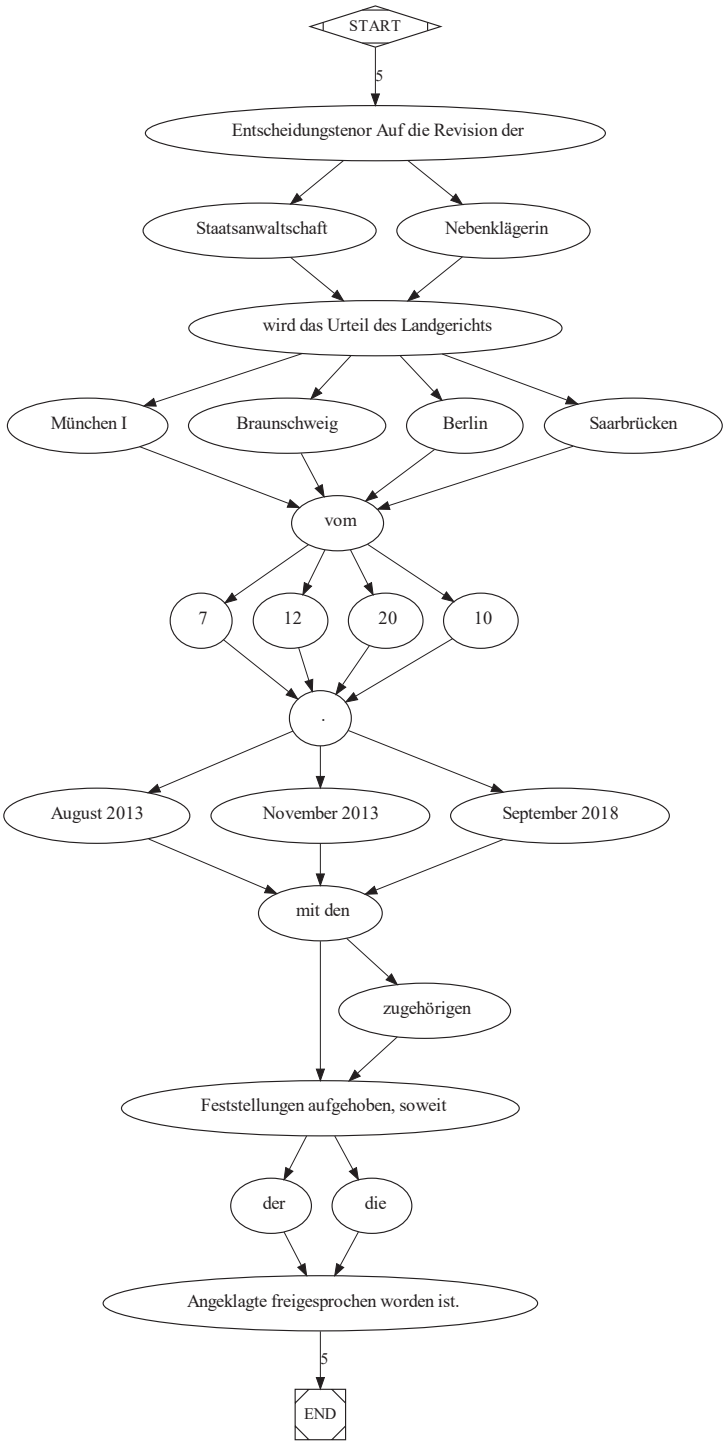

Abbildung 6: Wortgraph mit indiv. Ausprägungen

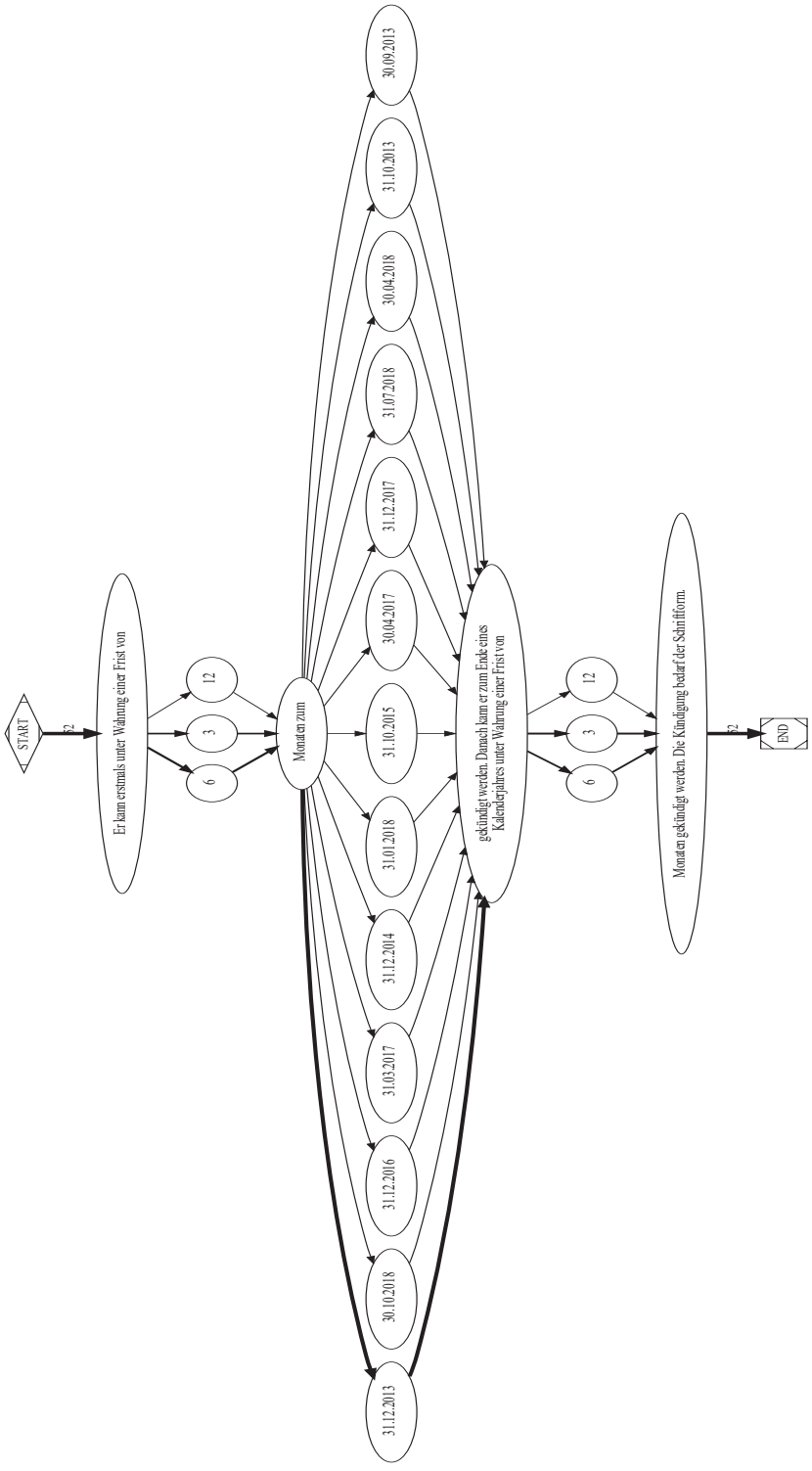

Abbildung 7: Wortgraph mit indiv. Ausprägungen

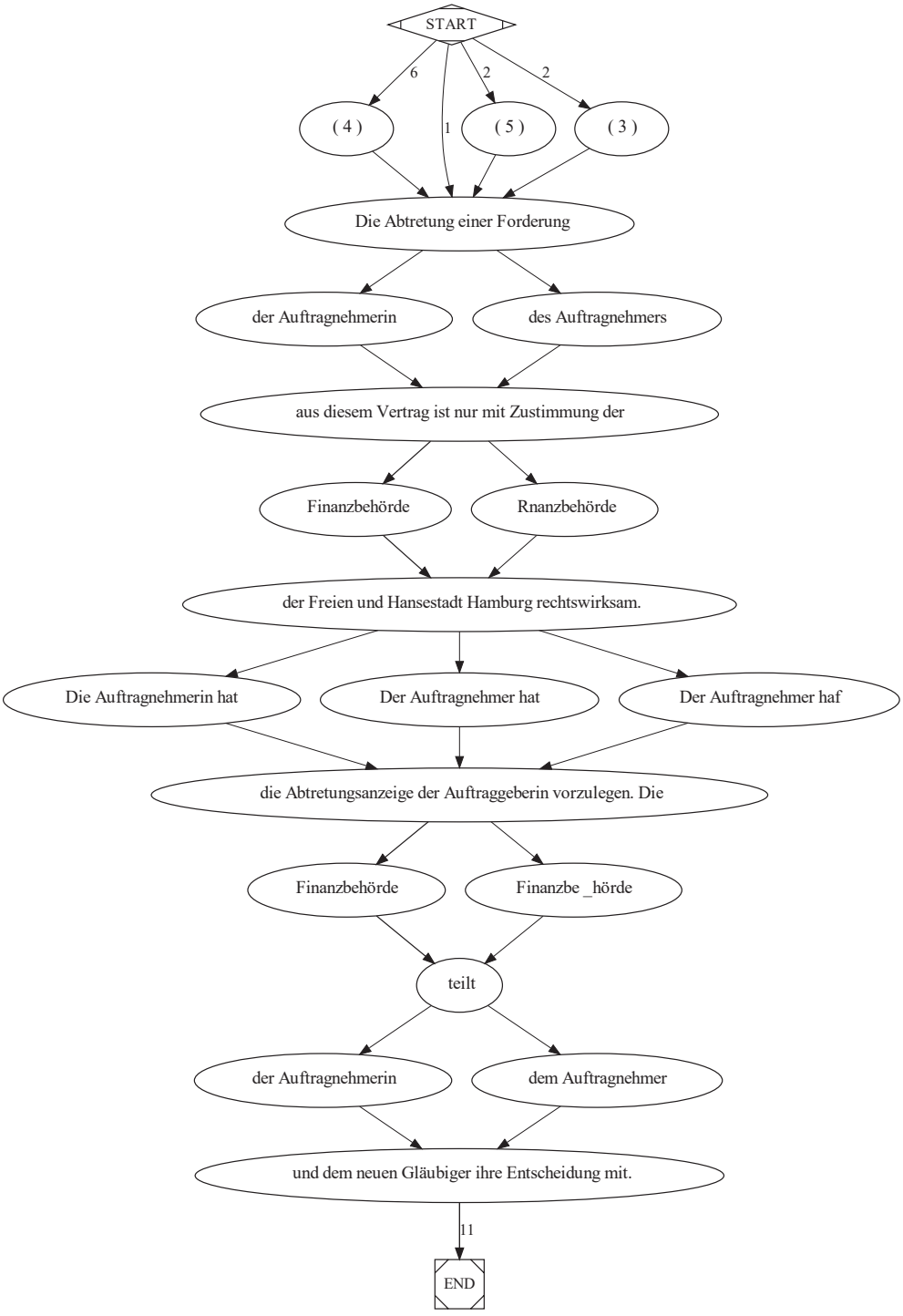

Abbildung 8: Wortgraph mit indiv. Ausprägungen

4.2.2.2. Individuelle Ausprägungen von formelhaften Textabschnitten im Korpus „Verträge aus Stadtverwaltungen und Hochschulen“

Die Abbildung 7, 8 und 9 zeigen die musterhaften Variationen, die in drei formelhaften Textabschnitten im Korpus „Verträge aus Stadtverwaltungen und Hochschulen“ vorgekommen sind. Die Knoten im Graph sind indexiert.

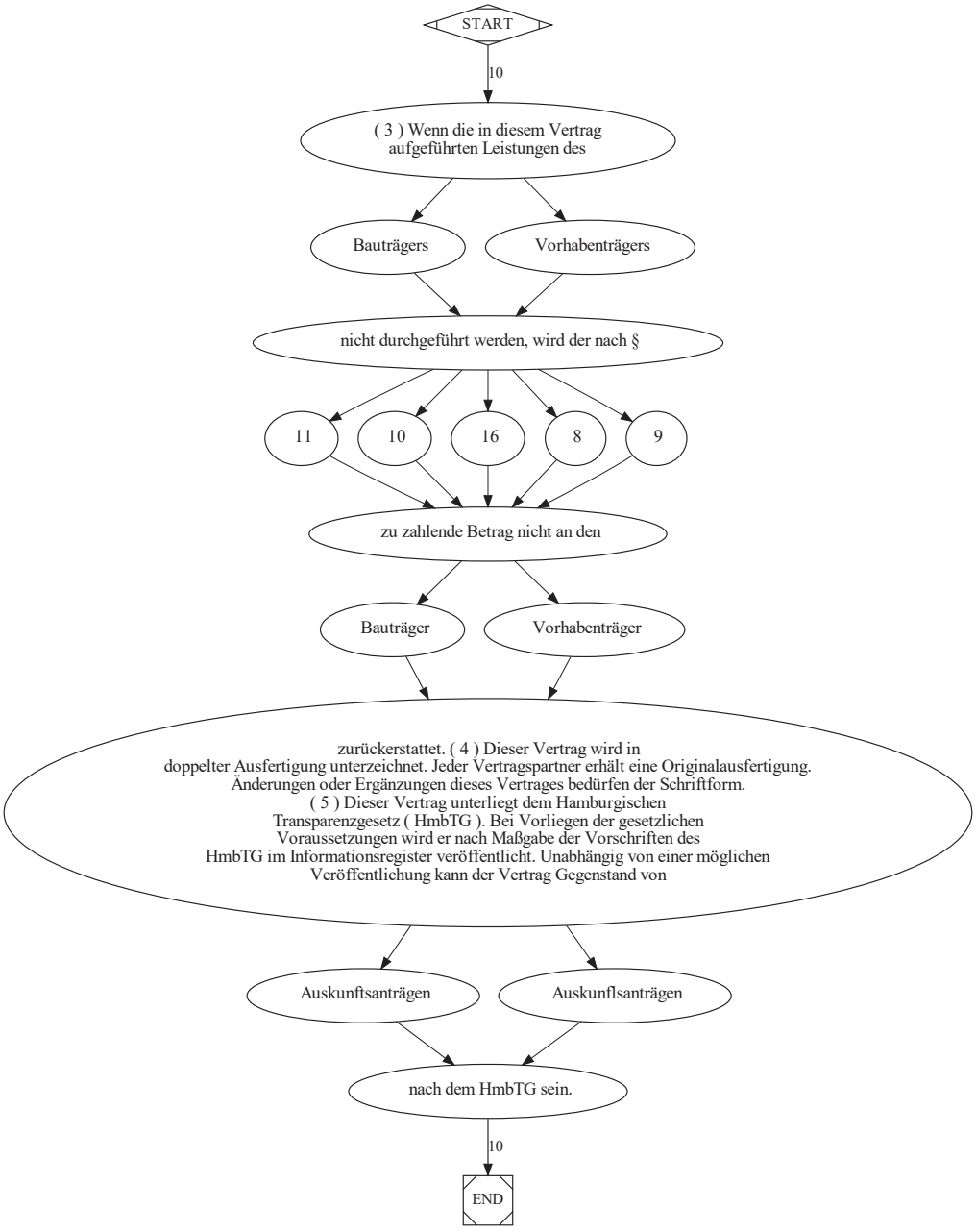

Abbildung 9: Wortgraph mit indiv. Ausprägungen

4.2.3. Erste Ergebnisse

Tabelle 1 zeigt die Ergebnisse für den aktuellen Stand des entwickelten Verfahrens. Die Satzcluster werden für beide Korpora gleich gebildet: Sätze werden als ähnlich gesehen, die bei der Übereinstimmung von Trigrammen auf Buchstabenebene mindestens einen Jaccard-Koeffizienten von 0.5 haben. Für die Identifizierung der gemeinsamen längsten Teilsequenz (siehe Abschnitt 4.1.) muss eine Sequenz im Korpus „Urteile aus dem Strafrecht“ achtmal vorkommen um weiter berücksichtigt zu werden und im Korpus „Verträge aus Stadtverwaltungen und Hochschulen“ mindestens zehnmal. Für die Berechnung der Anzahl von formelhaften Textabschnitten im Korpus gilt die Bedingung, dass jeder Satz in einem Dokument nur einer Sequenz zugeordnet werden darf.

	Korpus Urteile	**Korpus Verträge**
Anzahl Dokumente	4.250	2.129
Anzahl Sätze	308.781	192.712
Indiv. Sätze	197.713	125.038
Anzahl Satzcluster	172.435	95.361
Anzahl Sequenzmuster	899	2.308
Indiv. Formelhafte Textabschnitte (Seq. in Graphen)	317	411
Anzahl formelhafte Textabschnitte im Korpus (alle möglichen Sequenzmuster)	7.226	9.835
Anzahl formelhafte Textabschnitte im Korpus (mit Bedingung: jeder Satz wird nur einer Sequenz zugeordnet)	514	1.772
Indiv. Ausprägungen von formelhaften Textabschnitten im Korpus	78.279	691.296

Tabelle 1: Ergebnisse für Einzelbereiche des Verfahrens, getrennt nach Korpus.

5. Einsatzbereiche und Potential

Das vorgestellte Verfahren für die Identifizierung von formelhaften Textabschnitten, kann für die Analyse bestehender umfangreicher Korpora mit Texten aus den Rechtswissenschaften eingesetzt werden. Hierbei sehen wir Potential, da Verträge, die jährlich neu gezeichnet werden müssen, zumeist wenige Aktualisierungen gegen über dem Vorjahr aufweisen, die jedoch mitunter gravierend sein können. Für die Unterstützung im Genehmigungsprozess ist es daher sinnvoll die formelhaften Abschnitte, sowie die aktuelle Ausprägung im vorliegenden Vertrag zu erkennen und aufzuzeigen. Einen weiteren Einsatzbereich sehen wir für die Generierung von Vertragsdokumenten. Genehmigte und häufig eingesetzte Textabschnitte können für Vertragstexte mit individuellen aktuellen Angaben ergänzt werden. Die vorgegebenen Positionen in diesen formelhaften Abschnitten sind bekannt und auch die Art der zugelassenen Angabe kann festgelegt werden, beispielsweise Angaben, die den Vertragsinhalt beschreiben. Somit können aufzusetzende Verträge formalisiert werden.