Wörtliche Übereinstimmungen und Übernahmen in frühneuhochdeutschen Rechtstexten

Erkennung und Auswertung

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät
der Universität zu Köln
im Fach Informationsverarbeitung

vorgelegt von
Almuth Bedenbender

Erster Referent: Prof. Dr. Manfred Thaller

Zweiter Referent: Prof. Dr. Reinhard Förtsch

Datum der Disputatio: 3. 2. 2016

Vorwort

Der vorliegende Text ist die für die Publikation überarbeitete Fassung meiner Dissertation, die im Wintersemester 2015/16 von der Philosophischen Fakultät der Universität zu Köln angenommen wurde.

Mein besonderer Dank gilt zunächst meinem Doktorvater, Herrn Professor Dr. Manfred Thaller, für die geduldige Betreuung meines Projekts und für langjährige Unterstützung. Herrn Professor Dr. Reinhard Förtsch danke ich herzlich für die Übernahme des Zweitgutachtens.

Ohne die Anregungen, die ich in der Arbeit an den Forschungsstellen *Deutsches Rechtswörterbuch* und *Frühneuhochdeutsches Wörterbuch* sowie insbesondere im Projekt *DRQEdit* erfahren habe, und ohne die mit diesen Tätigkeiten verbundene Beschäftigung mit Quellentexten und mit dem Frühneuhochdeutschen hätte ich diese Untersuchung nicht – oder jedenfalls nicht in dieser Weise – durchführen können. Dafür danke ich Herrn Professor Dr. Andreas Deutsch, Frau Professor Dr. Anja Lobenstein-Reichmann, Herrn Professor Dr. Oskar Reichmann und Herrn Dr. Heino Speer.

Perl hat als Basis für meine eigene Programmierung gedient. Weitere Programme sowie Softwarekomponenten, die für die hier beschriebenen Untersuchungsschritte genutzt werden konnten, werden im Haupttext beziehungsweise in den Anmerkungen genannt. Darüber hinaus ist auf verschiedene Programme und Module hinzuweisen, die die Gestaltung dieser Arbeit in der vorgelegten Form ermöglicht haben. Neben LaTeX (mit einer Reihe von dafür entwickelten Paketen) sind die *Perl*-Module *GD::SVG* und *SVG* sowie das Programm *Batik* zu nennen, die ich für die Generierung und Einbettung von Vektorgraphiken nutzen konnte. Für die Graphvisualisierung habe ich *Graphviz* eingesetzt. Den Entwicklern all dieser Programme und Komponenten bin ich sehr dankbar.

Schließlich: Für Ermutigung und für Einzelhinweise, die ich in der hier vorgelegten Textfassung berücksichtigen konnte, danke ich meinem Kollegen- und Freundeskreis sowie meiner Familie.

Heidelberg, im Januar 2018 Almuth Bedenbender

Inhaltsverzeichnis

Einleitung

Die vorliegende Arbeit ist aus einer praktischen Fragestellung erwachsen: Wie lassen sich textuelle Abhängigkeiten zwischen frühneuhochdeutschen Rechtstexten mit Hilfe der EDV feststellen und analysieren? Die hier untersuchten Texte weisen in vielen Fällen wörtliche Übereinstimmungen auf, die so ausgeprägt sind, dass ein Zufall ausgeschlossen werden kann und offenbar Traditionszusammenhänge – aufgrund direkter Abhängigkeit oder auch mit Verbindungen über Zwischenglieder – festgestellt werden können. Die Ermittlung weitgehend wortgleicher Passagen ist aber auch mit einer gründlichen Quellenkenntnis ohne technische Unterstützung nur in Einzelfällen zu leisten und insbesondere für Textpaare schwierig, die nur in kleineren Teilen Entsprechungen aufweisen. Dementsprechend finden sich in der wissenschaftlichen Literatur zwar verschiedentlich Hinweise auf solche Abhängigkeitsverhältnisse, dabei werden aber vor allem bekanntere Texte als Vorlagen genannt, und in nicht wenigen Fällen wird das Ausmaß der Übereinstimmung irreführend beschrieben.

Die Frage nach Traditionslinien zwischen Texten ist als solche natürlich nicht neu. Bis vor wenigen Jahren standen aber allenfalls Einzeltexte in maschinenlesbarer Form zur Verfügung, und schon deshalb wäre es gar nicht möglich gewesen, mithilfe der EDV ein größeres Korpus zu untersuchen, das nicht nur vorab als besonders wichtig eingestufte Texte umfasst, sondern gerade auch solche, die bisher nicht im Fokus des Interesses standen. Erst die Digitalisierungsaktivitäten der letzten Zeit haben hier zu einer Änderung geführt. Die vorliegende Arbeit basiert auf den bisher im Rahmen des DFG-Projekts *DRQEdit – Deutschsprachige Rechtsquellen in digitaler Edition* erfassten Texten.[1]

Neben der Quellengrundlage ist aber auch ein technisches Verfahren erforderlich, das die Erkennung von Übereinstimmungen in einem größeren Textbestand effizient ermöglicht. Diese Aufgabe ist nicht trivial, da schon die Anzahl der Vergleichsoperationen bei einem naiven Verfahren quadratisch von der Anzahl der zu vergleichenden Einheiten abhängt und damit sehr schnell eine Größe erreicht, die auch auf heutigen Rechnern Probleme bereitet.

Zudem müssen für die Verarbeitung des hier untersuchten Korpus besondere Herausforderungen bewältigt werden: Das Frühneuhochdeutsche ist durch eine starke Varianz gekennzeichnet, die nicht nur orthographische Muster betrifft, sondern in vielen Fällen auch auf lautlichen Unterschieden der verschiedenen Dialekte beruht. Und schon eine Abgrenzung von Wörtern und Sätzen und damit die Bildung etwas größerer Einheiten als Basis für einen Vergleich ist problematisch, weil sowohl die Setzung von Leerzeichen als auch die Interpunktion vielfach nicht den heutigen Schreibregeln entspricht und beim Vergleich zweier Texte keineswegs mit einem stets einheitlichen Gebrauch gerechnet werden kann. Außerdem können gerade

[1] Vgl. *http://drw-www.adw.uni-heidelberg.de/drqedit/* sowie unten Kapitel 1.2.

Rechtstexte in nicht unerheblichem Maße feststehende Formulierungen aufweisen, und jedenfalls im hier untersuchten Textkorpus gibt es auch zahlreiche Übereinstimmungen größerer Länge, die auf Formulare – also Mustertexte – zurückzuführen sind. Soweit das Ziel einer Untersuchung von Übereinstimmungen nicht die Ermittlung etablierter Formulierungsmuster ist, sondern vielmehr die Erkennung von Abhängigkeitsbeziehungen zwischen bestimmten Texten beziehungsweise Textpassagen, kommt hier also einer Filterung der ermittelten Entsprechungen eine besondere Bedeutung zu.

Der für diese Arbeit gewählte Titel ist von der am Ende des letzten Absatzes beschriebenen Schwierigkeit beeinflusst, insbesondere aber von der Überlegung, dass mit dem hier entwickelten Instrumentarium zwar eine Erkennung von wörtlichen Übereinstimmungen – soweit sie gewissen Kriterien genügen – möglich ist, aber nur begrenzt eine Erkennung von wörtlichen Übernahmen. Bei etwas längeren Stücken gleichen Wortlauts kann zwar ausgeschlossen werden, dass es sich um zufällige Übereinstimmungen oder um sprachlich fest gefügte Muster handelt, umgekehrt kann aber für nur kurze gleiche Wortfolgen selbst bei hoher Vorkommenshäufigkeit nicht unbedingt sicher gesagt werden, dass die betreffenden Textstellen nichts miteinander zu tun haben.[2]

Zudem sind Aussagen über direkte Beziehungen zwischen zwei Texten oft problematisch. Da nur in einem kleinen Teil der Fälle explizit auf Ausgangstexte Bezug genommen wird, lässt sich zwar vielfach feststellen, dass es Übereinstimmungen gibt, die offenbar nicht nur zufällig beziehungsweise durch etablierte sprachliche Muster erklärbar sind, sondern auf Abhängigkeitsverhältnissen beruhen, daraus ergibt sich aber ohne weitere Anhaltspunkte keineswegs automatisch, dass einer der beiden Texte direkt auf dem anderen basiert. Vielmehr ist durchaus auch mit komplexeren Verwandtschaftsverhältnissen zu rechnen, etwa der gemeinsamen Verwendung eines dritten Textes oder auch einer Abhängigkeit über Zwischenglieder.

Die vorliegende Untersuchung gliedert sich in vier Teile. In Teil 1 geht es um geisteswissenschaftliche Aspekte. Er geht auf verschiedene Themenbereiche ein, bei denen wörtliche Übereinstimmungen und Übernahmen eine Rolle spielen, stellt das untersuchte Textkorpus und den damit verbundenen rechtshistorischen Untersuchungsgegenstand vor und gibt einen Überblick über Entwicklungslinien und Besonderheiten der frühneuhochdeutschen Schreibung und Lautung sowie über die zugehörigen dialektalen Großräume.

[2] Das lässt sich – ohne Bezug zum hier ausgewerteten Korpus – leicht erläutern am Beispiel geflügelter Worte, die sich zwar prinzipiell auf einen Verfasser zurückführen lassen, aber auch ohne Kenntnis des Ursprungs als Redensart gebraucht werden können. Das schließt jedoch für den Einzelfall nicht aus, dass es sich um ein bewusstes Zitat handelt.

Teil 2 beschäftigt sich mit dem technischen Hintergrund und verwandten Arbeitsgebieten aus dem Bereich der *Digital Humanities*. Er stellt Verfahren aus dem Bereich der Stringalgorithmik, der Bioinformatik und des *Information Retrieval* vor, die für den Vergleich von Strings und Texten genutzt werden können, und geht auf die Forschungsbereiche der Plagiatserkennung, der Untersuchung von *Text Reuse* sowie der automatischen Kollationierung ein.

Teil 3 nimmt den größten Raum ein und steht auch inhaltlich im Zentrum. Er erläutert das für diese Untersuchung entwickelte Verfahren, das eine effiziente Erkennung von wörtlichen Übereinstimmungen auch bei recht ausgeprägten Schreibungsunterschieden ermöglicht, geht dabei auch auf Probleme ein, die sich bei den einzelnen Verarbeitungsschritten stellen, beschreibt die dafür gefundenen Lösungen und stellt verschiedene Möglichkeiten vor, die ermittelten Übereinstimmungen zu analysieren, zu visualisieren und für den detaillierten Vergleich von Texten zu nutzen.

Teil 4 schließlich zeigt exemplarisch für einige Texte, welche Erkenntnisse sich mit dem entwickelten Instrumentarium erzielen lassen. Zu den untersuchten Texten gibt es zwar kurze Darstellungen in der wissenschaftlichen Literatur, die auch auf Abhängigkeitsverhältnisse eingehen, die darin aufgestellten Behauptungen sind aber, wie hier gezeigt wird, vielfach korrekturbedürftig. Die Untersuchungsergebnisse werden in unterschiedlichen Formen präsentiert, die von einem graphischen Gesamtüberblick über die festgestellten Übereinstimmungen mit anderen Texten des Korpus bis zu einem detaillierten synoptischen Vergleich reichen.

Im Haupttext werden die Quellen teils mit einem neuhochdeutschen oder lateinischen, nicht unbedingt auf dem Original basierenden Namen[3], teils mit einer Kurzfassung des Titels der zugrunde gelegten Ausgabe bezeichnet. Hier Einheitlichkeit erreichen zu wollen, wäre problematisch, weil es zum einen für die wichtigeren Quellen häufig bereits etablierte Bezeichnungen gibt und zudem die Bezugnahme auf eine bestimmte Ausgabe nicht in jedem Fall angemessen ist und weil sich zum anderen manche Werktitel nicht einfach durch Anwendung der heutigen Orthographie in eine neuhochdeutsche Form bringen lassen. Die für Quellen- und Literaturangaben in den Anmerkungen und den Abbildungen und Tabellen verwendeten Zitierweisen werden in den Vorbemerkungen zum Literaturverzeichnis erläutert. Das Glossar enthält Hinweise zu technischen Termini und Abkürzungen, die im Haupttext als Grundlagenwissen vorausgesetzt werden, aber für geisteswissenschaftlich ausgerichtete Leser nicht unbedingt unmittelbar verständlich sind.

[3] Die Bezeichnungen für Normtexte bestehen vielfach aus einer Angabe der Textsorte mit einem attributiven Zusatz zur Benennung des Geltungsraums. Da diese Bezeichnungen nur in Einzelfällen dem Wortlaut des Originaltitels folgen und sie sich in der Regel problemlos als beschreibende Textbenennung lesen lassen, sind sie recte gesetzt, sofern diese Art der Darstellung keinen Anlass zu Missverständnissen bietet.

1 Fragestellung und Untersuchungsgegenstand

Teil 1 führt von einem geisteswissenchaftlichen Blickwinkel aus in die Fragestellung und den Gegenstand dieser Untersuchung ein. Dabei dient Kapitel 1.1 der Orientierung in einem größeren thematischen Rahmen und Kapitel 1.2 der Vorstellung des ausgewerteten Textkorpus. Kapitel 1.3 bietet eine kurze Überblicksbeschreibung des Frühneuhochdeutschen im Hinblick auf Schreibung und Lautung und geht dabei insbesondere auf die Schreibungsvarianz ein.

1.1 Geisteswissenschaftliche Anknüpfungspunkte

Kapitel 1.1 soll verschiedene geisteswissenschaftliche Forschungsgebiete betrachten, bei denen es um die Beziehungen zwischen zumindest teilweise durch wörtliche Übernahmen miteinander verbundenen Texten und Textfassungen geht. Das Spektrum reicht dabei vom Vergleich mehr oder weniger eng verwandter Textfassungen im Rahmen editorischer Arbeit über die Klassifikation intertextueller Beziehungen bis hin zu Untersuchungen zur Entwicklung des Plagiatsbegriffs und zum Umgang mit Textvorlagen in früheren Zeiten.

1.1.1 Textvergleich in wissenschaftlichen Editionen

In diesem Unterkapitel geht es um die Frage, welche Rolle der Vergleich von Texten beziehungsweise Textfassungen oder Textzeugen im Rahmen wissenschaftlicher Editionen spielt und inwieweit das in der vorliegenden Untersuchung vorgestellte Verfahren dafür verwendbar ist. Das kann freilich nur in sehr oberflächlicher Form geschehen und überhaupt nur wenige Aspekte thematisieren – die editorischen Fragen und die üblicherweise angewandten Verfahren unterscheiden sich je nach Fachgebiet und wissenschaftlicher Schule erheblich.[4]

Kernaufgabe der kritischen Edition eines Textes[5] ist traditionellerweise die Herausgabe dieses Textes in meist einer einzigen Fassung[6], der in einer bestimmten Weise eine besondere Relevanz zugesprochen wird, sowie die Verzeichnung von Abweichungen in anderen überlieferten Fassungen oder auch – sofern sie nicht schon in

[4] Vgl. die umfassende Darstellung von Patrick Sahle (Sahle 2013, insbesondere Bd. 1 zur vordigitalen editorischen Theorie und Praxis).

[5] Dieser Begriff wird hier bewusst ohne genaue inhaltliche Füllung verwendet, da es nur um eine allgemeine Beschreibung geht. Vgl. zu den verschiedenen Verwendungsweisen von *Text* Sahle 2013, Bd. 3, S. 1 ff. Wie ebd. S. 76 ff. dargelegt wird, hängt die Aufgabe einer Edition wesentlich davon ab, welche Vorstellung vom Text zugrunde liegt.

[6] *Fassung* kann hier im Sinne von Sahle verstanden werden („Orthografische oder grafematische Unterschiede konstituieren, wenn sie wahrgenommen werden, verschiedene Fassungen“, Sahle 2013, Bd. 3, S. 46), also abweichend von Definitionen, nach denen sich Fassungen im Wortlaut unterscheiden und diese Unterschiede zudem intendiert sind, also nicht auf Textverderbnis beruhen (vgl. Schiewer 2005, S. 37–40).

die präsentierte Version eingeflossen sind – von Konjekturen des Herausgebers zur Verbesserung des edierten Textes.

Was die dabei zugrunde liegenden Leitvorstellungen und angewandten Praktiken sind, kann recht unterschiedlich sein. Insbesondere ist die Problemlage für Texte, die von Anfang an in gedruckter Form verbreitet wurden, natürlich eine andere als bei solchen mit einer zum Teil lang anhaltenden handschriftlichen Tradierung, da durch den Druck ein innerhalb einer Auflage im Großen und Ganzen einheitlicher Text vorliegt.[7] Auch für Werkausgaben zu neuzeitlichen Autoren ist prinzipiell mit Textvarianten zu rechnen (aufgrund von Textabweichungen zwischen Autograph und Erstdruck, aufgrund späterer Überarbeitungen durch den Verfasser sowie – bei textgenetischen Editionen – aufgrund von Änderungen im Laufe der Textentstehung), wesentlich größere Unterschiede im Wortlaut – ganz abgesehen von den dafür verwendeten Schreibungen – gibt es aber in aller Regel bei älteren Texten mit breiter handschriftlicher Überlieferung.

Nach welchen Kriterien jeweils die Auswahl – oder auch Konstruktion – des edierten Textes und der verzeichneten Varianten erfolgen kann, braucht im Zusammenhang dieser Untersuchung nicht dargestellt zu werden. Hier soll vielmehr auf drei Aufgaben hingewiesen werden, die sich bei der Erarbeitung einer kritischen Edition stellen beziehungsweise stellen können und die eine gewisse Verwandtschaft zum in dieser Arbeit untersuchten Problem aufweisen, nämlich auf die Ermittlung der Handschriften und Drucke, die für die Edition berücksichtigt werden, den Vergleich dieser Textfassungen sowie die Ermittlung von Quellen etwa von Zitaten.

Zunächst einmal ist die Voraussetzung für einen Vergleich verschiedener Textfassungen, dass diese Fassungen überhaupt bekannt sind. Während eine entsprechende Recherche für Drucke mit bekanntem Autor und einheitlichem Werktitel heutzutage aufgrund übergreifender Verzeichnisse wohl in vielen Fällen wenig problematisch sein dürfte, gestaltet sich die Lage schwieriger, wenn nicht in dieser Weise verlässlich nach Manuskripten oder Ausgaben mit bestimmten Metadaten gesucht werden kann, insbesondere also bei anonymen Werken ohne einen festen Titel.

Das in dieser Untersuchung vorgestellte Verfahren könnte theoretisch Unterstützung bieten für die Suche nach Handschriften oder Drucken eines Textes, die über die Sekundärliteratur beziehungsweise die bibliographischen Hilfsmittel nicht gefunden werden können. Dies ist allerdings jedenfalls auf dem gegenwärtigen Stand unrealistisch. Zum einen würde es voraussetzen, dass einigermaßen fehlerfreie maschinenlesbare Fassungen dieser Texte beziehungsweise Textzeugen vorliegen, zum anderen wäre es für diese Aufgabe wohl vor allem dann von Interesse, wenn wirklich große Datenbestände ausgewertet werden könnten, so dass auch

[7] Vgl. zum Problem der Pressvarianten zum Beispiel Plachta 2006, S. 64–66; ebd. S. 9 f. wird auf die „kategorial unterschiedliche editorische Methodologie und Praxis bei der Edition antiker oder mittelalterlicher Texte“ im Vergleich zu neueren Werken hingewiesen.

das berücksichtigt werden könnte, was bisher noch nicht einmal bibliographisch erschlossen worden ist.

Im Zentrum der Arbeit im Rahmen einer kritischen Edition steht der Vergleich der für die Edition berücksichtigten Handschriften und Drucke, gegebenenfalls die Ermittlung von Abhängigkeitsverhältnissen zwischen ihnen (die Bildung eines Stemmas), die Feststellung von Textvarianten sowie – jedenfalls bei einer gedruckten Edition, die nicht auf eine gleichrangige Präsentation verschiedener Fassungen zielt[8] – die Auswahl oder auch Erstellung einer Textfassung, die als Haupttext abgedruckt wird, sowie die Verzeichnung von Textvarianten im Apparat.

Insbesondere bei Texten der Antike und des Mittelalters, die eine breite und variantenreiche Überlieferung haben, aber keine erhaltene Originalversion, sind die Einordnung der erhaltenen Versionen in Traditionslinien und die Sichtung und Klassifikation der Varianten wesentliche Voraussetzungen für eine adäquate Auswahl des edierten Textes und der im Apparat verzeichneten anderen Lesarten.

Für diesen aufwendigen Arbeitsschritt – und im Anschluss daran für die Aufbereitung von Haupttext und Apparat – legt sich der Einsatz von Softwaretools nahe, und dementsprechend sind verschiedene Programme zum automatischen Kollationieren entwickelt worden, von denen einige unten in Kapitel 2.6 vorgestellt werden sollen. Schon hier sei allerdings angemerkt, dass sich für die automatische Kollationierung verschiedene Probleme auf unterschiedlichen Ebenen stellen und – soweit das Ziel nicht die vollständige Dokumentation auch kleinster Schreibungsunterschiede ist – eine Entwicklung editorischer Klassifikationskriterien und gegebenenfalls die Überarbeitung von Kollationierungsergebnissen erforderlich ist.

Das in dieser Untersuchung vorgestellte Verfahren zur Ermittlung von Textübereinstimmungen ist zwar prinzipiell auch dafür geeignet, in einer ähnlichen Weise wie bei einer Kollationierung einen Gesamtvergleich zweier Texte durchzuführen und eine synoptische Ansicht mit Hervorhebung der ermittelten Übereinstimmungen beziehungsweise Unterschiede zu erstellen, es ist allerdings nicht für die Kollationierung und die Erstellung von Apparaten optimiert.

Umgekehrt ist aber auch festzustellen, dass Kollationierungstools für einen Textvergleich, wie er im Rahmen dieser Untersuchung durchgeführt wird, kaum geeignet sind. Sie sind ausgerichtet auf die Ermittlung von Unterschieden zwischen Textfassungen oder Textzeugen, bei denen vorausgesetzt wird, dass sie jedenfalls in erheblichem Umfang miteinander übereinstimmen. Das ist eine deutlich andere Zielsetzung als das hier untersuchte Problem, Übereinstimmungen von oft auch nur geringem Umfang in einer größeren Zahl von Texten zu finden.[9]

[8] Für eine solche gleichrangige Präsentation ist die synoptische Wiedergabe etabliert.

[9] Vgl. zum Umfang des hier ausgewerteten Textkorpus unten S. 36. Die Schwierigkeiten bei der Verwendung von Kollationierungsprogrammen für den Vergleich von Texten aus dem hier untersuchten Korpus werden unten in Kapitel 2.6 näher beschrieben.

Während das Zusammenstellen und Weiterverarbeiten von Textvarianten also nicht beziehungsweise nur entfernt der hier untersuchten Fragestellung entspricht (und für diese Aufgaben sicherlich geeignetere Hilfsmittel vorhanden sind), gibt es doch einen recht unmittelbaren Bezug zur editorischen Arbeit, jedenfalls dann, wenn diese auch die Texterschließung im Hinblick auf den Nachweis von Zitaten und Quellen einschließt.[10]

Der Quellennachweis für als solche erkennbare Zitate, deren Herkunft im edierten Text nicht genau angegeben ist, dürfte heutzutage dank der Möglichkeiten, die Suchmaschinen bieten, und dank der Digitalisierung großer Buchbestände samt Erschließung über – freilich oft stark fehlerhafte – per OCR erstellte maschinenlesbare Volltexte in vielen Fällen weitaus weniger aufwendig sein als früher, da sich durch die Kombination signifikanter Wörter oder Wortfolgen entsprechende Textstellen in anderen Werken oft leicht eruieren lassen.

Wesentlich schwieriger verhält es sich allerdings, wenn aus dem Text nicht erkennbar ist, dass bestimmte Textstücke auf Vorlagen basieren. Es legt sich zwar wohl in vielen Fällen ein Vergleich mit bestimmten inhaltlich verwandten Texten nahe. Soweit Parallelen aber nicht an Stellen zu finden sind, die sich aus dem Sachzusammenhang und der Textstruktur ergeben, ist ein Vergleich ohne technische Unterstützung ausgesprochen mühselig und sicherlich allenfalls in kleinem Umfang zu leisten. Und auch das eben beschriebene Rechercheverfahren für den Nachweis von Zitaten kann hier nur begrenzt Abhilfe schaffen. Natürlich lässt es sich prinzipiell auch für andere Textstücke anwenden, der Aufwand ist aber um ein Vielfaches größer, wenn einfach alles überprüft wird, und zudem ist selbst für die Suche nach weitgehend wörtlichen Übernahmen keineswegs klar, welche Wörter für die Recherche auszuwählen sind, da die Abgrenzung dieser Übernahmen nicht erkennbar ist.

Dementsprechend ist davon auszugehen, dass ein Vergleichsverfahren wie das hier vorgestellte für die Edition von Texten, die auch im Hinblick auf Abhängigkeiten untersucht werden sollen, von Interesse ist – sowohl als Arbeitserleichterung für den Vergleich mit ausgewählten Texten, der andernfalls von Hand durchgeführt würde, als auch für die Erkennung von Übereinstimmungen mit Texten beziehungsweise Stellen, die sonst nicht berücksichtigt würden. Voraussetzung ist freilich, dass die Vergleichstexte in maschinenlesbarer Form vorliegen, und Kapazitätsprobleme machen es erforderlich, das Referenzkorpus einigermaßen gezielt zusammenzustellen, die dabei erreichte Größenordnung übersteigt aber bei weitem das, was durch lesenden Vergleich untersucht werden könnte.

In diesem Zusammenhang ist auf die im Bereich der Edition historischer Texte relativ verbreitete Praxis hinzuweisen, übernommene Textstücke durch Petitdruck

[10] Vgl. Plachta 2006, S. 123–125 mit Hinweisen auf entsprechende Forderungen im literaturwissenschaftlichen Bereich sowie Sahle 2013, Bd. 1, S. 55 zu geschichtswissenschaftlichen Editionen.

im Editionstext selbst kenntlich zu machen.[11] Dabei geht es allerdings – bei der Edition von Urkunden – wohl üblicherweise nur um die Abhängigkeit von Vorurkunden für denselben Empfängerkreis, insbesondere bei einer Rechtsbestätigung.[12] Eine solche Beschränkung ist aber auch kritisiert worden, und es kann hierzu je nach Fragestellung unterschiedliche Ansichten geben.[13] Hier mag zum Tragen kommen, dass eine digitale Edition (im Sinne Patrick Sahles[14]) nicht dem Zwang einer gedruckten Edition unterworfen ist, den zu edierenden Text in einer einzigen Form – oder allenfalls in einer kleinen Zahl synoptisch nebeneinander gestellter Fassungen – darzustellen, sondern vielmehr Material bereitstellen kann, das je nach Benutzerwunsch in unterschiedlicher Weise präsentiert werden kann. Entsprechend dieser Konzeption scheint es also durchaus plausibel, auch die Verdeutlichung von nach bestimmten Kriterien ausgewählten Übereinstimmungen mit anderen Texten als Darstellungsoption zu integrieren.

Während für Editionen aus dem Bereich der Rechtsgeschichte der Frühen Neuzeit der Vergleich verschiedener Überlieferungsvarianten wohl nur eine eher geringe Rolle spielt, ist in diesem Bereich die synoptische Darstellung mehrerer durch Textübernahmen verbundener Texte oder die zusammenfassende Edition eines Haupttextes und eines oder mehrerer davon abhängiger oder ihm als Vorlagen zugrunde liegender Texte, etwa bei Gesetzesrevisionen, eine schon seit langem gängige Praxis.[15] Dies zeigt ebenso wie die eben beschriebene editorische Verwendung des Petitdrucks deutlich, dass eine solche vergleichende Präsentation gerade für Texte, die der Setzung beziehungsweise Dokumentation von Rechtsverhältnissen dienen, von Interesse ist, da solche Texte häufig an Vorgängertexte anknüpfen.

In diesem Unterkapitel ging es um Textvergleich im Rahmen der editorischen Arbeit. Abschließend sei aber angemerkt, dass sich ein automatisierter Vergleich gerade

[11] Vgl. Sahle 2013, Bd. 1, S. 55 (Literaturhinweise in Anm. 148).

[12] Genaue Aussagen hierzu wären wohl nur im Rahmen umfassender Studien zu editorischen Praktiken in verschiedenen Großprojekten zu machen. Nach Sahle 2013, Bd. 1, S. 55 wurde der Petitdruck zur typographischen Kenntlichmachung von Abhängigkeitsverhältnissen von den *Monumenta Germaniae Historica (MGH)* eingeführt und dann in anderen Editionen übernommen. Ein Überblick über Aussagen zum Petitdruck in den MGH lässt sich – jedenfalls für die Bände mit editorischen Hinweisen in deutscher Sprache – über die Volltextsuche nach „Petitdruck“ oder auch „Petitsatz“ in der Online-Version (*http://www.dmgh.de/*) gewinnen.

[13] Vgl. die einleitenden Bemerkungen zur Edition der Urkunden Heinrichs II. und Arduins in den MGH von Harry Breßlau, in denen Theodor von Sickels Beschränkung des Petitdrucks auf „Entlehnungen aus einer Vorurkunde derselben Empfängergruppe“ als „dies neuerdings getadelte Verfahren“ beschrieben, aber ausdrücklich bestätigt wird, „denn die Unterscheidung zwischen gewöhnlichem und Petitdruck bei Urkunden hat in erster Reihe den Zwecken des Historikers, nicht denen des Diplomatikers zu dienen“ (MGH.DD H II, S. XI).

[14] Vgl. Sahle 2013, Bd. 2, insbesondere S. 148 ff.

[15] Für das hier untersuchte Korpus sind zum Beispiel die Edition von WürtLR. 1567 mit Verzeichnung der Abweichungen vom Vorgängertext WürtLR. 1555 in Reyscher (Hg.) 1831 oder die synoptische Edition von BambHalsGO. 1507 und PeinlGO. 1532 (1533) samt Apparat zu Abweichungen in BrandenbAnsbHalsGO. 1516 in Zoepfl (Hg.) 1883 zu nennen.

auch für solche Texte anbietet, die nicht kritisch ediert werden, aber immerhin in mehr oder weniger guter Qualität in maschinenlesbarer Form zugänglich sind. Die Massendigitalisierung, die seit einigen Jahren betrieben wird, eröffnet prinzipiell die Möglichkeit, Texte auch dann in dieser Weise auszuwerten, wenn kein Bedarf nach einer sorgfältigen editorischen Sichtung der Überlieferung besteht oder wenn dafür keine Mittel verfügbar sind. Die durch einen Vergleich ermittelten Übereinstimmungen bedürfen zwar häufig einer näheren Untersuchung, um daraus tragfähige Schlussfolgerungen über Abhängigkeitsverhältnisse ziehen zu können, sie sind aber in nicht wenigen Fällen aufgrund ihres Umfangs in bestimmten Abschnitten oder Gesamttexten so signifikant, dass auch rein quantitative Auswertungen ohne eine Betrachtung der Entsprechungen als Basis für eine erste Einordnung der betreffenden Texte dienen können. In diesem Sinne kann die automatisierte Erkennung von Textübereinstimmungen durchaus auch in den Kontext der heute gerne als *Distant Reading*[16] bezeichneten quantitativen Analyse großer Textmengen eingeordnet werden, auch wenn es dabei – jedenfalls nach der Intention der vorliegenden Untersuchung – letztlich um die jeweiligen Texte und nicht um die übergreifenden Muster geht.

1.1.2 Intertextualität

Der Begriff *Intertextualität*, der für die Überschrift dieses Unterkapitels gewählt wurde, wird in recht unterschiedlichen Weisen verwendet und ist deshalb in besonderem Maße erklärungsbedürftig. 1967 wurde das französische Pendant „intertextualité“ von Julia Kristeva geprägt und im Sinne des Poststrukturalismus als generelle Eigenschaft von Texten erläutert, da jeder Text ein „mosaïque de citations“ sei,[17] wobei dies natürlich nicht im wörtlichen Sinne zu verstehen ist, sondern vielmehr zum Ausdruck bringen soll, dass Texte in einem Abhängigkeitsverhältnis zu ihrem kulturellen Umfeld stehen.[18] Er hat sich aber in einem anderen Verständnis auch zur Beschreibung konkret fassbarer Bezüge zwischen Texten etabliert.[19]

In diesem Sinne soll *Intertextualität* auch hier verstanden werden. Die verschiedentlich vorgenommene Eingrenzung auf den literarischen Bereich[20] ist dabei

[16] Vgl. zum Beispiel Schubert 2015, S. 1 f. Der Begriff wurde von Franco Moretti geprägt und lässt sich (zumindest) bis auf das Jahr 2000 zurückführen, vgl. Moretti 2000, S. 56 ff., allerdings mit einer etwas anderen Akzentuierung, da Moretti ebd. S. 57 vom Verzicht auf Lektüre der literarischen Texte und von einem „patchwork of other people's research“ spricht.

[17] Kristeva 1967, S. 440 f.

[18] Vgl. zu Kristevas Position zum Beispiel Pfister 1985, S. 5–11, Tegtmeyer 1997, S. 51–56 und Berndt/Tonger-Erk 2013, S. 34–46.

[19] Diese beiden Verwendungsweisen werden in Broich 2000, S. 175 f. sowie Pfister 1985, S. 11 ff. (zusammengefasst ebd. S. 25) einander gegenübergestellt. Pfister 1985, S. 15 konstatiert, dass die Verwendung zur Bezeichnung von greifbaren Textbeziehungen jedenfalls dann dominiere, wenn es um „konkrete Textanalyse“ gehe.

sicherlich nicht zwingend[21] und wird hier nicht zugrunde gelegt; vielmehr soll es gerade auch um Beziehungen zwischen Gebrauchstexten gehen, wie sie im hier untersuchten Korpus vorkommen. Ein in diesem Sinne verwendbarer anderer Terminus steht anscheinend nicht zur Verfügung.[22]

Der Begriff, der verschiedentlich auch noch in anderer Hinsicht eingegrenzt wird,[23] soll hier ohne eine differenzierte literaturwissenschaftliche Konzeption der Zusammenfassung der verschiedenen Arten dienen, in denen Texte in Beziehung zu bestimmten anderen Texten stehen. Der Intention nach geht es dabei um objektive Zusammenhänge, nicht um solche, die in freier Interpretation vom Leser hergestellt werden.[24] Das bedeutet aber natürlich nicht, dass generell eine eindeutige Aussage über solche intertextuellen Beziehungen möglich ist; vielmehr ist gerade im Hinblick auf die in dieser Untersuchung betrachteten wörtlichen Übereinstimmungen in vielen Fällen mit mehr oder weniger großen Unsicherheiten zu rechnen, insbesondere wenn es um die Feststellung der direkten Verwendung eines bestimmten Textes bei der Abfassung eines bestimmten anderen geht.

Auch wenn in diesem Rahmen keine theoretische Erörterung oder systematische Entfaltung dessen, was in diesem Sinne als Intertextualität zu bezeichnen ist, erfolgen kann, soll doch anhand der von Gérard Genette entwickelten Typologie betrachtet werden, bei welchen Arten von Intertextualität (oder in Genettes Terminologie „Transtextualität“[25]) wörtliche Übereinstimmungen eine Rolle spielen.[26]

Genette geht von fünf Typen aus. Nur den ersten bezeichnet er als „Intertextualität“ und beschreibt ihn „als Beziehung der Kopräsenz zweier oder mehrerer Texte, d. h. in den meisten Fällen [...] als effektive Präsenz eines Textes in einem anderen“. Er ordnet Zitat, Plagiat und Anspielung diesem Typus zu, wobei er neben dem Zitat auch das Plagiat als „wörtliche Entlehnung“ charakterisiert und als Unterscheidungskriterium die Kennzeichnung durch Anführungszeichen nennt.[27] Ob Letzteres wirklich so gemeint ist, sei hier dahingestellt – bei der Verwendung

[20] So wird Intertextualität zum Beispiel als „Eigenschaft von insbes. literar. Texten, auf andere Texte bezogen zu sein“ beschrieben (Aczel 2008, S. 330). Vgl. auch Pfister 1985, S. 13 f. und 17.

[21] Vgl. verschiedene Beschreibungen, die den Begriff auf Texte allgemein und auch auf mündliche Äußerungen anwenden (Pfister 1985, S. 12) sowie Klein/Fix (Hg.) 1997 als Sammelband, in dem es insbesondere auch um linguistische Perspektiven und die Untersuchung von Sachtexten geht.

[22] Vgl. Pfister 1985, S. 1 sowie Broich 2000, S. 176 (ohne klare Ein- oder Ausgrenzung nichtliterarischer Texte).

[23] So ist zum Beispiel strittig, inwieweit der Rückgriff auf einzeltextübergreifende Formen als Intertextualität zu betrachten ist (vgl. Pfister 1985, S. 17–19) und ob die bewusste und erkennbare Bezugnahme für Intertextualität erforderlich ist (so das Verständnis in Broich 1985, vgl. ebd. S. 31).

[24] Vgl. zum entgegengesetzten poststrukturalistischen Verständnis Pfister 1985, S. 20 ff.

[25] Genette 1993, S. 9.

[26] In Berndt/Tonger-Erk 2013, S. 99–155 werden verschiedene Intertextualitätstypologien vorgestellt. Ebd. S. 99 wird erklärt, dass Genettes Entwurf „unseres Erachtens den Maßstab bildet“.

[27] Genette 1993, S. 10.

bekannter Formulierungen fehlt eine formale Kennzeichnung wohl häufig, ohne dass dabei vernünftigerweise von einer plagiatorischen Intention auszugehen ist. Ebenso soll an dieser Stelle offen bleiben, inwieweit ein Plagiat wirklich nur bei gleicher Wortwahl vorliegt und ob Genette auch Plagiate ganzer Texte im Blick hat oder ob er hier nur an kurze Übernahmen denkt, wie es die Zusammenstellung mit Zitat und Allusion nahelegt.[28] Jedenfalls ist festzuhalten, dass es nach dieser Beschreibung bei Plagiaten ebenso wie bei Zitaten um Übereinstimmungen geht, die in der vorliegenden Untersuchung als wörtliche Übernahmen bezeichnet werden.

Davon abzugrenzen sind intertextuelle Allusionen, jedenfalls wenn diese gerade durch Abweichungen in der Wortwahl bestimmt sind[29] (wobei es sich natürlich nicht allgemein um Umformulierungen handelt, sondern um punktuelle und nur angedeutete Bezugnahmen). Inwieweit auch ungekennzeichnete übereinstimmende Wortverbindungen bei Voraussetzung eines bestimmten Bildungshorizonts als Anspielung betrachtet werden können beziehungsweise sollten, muss hier nicht entschieden werden, zumal es im Rahmen dieser Arbeit nicht um literarische, sondern um Gebrauchstexte geht.

Paraphrasen werden von Genette nicht genannt, sind aber für intertextuelle Beziehungen insbesondere von Sachtexten wie etwa wissenschaftlicher Literatur natürlich wichtig und können wohl hier eingeordnet werden.[30] Da auch für sie kennzeichnend ist, dass sie den Wortlaut des Bezugstextes nicht oder nur sehr partiell übernehmen, liegt ihre Erkennung ebenso wie die von Allusionen jenseits der Ziele der vorliegenden Untersuchung.[31]

In diesem Zusammenhang sind außerdem noch explizite Verweise auf andere Texte beziehungsweise Textstellen zu nennen, wobei solche Verweise zwar häufig, aber natürlich nicht immer als Erläuterung zu Zitaten oder Paraphrasen auftreten.[32] Literaturverweise sind für einen der jeweiligen Sprache mächtigen und über relevante Abkürzungen informierten Leser als solche zwar in aller Regel gut zu erkennen, das gilt aber nicht unbedingt für das Verweisziel, das gerade in älterer Literatur häufig nicht mit einer Präzision bezeichnet wird, die eine leichte Ermittlung des bezeichneten Werks und im Idealfall auch der verwendeten Ausgabe

[28] Vgl. dazu auch das unten in Anm. 47 angeführte Zitat. Der Plagiatsbegriff wird unten in Unterkapitel 1.1.3 noch näher betrachtet. Die plagiatorische Übernahme eines ganzen Werks mit gewissen Änderungen wäre nach Genettes Systematik wohl als Transformation zu betrachten und damit auch oder ausschließlich dem Typus der „Hypertextualität“ zuzuordnen.

[29] Genette 1993, S. 10 spricht von einer „weniger wörtlichen Form“.

[30] Susanne Holthuis und Henning Tegtmeyer stellen in dieser Weise Zitat, Allusion und Paraphrase nebeneinander, vgl. Tegtmeyer 1997, S. 64 und 79.

[31] Vgl. zum Projekt *Tesserae*, in dem es um die automatisierte Ermittlung von Allusionen geht, unten S. 97–99.

[32] Inwieweit die Bezugnahme auf wissenschaftliche Literatur ohne nähere Angaben zum dort zu findenden Inhalt noch als Paraphrase bezeichnet werden kann, sei dahingestellt. Zumindest Literaturverzeichnisse (abgesehen vom Sonderfall der kommentierten Verzeichnisse) stellen jedenfalls Sammlungen von Verweisen ohne direkten Bezug zu Zitaten oder Paraphrasen dar.

ermöglichen würde. In Bezug auf die Entschlüsselung ungenauer bibliographischer Angaben, die nicht einem Zitat zugeordnet sind, ist freilich von einem Verfahren zur Feststellung wörtlicher Übereinstimmungen kaum eine Unterstützung zu erhoffen, da allenfalls bei einem nicht zu stark verkürzten Titel (und natürlich bei Zugehörigkeit des referenzierten Werks zum Korpus) eine Entsprechung zum Titel des Bezugstextes festgestellt werden könnte – und für eine solche Suche nach dem gemeinten Werk ist die Verwendung bibliographischer Rechercheinstrumente mit Sicherheit effektiver.[33]

Die Identifikation von Verweiszielen kann zwar bei einer Zitierweise, die nicht den heutigen Standards entspricht, unter Umständen nicht unerhebliche Probleme bereiten, die Existenz eines Verweises ist aber ein klares Kriterium dafür, dass ein Text Bezug auf einen anderen nimmt, so dass jedenfalls ein Anhaltspunkt für eine Überprüfung besteht. Ziel der vorliegenden Untersuchung ist hingegen insbesondere auch die Ermittlung von Abhängigkeiten, die nicht in dieser Weise kenntlich gemacht und auch für Leser mit einer guten Textkenntnis vielfach wohl nicht offensichtlich sind. Da in diesen Fällen gerade nicht explizit Bezug genommen wird, kann wohl hinterfragt werden, inwieweit man hier tatsächlich im Sinne des angeführten Zitats von Genette von einer *Präsenz* des Basistextes sprechen kann – passender ist vielleicht die Beschreibung als *Verwendung*. Aber auch wenn es bei dieser Verwendung nicht darum geht, einen Bezug herzustellen, liegt doch eine Beziehung – nämlich ein Rezeptionsverhältnis – vor. Dass wörtliche Übernahmen in früheren Zeiten nicht unüblich waren, soll in Unterkapitel 1.1.3 im Zusammenhang mit dem Plagiatsbegriff erörtert werden.

Von den übrigen von Genette vorgestellten Typen ist im Rahmen der Fragestellung der vorliegenden Untersuchung noch die „Hypertextualität“ von Interesse.[34] Genette beschreibt sie als Ableitungsverhältnis und stellt als Unterkategorien (einfache) Transformation und Nachahmung einander gegenüber.[35] Während bei nachahmenden Werken wörtliche Übereinstimmungen mit dem Basistext wohl dem schon

[33] Der Vollständigkeit halber sei auch die automatisch durchgeführte Zitationsanalyse erwähnt, die zum einen Regeln für die Erkennung erfordert, was überhaupt ein Literaturverweis ist, zum anderen Verfahren, um aus den darin enthaltenen Angaben möglichst präzise und gleichzeitig tolerant gegenüber unterschiedlichen Zitierweisen auf die gemeinten bibliographischen Einheiten zu schließen beziehungsweise – wenn es nur darum geht – den jeweiligen Autor zu ermitteln. Hier gilt also noch mehr als für eine Hilfestellung bei der Identifikation der referenzierten Werke, dass die Erkennung wörtlicher Übereinstimmungen ungeeignet ist und vielmehr der Berücksichtigung typischer Zitationsmuster sowie der Entwicklung von Kriterien für die Absicherung von Zuordnungen eine zentrale Bedeutung zukommt.

[34] Daneben gibt es noch die „Paratextualität“, die die Beziehung eines Textes zu dem bezeichnet, was ihm im Rahmen seiner Publikation beigegeben wird (Genette 1993, S. 11 nennt unter anderem Titel, Vorworte und Illustrationen), die „Metatextualität“, bei der es um kommentierende beziehungsweise kritische Äußerungen zu Texten geht (ebd. S. 13) sowie die „Architextualität“, das Verhältnis eines Textes zu seiner Gattung (ebd. S. 13 f.).

[35] Ebd. S. 15–18.

beschriebenen ersten Typus (also der „Intertextualität“ in Genettes Sprachgebrauch) zuzuordnen und eher punktuell sind, ist bei Transformationen mit wesentlich umfangreicheren wortgleichen Passagen zu rechnen. Schon „das Herausreißen einiger Seiten“ stellt für Genette eine solche Transformation dar.[36] Diesem Subtypus sind unter anderem die „Parodie“[37], Übersetzungen[38] sowie kürzende, erweiternde und substituierende Bearbeitungen eines Textes[39] zuzuordnen. Im Zusammenhang mit Erweiterungen führt Genette auch die „Mischung zweier (oder mehrerer Hypotexte)“ (die „Kontamination“)[40] an.

Ob Textbearbeitungen in diesem Sinne als hypertextuell zu klassifizieren sind, hängt wohl davon ab, ob der Ausgangsform und dem Resultat eine Art gemeinsame Textidentität zuzuschreiben ist oder nicht. Eine Entscheidung darüber ist sicherlich in nicht wenigen Fällen von einer theoretischen Konzeption abhängig. Hier soll zwar keine Text- oder Intertextualitätstheorie entwickelt oder dargelegt werden, die Betrachtung einiger Konstellationen mag aber dazu beitragen, das breite Spektrum von Beziehungen zwischen Texten beziehungsweise Textfassungen, die ein hohes Maß an wörtlicher Übereinstimmung aufweisen, besser in den Blick zu bekommen.

Generell können Texte einen komplexen Entstehungsprozess mit erheblichen Umarbeitungen haben. Es wäre wahrscheinlich wenig sinnvoll, das Verhältnis der publizierten Fassung zu einer noch nicht als abgeschlossen betrachteten Vorform als hypertextuell zu charakterisieren.[41]

Etwas anders mag der Fall liegen, wenn es um die Überarbeitung einer zuvor schon publizierten Textfassung durch den Autor selbst geht. Soweit es sich nicht um für einen konkreten Anlass vorgenommene Änderungen handelt[42], sondern um solche, die im Zusammenhang mit einer erneuten Publikation des Textes unter gleichem Titel erfolgen, spricht einiges dafür, auch dies als Weiterentwicklung eines einzigen Textes zu betrachten, und dementsprechend kann eine Edition auf dem Prinzip der Fassung letzter Hand basieren, also primär die letzte vom Autor stammende beziehungsweise autorisierte Version präsentieren und Abweichungen davon nur im Apparat dokumentieren. Nach einer anderen Konzeption kann bei einer Edition zum Beispiel die erste publizierte Fassung und damit der Beginn der Wirkungsgeschichte die Grundlage für den edierten Text bilden.[43] Eine solche

[36] Ebd. S. 16.
[37] Hierunter versteht Genette „die Bedeutungsänderung durch minimale Transformation eines Textes“ (ebd. S. 40).
[38] Vgl. ebd. S. 289–294.
[39] Vgl. ebd. S. 313–382.
[40] Ebd. S. 359 (Klammersetzung des ersten Zitats dort so).
[41] Genette geht speziell auf die Ausarbeitung von Skizzen und Entwürfen ein und erklärt, dass viele „Werke, die im Prinzip in keiner Weise als hypertextuell gelten können“, durch „Amplifizierung“ entstünden (ebd. S. 381).
[42] Genette weist im Zusammenhang mit Kürzungen auf nicht dokumentierte Bühnenfassungen von Theaterstücken hin (ebd. S. 318).
[43] Vgl. zu Fassungen früher und später Hand Plachta 2006, S. 77–80.

editorische Entscheidung für die Wiedergabe nur einer einzigen Textfassung in voller Form ist sicherlich nicht in jedem Fall als adäquat zu betrachten,[44] wenn sie aber im Rahmen einer Gesamtausgabe getroffen wird, ist dies wohl ein deutliches Zeichen, dass der Herausgeber davon ausgeht, damit kein nach der Autorintention separates Werk zu übergehen.

Die eben angestellten Überlegungen gehen von der Annahme voraus, dass ein Text in verschiedenen Fassungen von einem einzigen Autor stammt und dass diesem quasi die persönliche Entscheidungsgewalt über die Textgestalt auch im Hinblick auf Veränderungen zukommt – eine Vorstellung, wie sie zum Beispiel dem deutschen Urheberrechtsgesetz entspricht. Diese Voraussetzungen sind aber keineswegs allgemeingültig, vielmehr treffen sie auch auf viele Texte der heutigen Zeit nicht zu. So sind insbesondere Gebrauchstexte in vielen Fällen keine individuellen Schöpfungen, sondern werden von mehreren Beteiligten gemeinsam verfasst beziehungsweise redigiert, oft ohne dass diese namentlich in Erscheinung treten, und die Erstellung überarbeiteter Fassungen kann durch andere Personen erfolgen. Als Beispiele lassen sich etwa aus dem juristischen Bereich Gesetze und Vertragstexte anführen.

Zudem ist eine Unterscheidung zwischen Textfassungen, die vom ursprünglichen Verfasser stammen, und späteren Bearbeitungen zumindest dann nicht anwendbar, wenn die originalen Fassungen nicht erhalten sind und die überlieferten Versionen voneinander abweichen, ohne dass es sich erklärtermaßen um Bearbeitungen handelt – eine typische Fallkonstellation etwa für Texte des Mittelalters. Die Vorstellung, auch kleinste Änderungen des Wortlauts stellten generell eine hypertextuelle Transformation dar, würde wohl an der Realität der Textüberlieferung älterer Werke vorbeigehen – wenn damit mehr ausgesagt werden soll als eben das simple Faktum, dass Abschriften in aller Regel nicht völlig exakt sind. Natürlich kann es sinnvoll sein, die Spezifika einer bestimmten Textfassung näher zu untersuchen. Zugleich ist es aber eine offenkundig sinnvolle Strukturierung der Überlieferung, Textfassungen jedenfalls dann, wenn sie nicht zu stark voneinander abweichen, als Realisierungen eines einzigen Textes zu betrachten und zum Beispiel in aller Regel nicht jede Handschrift separat zu edieren, sondern vielmehr in einer Edition Zusammenhang und Unterschiede zwischen den verschiedenen Versionen zu verdeutlichen.[45]

Neben kleineren Änderungen des Wortlauts und unterschiedlichen Schreibungen einschließlich dialektal bedingter Abweichungen gibt es freilich vielfach auch größere Unterschiede, etwa im Textumfang oder -aufbau. In solchen Fällen stößt eine zusammenfassende Edition jedenfalls in gedruckter Form schnell an darstellerische Grenzen, und dementsprechend kann es – bei hinreichender Wichtigkeit –

[44] Vgl. zum Beispiel ebd. S. 80 f. zu den verschiedenen Fassungen von Brechts *Leben des Galilei*.

[45] Hierfür ist insbesondere die Erstellung eines Variantenapparats etabliert, daneben auch – insbesondere bei stärker ausgeprägten Unterschieden – die synoptische Textdarstellung.

sinnvoll sein, die verschiedenen Hauptformen jeweils separat zu edieren.[46] Wenn man davon ausgeht, dass eine Edition einen Text repräsentiert, kann man vielleicht sagen, dass im Fall der separaten Edition verschiedener Fassungen diese jeweils als so eigenständig erscheinen, dass von einer hypertextuellen Beziehung der einen Fassung zur anderen beziehungsweise von beiden zu einem gemeinsamen Hypotext zu sprechen ist. Ob eine Eingrenzung von Hypertextualität auf Textbeziehungen, die die Grenze einer Edition überschreiten, texttheoretisch sinnvoll ist, muss hier nicht geklärt werden. Wesentlich scheint aber, dass mit der Betrachtung hypertextueller Beziehungen auch Textabhängigkeiten in den Blick kommen, deren Darstellung in einer Edition wohl allenfalls fakultativ erfolgt.

Die hier angestellten Überlegungen zu Genettes Kategorien der „Intertextualität“ und der „Hypertextualität“ gehen von der Annahme aus, dass mit Ersterer speziell punktuelle Bezüge bezeichnet werden sollen, bei Letzterer hingegen die Texte als ganze in Beziehung zueinander stehen.[47] Offenkundig handelt es sich dabei allerdings um Eckpunkte einer breiten Skala. Natürlich kann ein Text aber auch zum Beispiel in einem größeren Teilbereich von einem anderen Text abhängig sein, in einem anderen Teilbereich aber von einem dritten – für Genette, wie schon angeführt, eine „Kontamination“ –, oder ein Teil eines Textes basiert auf einer Vorlage, der Rest hingegen ist (mehr oder weniger) neu verfasst.

Und grundsätzlich mag die Frage sein, wie gut ein vor allem an literarischen Texten[48] entwickeltes Klassifikationsschema für die Untersuchung von Sachtexten geeignet ist, insbesondere wenn es sich dabei um Texte handelt, die für eine Institution oder Ähnliches und ohne den Anspruch einer persönlichen Schöpfung verfasst worden sind.

1.1.3 Plagiat, Autorschaft und Urheberrecht

Im letzten Unterkapitel war im Zusammenhang mit Genettes Klassifikation von Arten der „Transtextualität“ schon vom Plagiat die Rede, und natürlich geht es bei der Beschreibung von Texten oder Textpassagen als Plagiate um die Feststellung einer bestimmten Art von Beziehungen zu anderen Texten, also um einen Teilbe-

[46] Als Beispiel aus der Rechtsgeschichte seien die verschiedenen Fassungen des *Schwabenspiegels* genannt (vgl. Johanek 1992, Sp. 899 f.).

[47] Genette erklärt, dass die von ihm gebildeten fünf Kategorien „in der Reihenfolge zunehmender Abstraktion, Implikation und Globalität“ geordnet seien (Genette 1993, S. 10), was eine solche Vorstellung wohl nahelegt.

[48] Genette weist zwar an einzelnen Stellen auch auf nichtliterarische Texte hin, und insbesondere beschreiben einige Kategorien wie etwa die „Paratextualität“ und die „Metatextualität“ Beziehungen, bei denen in der Regel zumindest einer der beteiligten Texte nicht als literarisch zu bezeichnen ist, die große Masse seiner Beispiele behandelt aber die Beziehungen zwischen literarischen Werken. Und auch der Untertitel von Genette 1993, „Die Literatur auf zweiter Stufe“, zeigt diese Ausrichtung.

reich der Intertextualität im hier zugrunde gelegten Verständnis. Wesentlich für den Begriff ist allerdings die damit verbundene Bewertung. Nicht nur die rechtliche Beurteilung von Übernahmen aus anderen Texten, sondern auch die Frage, was als Plagiat betrachtet wird, ist aber abhängig von verschiedenen Rahmenbedingungen, zum einen von der Textsorte, zum anderen vom kulturellen Kontext und auch vom Verständnis der Rolle des Autors. Deshalb sollen die damit zusammenhängenden Fragen hier etwas näher betrachtet werden.[49]

Das Wort *Plagiat* knüpft an eine Stelle in einem Epigramm Martials an, wurde aber – ebenso wie die im Deutschen gebräuchlichen Wörter derselben Wortfamilie – erst in der Frühen Neuzeit gebildet. In dem Epigramm ist von einem *plagiarius* die Rede, was eigentlich einen Menschenräuber bezeichnet, hier aber zur Beschreibung von jemandem dient, der Werke Martials als seine eigenen ausgegeben hatte.[50] Auch wenn die Verwendung des Wortstamms *plag* mit der heutigen Bedeutung also nicht als normaler Sprachgebrauch des klassischen Latein beschrieben werden kann, taucht der Vorwurf, ein Text stamme nicht beziehungsweise nicht vollständig vom erklärten Autor, in Schriften der Antike wiederholt auf. Gängig war dabei die Bezeichnung als κλοπή beziehungsweise *furtum*, also als Diebstahl.[51] Trotzdem handelte es sich aber nicht um ein justitiables Vergehen, sondern diese Wortwahl beruhte nur auf der moralischen Verurteilung von Plagiaten.[52]

Die Vorwürfe bezogen sich dabei generell nicht auf inhaltliche Übereinstimmungen, sondern auf mehr oder weniger wörtliche Übernahmen.[53] Die *imitatio* hingegen, also die freiere Orientierung an literarischen Vorbildern, war üblich und anerkannt. Das musste nicht unbedingt eine umfassende Umformung bedeuten, sondern das Resultat konnte durchaus auch deutliche wörtliche Anlehnungen an die Vorlage beinhalten; dabei wurde aber eine eigene künstlerische Leistung vorausgesetzt.[54]

Während die Bewertung von Abhängigkeitsverhältnissen im literarischen Bereich als *imitatio* beziehungsweise *furtum* nicht immer eindeutig war, gab es bei Ge-

[49] Zum Plagiatsbegriff und seiner Entwicklung, zur Autorschaft im Mittelalter sowie zur Geschichte des Urheberrechts gibt es eine Vielzahl von Publikationen. Abgesehen von den in den folgenden Anmerkungen angeführten Werken sei verwiesen auf Kewes (Hg.) 2003 und Goltschnigg/Grollegg-Edler/Gruber (Hg.) 2013 (mit Bibliographie S. 233–249).

[50] Vgl. zum Beispiel McGill 2012, S. 9 mit Anm. 28 sowie zur Wortgeschichte *http://www.cnrtl.fr/etymologie/plagiat* und *https://www.dwds.de/wb/Plagiat#et-1*.

[51] McGill 2012, S. 8.

[52] Vgl. ebd. S. 10 über das alte Rom. Auch für den griechischen Kulturkreis gibt es wohl keinen Grund, anderes anzunehmen. Vgl. zum Bericht Vitruvs über einen Plagiatsfall in Alexandria mit Ankündigung eines Prozesses die Bewertung in Theisohn 2009, S. 51.

[53] McGill 2012, S. 3 Anm. 8 nennt als einzige Quelle, in der die Verschleierung der Herkunft von philosophischen Ideen als Diebstahl dargestellt wird, eine Stelle aus Ciceros Werk.

[54] Vgl. ebd. S. 19–22. Daneben konnte ein Text als plagiatorisch bewertet werden, wenn dem Verfasser unterstellt wurde, die Abhängigkeit von anderen Texten verheimlichen zu wollen (vgl. ebd. S. 23 f.)

schichtsdarstellungen und technischen Abhandlungen im alten Rom einen recht freizügigen Umgang mit Übernahmen aus älteren Schriften; die Nennung dieser Quellen war keineswegs allgemein üblich und diente, soweit sie erfolgte, in der Regel der Bestärkung der eigenen Glaubwürdigkeit oder der Abgrenzung von Gegenpositionen. Plinius der Ältere forderte zwar den Hinweis auf verwendete Werke, aber auch er beschränkte sich im Wesentlichen auf eine Auflistung dieser Texte.[55]

Die Vorstellung, in gelehrten beziehungsweise wissenschaftlichen Darstellungen die Herkunft übernommener Erkenntnisse durch minutiöse Anmerkungen dokumentieren zu müssen,[56] entwickelte sich erst spät. Zwar finden sich schon bei den römischen Juristen präzise Stellenangaben, und im Mittelalter wurden entsprechende Referenzsysteme auch in anderen Wissensbereichen entwickelt, dabei ging es aber um die Bezugnahme auf autoritative Texte.[57]

Gerade die universitäre Gelehrsamkeit war lange Zeit der Intention nach nicht auf die Gewinnung völlig neuer Erkenntnisse ausgerichtet, sondern vielmehr auf die Erschließung der für das jeweilige Fachgebiet kanonischen Schriften – für den juristischen Bereich waren das das *Corpus Iuris Civilis* einschließlich der *Libri Feudorum* und das *Corpus Iuris Canonici*.[58] Zwar erlangten insbesondere die in die *Glossa Ordinaria* – also in die den Text umrahmenden Erläuterungen – aufgenommenen Auslegungen selbst Autorität und konnten aufgrund ihrer Verknüpfung mit dem Grundtext in ähnlicher Weise wie dieser leicht referenziert werden, eine explizite Bezugnahme auf zeitgenössische Autoren ist aber jedenfalls im hier ausgewerteten Korpus selten,[59] ganz im Gegensatz zu den Verweisen auf das *Corpus Iuris Civilis* und die zugehörige Auslegungsliteratur. Inwieweit das auch für die Zitationspraxis in zeitgenössischen lateinischen Werken gilt, die sich an ein gelehrtes beziehungsweise studentisches Publikum richten, kann hier nicht untersucht werden. Es lässt sich aber vermuten, dass auch von den Verfassern dieser Texte entsprechend dem damaligen Wissenschaftsverständnis kein prinzipieller Bedarf gesehen wurde, die Übernahme eines Gedankens oder eines Befundes zu dokumentieren, soweit als Quelle nicht eine anerkannte Autorität genannt werden konnte, die der eigenen Darstellung mehr Gewicht verlieh.

[55] Vgl. ebd. S. 24–26 sowie 51 f.

[56] Auch im akademischen Bereich gibt es natürlich je nach Textsorte und Fachkonventionen unterschiedliche Praktiken, so etwa den weitgehenden Verzicht auf einen detaillierten Nachweis zu in der Fachwelt anerkannten Aussagen in Lehrbüchern (vgl. Posner 2007, S. 18 f.) oder Lexika, wo Literaturhinweise meist zusammenfassend erfolgen.

[57] Vgl. Grafton 1997, S. 29–31.

[58] Vgl. zum Inhalt und zur Gestaltung der mittelalterlichen und frühneuzeitlichen Lehrveranstaltungen Paulsen 1919, insbesondere S. 35–40 und 263–275, zum juristischen Studium in Mittelalter und Früher Neuzeit Coing 1973, S. 69–80 und Coing 1977, S. 29–53 sowie zur Integration der *Libri Feudorum* ins *Corpus Iuris Civilis* Dilcher 2014, Sp. 971.

[59] Ein Beispiel ist der Verweis auf Andreas Perneder und Ulrich Zasius in Gobler,Rsp. 1550, Bl. 189 r.

Nach diesen kurzen Bemerkungen zum Plagiatsbegriff im alten Rom und zu fußnotenartigen Referenzen in Sachtexten soll ein Blick auf die volkssprachliche mittelalterliche Literatur geworfen werden.[60]

Inwieweit es sinnvoll ist, poststrukturalistische Theorien bei der Analyse zum Beispiel von mittelalterlichen Texten heranzuziehen, sei hier dahingestellt.[61] Offensichtlich lässt sich aber der moderne Werkbegriff, der von einem durch den Autor fixierten und ihm als persönliche Leistung zugerechneten Text ausgeht, nur begrenzt auf diese Literatur anwenden.[62]

Dass die heutigen Vorstellungen vom Autor nicht einfach als für die volkssprachliche Literatur des Mittelalters passend vorausgesetzt werden können, zeigt sich schon darin, dass erhebliche Teile dieser Literatur ohne Nennung des Verfassers oder nur mit einer erst nachträglich zugefügten Zuschreibung überliefert sind.[63] Hier lässt sich eine Entwicklung feststellen von einer mündlichen hin zu einer schriftlichen Kultur. Wenn Texte für den mündlichen Vortrag verfasst und von den Verfassern selbst vorgetragen wurden, bestand kein Bedarf und auch je nach Gattung zum Teil kaum die Möglichkeit, die Urheberschaft im Text selbst kenntlich zu machen.[64] Es ist aber davon auszugehen, dass Vortragstexte auch von anderen Interpreten übernommen und entsprechend der allgemeinen textuellen Variabilität verändert werden konnten.[65] Solche Übernahmen lassen sich etwa für die Sangspruchdichtung feststellen.[66]

[60] Schnell 1998, S. 46 beschreibt als eine der in der Forschung zur Entwicklung des Autor- und Werkbegriffs vertretenen Positionen, beides habe es schon im Mittelalter gegeben, aber nur im „lateinischen Literaturbetrieb“. Worin die lateinische Literatur abweicht und ob dies möglicherweise nur für bestimmte Textgattungen gilt, kann hier nicht weiter untersucht werden.

[61] Vgl. dazu zum Beispiel kritisch Schnell 1998, insbesondere S. 43–45 und 72, sowie mit einer gewissen Anerkennung von Anregungen durch poststrukturalistische Ansätze Bein 1998, S. 76–106, insbesondere S. 77 f. und 106.

[62] Er ist auch in der heutigen Zeit für Texte nicht angemessen, bei denen die Kriterien für einen urheberrechtlichen Schutz nicht erfüllt sind. Im Urheberrechtsgesetz wird ausdrücklich eine Einschränkung auf „persönliche geistige Schöpfungen“ vorgenommen (UrhG § 2 Abs. 2), woraus zu entnehmen ist, dass dies keineswegs für alle Texte gilt.

[63] Haferland 2011, S. 58 f. bietet einen Überblick, für welche Gattungen der deutschsprachigen mittelalterlichen Literatur die Nennung des Verfassernamens typisch ist und für welche die Namenlosigkeit. Vgl. zur nachträglichen Zuschreibung als Ordnungsprinzip unten Anm. 68.

[64] Kühnel 1976, S. 312–314 stellt idealtypisch archaische und moderne Texte einander gegenüber und beschreibt mittelalterliche volkssprachliche Texte als Zwischenstufe; er weist auf Gestaltungsmöglichkeiten beim Vortrag hin, aber auch darauf, dass Texte auch von anderen als den Verfassern vorgetragen werden konnten (vgl. dazu auch Schnell 2001, S. 105–109). Vgl. zur namenlosen Folklore als dem „kulturgeschichtlichen Untergrund von Literatur“ Haferland 2011, S. 53–58 (zitierte Stelle ebd. S. 55). Bein 1999, S. 306 weist für die Lyrik auf die Schwierigkeit einer textinternen namentlichen Kennzeichnung sowie auf den ursprünglich durch den Autor erfolgenden Vortrag hin. Die Variabilität mittelalterlicher volkssprachlicher Texte und die Grenzen dieser Variabilität werden (unter der Bezeichnung *unfester* beziehungsweise *offener Text*) verschiedentlich erörtert; vgl. neben Kühnel 1976 zum Beispiel Bumke 1996, Schnell 1998 und Bein 1998, S. 90–102 sowie mehrere Beiträge in Teil 1 von Peters (Hg.) 2001.

[65] Bein 1998, S. 33 erklärt die Übernahme von Textgut insbesondere nach dem Tod des Urhebers für „mehr als wahrscheinlich“. Vgl. auch die diesbezüglich eben in Anm. 64 genannte Literatur.

Mit zunehmendem Eigengewicht der schriftlichen Fassung ergab sich ein verstärktes Interesse der Autoren, durch eine Namensnennung als Verfasser erkennbar zu bleiben.[67] Zudem wurden auch nachträgliche Zuschreibungen vorgenommen, wobei sich als Motivation insbesondere auch die Verwendung der Autorennamen als Ordnungskriterium zur Strukturierung der Überlieferung erkennen lässt.[68] In diesem Zusammenhang ist auch auf das in den ältesten Drucken noch unbekannte, um 1500 aber weitgehend etablierte Titelblatt hinzuweisen, auf dem zunehmend auch die Nennung des Verfassers üblich wurde – in der *Reichspoliceyordnung* von 1548 wurde sie schließlich im Zusammenhang mit Zensurbestimmungen vorgeschrieben.[69]

Wie schon für das alte Rom lassen sich auch für das Mittelalter und die Frühe Neuzeit vermutlich insbesondere im Bereich der nichtliterarischen Texte wörtliche Übernahmen feststellen – und auch heute noch gilt das jedenfalls für viele Texte, die im Rahmen des Schriftverkehrs erstellt werden. Es lässt sich wohl vermuten, dass in der älteren Literatur insbesondere bei der Wiedergabe tradierter Wissensbestände und Lehren der Individualität der Formulierung kein Gewicht beigemessen wurde.[70]

In diesem Zusammenhang ist auf den weiten Bereich sammelnder und kompilatorischer Literatur hinzuweisen. Unter *Kompilationen* sollen hier allgemein Werke verstanden werden, die sich im Wesentlichen (also zum Beispiel unter Ausklammerung von einleitenden Bemerkungen) als Zusammenstellung beziehungsweise Zusammenführung von Passagen anderer Texte (oder vielleicht auch von sehr kurzen Texten) beschreiben lassen.[71] Diese Definition schließt eine gewisse Un-

[66] Vgl. Haustein/Stackmann 1998 sowie Stackmann 1998. Schnell 1998, S. 64 Anm. 200 verweist auf die vielfachen Abhängigkeitsverhältnisse in der Sangspruchdichtung des 13. Jahrhunderts.

[67] Vgl. Haferland 2011, S. 62 f. Nach Eisenstein 1997, S. 77–79 kommt insbesondere dem Buchdruck dabei erhebliche Bedeutung zu. Dazu passt auch die Forschungsmeinung, die die Entstehung des Autor- und Werkbegriffs mit dem Buchdruck verknüpft (vgl. Schnell 1998, S. 45).

[68] Vgl. zu Autorzuweisungen in deutschsprachigen literarischen Handschriften des 14. Jahrhunderts Bein 1998, S. 193–233. Bein weist auf abweichende Prinzipien im 15. Jahrhundert hin (ebd. S. 196) und darauf, dass bei Sachtexten Autornamen teils keine Rolle spielen, teils unter dem Aspekt der Autorität wichtig sind (ebd. S. 197).

[69] RPolO. 1548, Bl. 29 r (vgl. Giesecke 1991, S. 442). Vgl. zur Entwicklung und Verbreitung des Titelblatts Kiessling 1930, Geldner 1978, S. 107–111 sowie Rautenberg 2004 (statistische Daten ebd. S. 10 f.; für den Hinweis auf diese Publikation danke ich Herrn Professor Dr. Manfred Thaller).

[70] Schnell 1998, S. 63 hält eine Unterscheidung zwischen „Texte[n], die vor allem durch ihre wissensvermittelnde Funktion bestimmt werden“ und „Texte[n], deren Verfasser den Anspruch auf einen ästhetischen ‚Überschuß‘ erkennen lassen“ für sinnvoll. Dabei geht es zwar um die Frage, inwieweit mittelalterliche Texte *offen*, also im Wortlaut nicht oder jedenfalls nicht völlig fixiert waren, man kann aber natürlich umgekehrt – jedenfalls bei nicht unerheblichen Unterschieden – auch die jeweilige Individualität der Texte beziehungsweise Textfassungen betonen, so dass nicht die Varianz als Abweichung vom Ausgangstext, sondern vielmehr die Übereinstimmung als Übernahme zu beschreiben ist.

[71] Diese Definition ist weiter als die von Kallweit 2000, S. 317. Dort wird die Kompilation beschrieben als „Textsammlung durch Ausschreiben und Zusammenstellen der Quellenbestände zu einem

schärfe in der Abgrenzung von der Textsammlung ein, die im Wesentlichen aus in sich abgeschlossenen und vollständigen Einzeltexten besteht (zum Beispiel eine Anthologie oder ein Kochbuch). Texte, die nur teilweise auf der Zusammenführung verschiedener Vorlagen basieren, sind nach der eben gewählten Definition – wohl entsprechend der üblichen Begriffsverwendung[72] – zwar als solche keine Kompilationen, sollen aber durch das Adjektiv *kompilatorisch* (im Sinne von: zumindest teilweise auf Kompilation beruhend) mit abgedeckt sein.

Die Unterscheidung zwischen Kompilationen und Textsammlungen ist für die hier behandelten Fragen kaum von Bedeutung und beruht auch nicht auf einem durchdachten Klassifikationsschema, sondern wurde *ad hoc* vorgenommen, da der Begriff *Kompilation* wohl nicht gut für die Bezeichnung der Zusammenstellung vollständiger Texte verwendbar ist, die Untersuchung solcher Zusammenstellungen aber durchaus in ähnlicher Weise wie die von Kompilationen von Interesse sein kann, wenn es etwa um die Klärung von Abhängigkeitsverhältnissen geht.

Während wohl recht unmittelbar verständlich ist, was sich als Textsammlung beschreiben lässt, ist im Hinblick auf kompilatorische Literatur vermutlich eine etwas nähere Erläuterung anhand einiger Beispiele sinnvoll.

Zunächst einmal sind hier die Florilegien einzuordnen, also Zusammenstellungen von Exzerpten aus verschiedenen Texten, wobei die Textfragmente nach der Reihenfolge in den Vorlagetexten, aber auch nach inhaltlichen oder formalen Kriterien angeordnet und durch Überschriften strukturiert sein können.[73] Obwohl sich die Bezeichnung *Florilegium* ebenso wie *Anthologie* mit *Blütenlese* übersetzen lässt, sind die Begriffe nicht inhaltsgleich. Wesentliches Unterscheidungsmerkmal ist, ob kurze Auszüge oder in der Regel geschlossene Einzeltexte wiedergegeben werden.[74] Eine genaue inhaltliche Fassung kann hier nicht geleistet werden und dementsprechend auch keine Abgrenzung zu anderen Bezeichnungen für Werke, die

Thema; als Büchertypus verbreitet vor allem im Kontext der Wissenschaftslehre des 16. und 17. Jhs." (der Abschnitt zur Sachgeschichte ebd. S. 319 f. umfasst allerdings die Zeit von der Antike bis zum 18. Jahrhundert; vgl. zum 13. und 14. Jahrhundert Minnis 1979, wo wohl ein ähnliches Verständnis von *Kompilation* wie bei Kallweit vorausgesetzt wird). Hier hingegen sollen auch Werke ohne eine thematische Auswahl beziehungsweise Ordnung mit eingeschlossen sein. Ein sehr weiter Begriff von *Kompilation*, der auch Textsammlungen einschließt, liegt anscheinend Meierhofer 2010 zugrunde (vgl. ebd. S. 8 ff.).

[72] Hier lässt sich auf das berühmte Zitat von Bonaventura verweisen, in dem *scriptor*, *compilator*, *commentator* und *auctor* einander gegenübergestellt werden und es über den *compilator* heißt: „Aliquis scribit aliena, addendo, sed non de suo" (zitiert nach Minnis 1979, S. 415). Minnis weist ebd. S. 416 allerdings darauf hin, dass Kompilatoren wie Vinzenz von Beauvais durchaus auch selbst formulierte Passagen hinzufügten, wobei ihnen die Abgrenzung von den übernommenen Äußerungen der *auctores* ein Anliegen war.

[73] Vgl. Munk Olsen 1982, S. 152–154, wo zusammengestellt ist, in welcher Hinsicht mit einer Formung des Materials durch die Kompilatoren zu rechnen ist. Grubmüller 1997 bietet unter anderem einen Überblick über die Entwicklung der Gattung sowie Literaturhinweise.

[74] Vgl. Grubmüller 1997, S. 606, wo allerdings auch auf teilweise abweichenden Wortgebrauch hingewiesen wird.

auf Exzerption beruhen, wie *Analekten*, *Kollektaneen* oder *loci communes* – wenn eine solche Abgrenzung überhaupt möglich und sinnvoll ist.[75]

Immerhin soll aber an dieser Stelle festgehalten werden, dass derartige Sammlungen von Textfragmenten lange Zeit sehr verbreitet waren, und zwar nicht nur als publizierte Werke, sondern auch als jeweils persönlich erstellte Notizen. Sie dienten als offenbar regelmäßig verwendetes Hilfsmittel für die Abfassung literarischer oder auch gelehrter Werke.[76] Für die Exzerption und die Ordnung der Exzerpte wurden in der Frühen Neuzeit ausgefeilte Techniken entwickelt.[77]

Dass in dieser Weise erstellte Textauszüge aus zeitgenössischen Schriften gegebenenfalls im Wortlaut und ohne Quellenangabe – oder nur mit schon in der Vorlage vorhandenen Quellenangaben – in eigene Publikationen übernommen, aber auch nach Belieben verändert werden konnten, passt wohl gut ins Bild einer Textkultur, die keinen Wert auf individuelle Formulierung legte und in der Stellenangaben zwar zur Anknüpfung an die anerkannten Autoritäten wichtig waren, nicht aber um den Eindruck eines falschen Originalitätsanspruchs zu vermeiden, da es jedenfalls in weiten Bereichen der gelehrten und erst recht in der übrigen wissensvermittelnden Literatur gar nicht um neue Erkenntnisse im Sinne moderner Forschung ging.

Während Florilegien und ähnliche Werke wohl typischerweise als dauerhafte Sammlungen von Textfragmenten gedacht waren, die als mehr oder weniger zeitlos von Interesse betrachtet wurden, war und ist das Nachrichtenwesen[78] gerade umgekehrt von oft nur kurzlebigen Informationen bestimmt, die aber ebenfalls vielfach in wörtlichen Übernahmen weitergegeben werden.[79] Auch heute noch sind Übernahmen in diesem Bereich recht verbreitet und auch schon verschiedentlich untersucht worden.[80] Entsprechendes lässt sich wohl auch für frühere Zeiten

[75] Diese Begriffe sind im RLW nicht als Stichwörter verzeichnet. Vgl. zu *Analekten* und *Kollektaneen* Wilpert 1979, S. 25 und 414, zu *loci communes* Zedelmaier 1992, S. 69 ff. Im englischen Sprachraum ist offenbar die Bezeichnung *commonplace books* üblich. Sie findet sich auch in deutscher Literatur dazu, vgl. zum Beispiel Plett 1990, S. 173–175. Handschriftliche Sammlungen werden ebd. S. 175 als „*loci-communes*-Hefte" bezeichnet, so dass wohl davon auszugehen ist, dass die englische Wortwahl nicht durch einen inhaltlichen Unterschied, sondern nur durch die in der anglistischen Literatur eingeführte Terminologie motiviert ist. Plett 1994, S. 19 führt „*Kollektaneen, Florilegien, Thesauri* oder *Commonplace Books*" als Bezeichnungen für „*Topische* Poetiken" an. Offenbar ist also von einer zumindest partiellen Austauschbarkeit dieser Termini auszugehen.

[76] Vgl. Plett 1990, S. 174, Wilpert 1979, S. 414 sowie Zedelmaier 1992, S. 72.

[77] Vgl. Zedelmaier 1992, S. 103–106, Zedelmaier 2000, S. 84–90, Zedelmaier 2001, S. 21–23 sowie Neumann 2001, S. 57–61.

[78] Der Begriff ist hier in einem weiten Sinn gemeint und soll allgemein Mitteilungen über Neuigkeiten verschiedenster Art umfassen.

[79] Ob dies als Zusammenstellung von Textfragmenten einzustufen ist oder als Sammlung von Einzeltexten, hängt wohl auch von der Betrachtungsebene ab, nämlich davon, ob die Übernahme von einzelnen Stücken einer anderen Textzusammenstellung (zum Beispiel einer Zeitung) von diesen Einzeltexten aus gedacht wird oder vom Gesamttext der Vorlage aus.

[80] Vgl. unten S. 89.

feststellen[81] – allerdings ohne Presseagenturen und ohne Rechtsregeln für den Nachdruck.[82]

Sowohl bei Exzerptsammlungen als auch bei Zeitungen und Zeitschriften, die einzelne Mitteilungen übernehmen, ist im Regelfall davon auszugehen, dass die übernommenen Textstücke als separate Einheiten erkennbar sind; auch mit einer Quellenangabe ist in vielen Fällen zu rechnen, für Florilegien etwa mit der Nennung von Autoritäten.[83] Daneben gibt es aber auch Werke, in denen Textstücke von Vorlagen ohne klare Abgrenzung zusammengefügt sind, so dass der kompilatorische Charakter nicht unmittelbar erkennbar ist. Hier lassen sich etwa mystische Komposit- und Mosaiktraktate anführen.[84]

Mit den vorgestellten Beispielen sollte zumindest ansatzweise verdeutlicht werden, dass in vielen Bereichen kompilatorischen Praktiken festzustellen sind. Darüber hinaus ist natürlich mit einem weiten Spektrum der Verwendung von Vorlagetexten zu rechnen, bei der nicht einfach nur Textpassagen übernommen werden, sondern eine stärkere Bearbeitung erfolgt, zum Beispiel durch die Einfügung eigener Kommentare oder anderer Zusätze oder durch die partielle Veränderung des Wortlauts und teilweise auch des Inhalts. Gerade solche adaptierenden Übernahmen lassen sich bei den hier untersuchten Abhängigkeitsverhältnissen zwischen Rechtstexten vielfach feststellen. Wie bei Kompilationswerken können dabei mehrere Texte als Vorlage dienen, aber natürlich kann auch nur ein einziger Basistext verwendet werden. Teil 4 dieser Untersuchung bietet einige Beispiele für etwas komplexere Textbeziehungen.

Auch über den Bereich des Rechtswesens hinaus dürften Gebrauchstexte vielfach zumindest partiell als Bearbeitung einer Textvorlage – oder auch als Anwendung eines möglicherweise erlernten Textmusters – zu erklären sein. Mit der Abfassung

[81] Auch in diesem Punkt muss die vorliegende Darstellung weitgehend im Spekulativen bleiben, da sie sich nur auf Zufallsfunde in der Literatur stützen kann.

[82] Ein Beispiel stellen die in den *Gelehrten Zeitungen* der Aufklärungszeit enthaltenen Informationen über Ereignisse der gelehrten Welt und insbesondere auch über Neuerscheinungen dar, wobei diese Informationen zum einen selbst die Form von Exzerpten haben und zum anderen vielfach – mit oder ohne Quellenangabe – von anderen Zeitschriften übernommen wurden – vgl. Gierl 2001, S. 71–80. Ebd. S. 76 wird als Desiderat der Vergleich der *Gelehrten Zeitungen* mit anderen Rezensionsorganen genannt, freilich auch die Schwierigkeit der Umsetzung artikuliert: „Ein endloses Geschäft!“ Offenbar könnte ein automatisiertes Vergleichsverfahren hier Unterstützung leisten.

[83] Dies muss allerdings nicht immer der Fall sein. Schmidt 2000, S. 21 weist auf die Schwierigkeit des Quellennachweises für mittelalterliche Zitate in zwei Florilegien-Handschriften hin, da die Verfassernamen teilweise nicht oder falsch angegeben seien. Auch hier zeigt sich also ein möglicher Anwendungsfall für eine automatisierte Erkennung von Textübereinstimmungen.

[84] Vgl. dazu zum Beispiel Hasebrink 2000 (zur Unterscheidung zwischen Komposit- und Mosaiktraktaten ebd. S. 80). Auch in diesen Texten kann es Hinweise auf Autoritäten geben (vgl. zum Beispiel ebd. S. 76 zur Nennung Meister Eckharts in *Spamers Mosaiktraktaten*), dies dürfte allerdings vielfach wohl nicht für eine Abgrenzung der Quellenstücke voneinander ausreichen, zumal schon in der jeweiligen Vorlage Zitate enthalten sein können (vgl. entsprechende Beispiele ebd. S. 81 f.). Für den Hinweis auf diese Texte und auf Publikationen dazu danke ich Dr. Carola Redzich.

auf der Basis verschiedener Vorlagen ist wohl am ehesten bei komplexen und als zumindest einigermaßen wichtig eingestuften Texten zu rechnen. Beispiele hierfür lassen sich vermutlich im Bereich der Lexika finden.[85] Aber auch wenn jeweils nur eine einzige Quelle verwendet wurde, können sich Traditionszusammenhänge ergeben, die für auf diese Weise miteinander verbundene größere Textgruppen nur schwer zu durchschauen sind.[86]

Schließlich soll in diesem Unterkapitel noch kurz auf urheberrechtliche Gesichtspunkte eingegangen werden, da Rechtsfragen bei der Beurteilung von Textübernahmen natürlich eine nicht unerhebliche Rolle spielen. Auch nach den heute gültigen Regeln sind wörtliche Übernahmen keineswegs generell untersagt. Vielmehr ist der urheberrechtliche Schutz an verschiedene Bedingungen geknüpft, insbesondere (nach dem deutschen Recht) daran, dass es sich um „persönliche geistige Schöpfungen“[87] handelt und dass die 70 Jahre nach dem Tod des Urhebers endende Schutzfrist[88] noch nicht abgelaufen ist. „Gesetze, Verordnungen, amtliche Erlasse und Bekanntmachungen sowie Entscheidungen und amtlich verfaßte Leitsätze zu Entscheidungen“ sind ausdrücklich vom urheberrechtlichen Schutz ausgenommen.[89] Zudem besteht natürlich die Möglichkeit des Zitats.[90] Auf die komplexen Fragen, was der urheberrechtliche Schutz eines Werks für Konsequenzen hat, muss hier nicht eingegangen werden. Es ist aber jedenfalls festzuhalten, dass der Begriff *Plagiat* im Urheberrechtsgesetz nicht vorkommt, dass keineswegs jedes Plagiat (was auch immer genau darunter zu verstehen ist) eine Urheberrechtsverletzung darstellt[91] und dass wörtliche Übernahmen zum Beispiel aus Gesetzestexten auch heutzutage urheberrechtlich unproblematisch sind.

Insbesondere aber ist im Hinblick auf den hier untersuchten Gegenstand zu konstatieren, dass sich die Vorstellung vom geistigen Eigentum, das rechtlich zu schützen ist, erst spät nachweisen lässt. Schon um 1500 entwickelte sich zwar das Instrument des Druckprivilegs, das anderen den Nachdruck jeweils eines bestimm-

[85] Dass gerade bei Lexika mit der Verwendung von Vorlagetexten zu rechnen ist, liegt wohl auf der Hand. Dabei kann es sich um wissenschaftliche Literatur, aber natürlich auch um andere Lexika handeln. Gierl 2001, S. 82–84 stellt am Beispiel von Wortartikeln zum Buchstaben *a* vor, wie sich bestimmte Informationen und auch inhaltliche Irrtümer wiederholen. Vgl. zum *Grossen vollständiges Universal-Lexikon* von Zedler – einem Werk, das bekanntlich in erheblichem Maße auf Kompilation beruht – ebd. S. 87.

[86] Schnell 1998, S. 63 f. weist darauf hin, dass sich etwa bei Dekalogerklärungen „so viele Übereinstimmungen, Abhängigkeiten und Interferenzen“ fänden, dass eine Würdigung der Einzeltexte nur im Gesamtvergleich möglich wäre. Weitere ebd. genannte Beispiele für verändernde Textübernahmen sind die Sangspruchdichtung sowie Predigten.

[87] UrhG § 2 Abs. 2.

[88] Ebd. § 64.

[89] Ebd. § 5 Abs. 1.

[90] Ebd. § 51.

[91] Dies ergibt sich schon aus der zeitlichen Befristung des urheberrechtlichen Schutzes (vgl. zum Beispiel Nitsche 2013, S. 83 f.).

ten Werks für begrenzte Zeit untersagte,[92] ein solches Privileg wurde aber nur in besonderen Fällen erteilt,[93] insbesondere bei Texten, deren erstmalige Publikation als gemeinnützig betrachtet wurde. Dies konnten auch Texte sein, die heute nicht unter urheberrechtlichen Schutz fallen würden – im hier untersuchten Korpus gilt das zum Beispiel für zahlreiche Reichsabschiede.[94] Die Privilegien wurden teils den Autoren beziehungsweise Bearbeitern, teils den Druckern erteilt und damit begründet, dass ihnen der Gewinn aus ihrem Werk nicht entzogen werden beziehungsweise aus diesem Werk kein Schaden erwachsen sollte.[95] Allerdings war in der Frühzeit des Buchdrucks ein Autorenhonorar noch nicht üblich, und auch in den Privilegien für Autoren wird nicht auf die schöpferische Leistung, sondern auf den Aufwand abgehoben.[96] Erst im 18. Jahrhundert ist von einem „Eigentum“ der Autoren an ihren Texten die Rede.[97]

Von den aus dem heutigen Urheberrecht erwachsenden Rechtsansprüchen ist im vorliegenden Zusammenhang wohl vor allem zweierlei von Interesse: zum einen das Verbot, Bearbeitungen oder Umgestaltungen ohne Zustimmung des Verfassers zu publizieren, zum anderen der Anspruch des Urhebers auf eine Nennung unter anderem bei Zitaten.[98] Abgesehen von den Druckprivilegien, die mit dem Schutz materieller Interessen eine andere Ausrichtung hatten und zudem nur für einzelne Drucke und für begrenzte Zeit galten,[99] lässt sich ein entsprechender Schutz für den hier untersuchten Zeitraum nicht feststellen. Zwar gab es einzelne Äußerungen von Autoren, die über unautorisierte und fehlerhafte Nach- oder auch Erstdrucke klagten, diesbezügliche gerichtliche Auseinandersetzungen lassen sich aber für die Frühzeit des Buchdrucks anscheinend nicht nachweisen.[100] Grundsätzlich ist davon auszugehen, dass das Nachdrucken von Texten zunächst „etwas Selbstverständliches“[101] war.

[92] Vgl. dazu allgemein Gieseke 1995, S. 39–90.

[93] Vgl. ebd. S. 41–52 mit einer Auflistung der etwa 70 bis zum Jahr 1530 erteilten Privilegien für im Heiligen Römischen Reich erschienene Drucke.

[94] Als Grund für den Schutz wird dabei in einigen der Privilegien die schnelle Publikation angeführt (ebd. S. 63).

[95] Vgl. ebd. S. 58–66.

[96] Vgl. Gieseke 1957, S. 20 f. und 35.

[97] Vgl. ebd. insbesondere S. 78–82. Die Formulierung *geistiges Eigentum* findet sich laut Bein 1998, S. 113 (mit Verweis auf eine Untersuchung von Heinrich Bosse) erstmals in einem Text von 1784.

[98] Vgl. UrhG § 23 und § 63.

[99] Auf die Weiterentwicklung hin zu Generalprivilegien für alle von einem Drucker oder Verleger herausgebrachten Werke und zur Verlängerung von Privilegien braucht hier nicht eingegangen zu werden. Vgl. dazu Gieseke 1995, S. 75–77.

[100] Vgl. ebd. S. 20–26 und 29–33 zu Äußerungen von Sebastian Brant, Martin Luther und Johannes Sleidanus sowie zum Streit zwischen dem Drucker Christian Egenolf und Conrad Lagus über den Druck einer Vorlesungsmitschrift gegen Lagus' Willen. Lagus drohte zwar mit einer Klage, verstarb aber bald danach (ebd. S. 30). Luthers Beschwerde beim Rat von Nürnberg wegen des – tatsächlich dort gar nicht erfolgten – Drucks eines gestohlenen Manuskripts führte nur zu einem Beschluss, dass Luthers Werke dort nur nach Korrektur und ohne die Behauptung, in Wittenberg gedruckt zu sein, erscheinen sollten (ebd. S. 25.).

Und insbesondere fehlen Hinweise darauf, dass die ungekennzeichnete Übernahme von Textpassagen nicht statthaft gewesen wäre. Vielmehr besteht nach dem, was in diesem Unterkapitel dargestellt wurde, wohl Anlass zur Vermutung, dass dies im Bereich der Sachtexte als lange Zeit durchaus normale Praktik bei der Texterstellung zu betrachten ist. Eine genauere Einschätzung, wie verbreitet dies war, ist allerdings wohl nur möglich, wenn mit Hilfe automatisierter Verfahren größere Textmengen verglichen werden können.

1.2 Textkorpus und rechtshistorische Zusammenhänge

Dieses Kapitel soll in das Quellenkorpus einführen, auf dem diese Arbeit basiert. Zunächst wird das Projekt DRQEdit vorgestellt, in dessen Rahmen das im Folgenden ausgewertete Korpus erstellt wurde. Dabei geht es insbesondere um das Gesamtkorpus des Projekts, die Kriterien für seine Zusammenstellung, seine rechtshistorische Einordnung und einen ersten Überblick über die darin enthaltenen Texte. Im zweiten Unterkapitel wird das Teilkorpus, das bereits in maschinenlesbarer Form vorliegt und deshalb als Textgrundlage für die Untersuchungen in Teil 3 und 4 dient, etwas näher beschrieben; dabei werden auch Abhängigkeitsverhältnisse thematisiert.

1.2.1 Das Projekt DRQEdit

Die vorliegende Untersuchung ist in zweierlei Hinsicht eng mit dem Projekt *DRQEdit – Deutschsprachige Rechtsquellen in digitaler Edition*[102] verbunden: Zum einen liegen ihr die Texte zugrunde, die bisher im Rahmen dieses Projekts erfasst wurden. Zum anderen ist die Fragestellung wesentlich dadurch motiviert, dass es zwischen den Texten, die den Gegenstand von DRQEdit bilden, eine Vielzahl enger Beziehungen und insbesondere auch wörtlicher Übereinstimmungen gibt, dass diese Beziehungen aber selten in den Texten selbst deutlich gemacht werden[103] und auch in der wissenschaftlichen Literatur nur sehr begrenzt aufgearbeitet sind. Deshalb sollen das Projekt und sein Quellenkorpus in diesem Unterkapitel etwas näher vorgestellt und rechtshistorisch eingeordnet werden.

DRQEdit hat das Ziel, die im 15. und 16. Jahrhundert erstmals im Druck erschienenen deutschsprachigen Rechtsquellen unter Ausschluss von Kirchenordnungen, policeyrechtlichen Sonderordnungen sowie Publikationen zu einzelnen rechtlichen Auseinandersetzungen und einschließlich von Autorenwerken und substantiell veränderten Fassungen als Korpus verfügbar zu machen und zu erschließen.[104]

[101] Gieseke 1995, S. 14. Gieseke weist deshalb darauf hin, dass die Bezeichnung von Nachdrucken als „Raubdrucke“ nicht sachgerecht sei (ebd. S. 33).
[102] *http://drw-www.adw.uni-heidelberg.de/drqedit/*.
[103] Am ehesten sind entsprechende Angaben in den Vorreden zu finden.

Die Kriterien für die Auswahl der zum Korpus gehörenden Texte und Ausgaben bedürfen wohl der Erläuterung. Zunächst einmal: Der Intention nach sollen sämtliche diesen Kriterien entsprechenden Drucke aufgenommen werden, und dementsprechend kann es zu Erweiterungen des Korpus kommen, wenn sich neue Erkenntnisse dazu ergeben. Eine Einbindung von Faksimiles ist aber nicht bei allen Drucken möglich. Zum einen sind nicht alle Ausgaben heute noch erhalten (beziehungsweise über die einschlägigen Nachweisinstrumente auffindbar), zum anderen kann eine Digitalisierung zum Beispiel an konservatorischen Bedenken scheitern. Um den Gegenstandsbereich des Projekts möglichst umfassend abzubilden, umfasst das Korpus auch solche nicht verfügbaren Drucke und gegebenenfalls zusätzlich eine später erschienene Ausgabe, um jedenfalls den Text als solchen einbeziehen zu können.

Eine konsequente Umsetzung der Beschränkung auf Erstausgaben war auch in anderer Hinsicht nicht möglich. Auch die zugrunde gelegten bibliographischen Hilfsmittel, vor allem der *Incunabula Short Title Catalogue (ISTC)* für das 15. und das *Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts (VD 16)* für das 16. Jahrhundert, bieten keine wirklich vollständigen Informationen, vielmehr werden sie gegebenenfalls immer wieder ergänzt. Und auf der Basis der darin enthaltenen Angaben lässt sich bei mehreren Ausgaben aus demselben Jahr beziehungsweise mit unklarer Datierung oft nicht entscheiden, welche davon tatsächlich die älteste ist.[105] Selbst wenn prinzipiell eine genauere zeitliche Einordnung möglich sein sollte, konnten diesbezügliche Untersuchungen im Rahmen dieses Projekts nicht betrieben werden. Das Grundanliegen, eine möglichst originale Textfassung auszuwählen, dürfte dadurch freilich nicht allzu sehr beeinträchtigt sein.

Eine weitere Schwierigkeit ergibt sich bei der Abgrenzung, was überhaupt als eigener Text beziehungsweise als substantielle Bearbeitung eines anderen Textes zu betrachten und somit ins Korpus aufzunehmen ist und was als im Wesentlichen unveränderte Ausgabe. Dass es auf den Titelblättern von Neuauflagen oft zum Beispiel heißt, der Text sei „gebessert“, „corrigirt“ oder „gemehret“[106], zeigt, dass durchaus Anlass zu einer sorgfältigeren Prüfung bestehen mag, vielfach ist aber

[104] Das im Projekttitel vorkommende Wort *Rechtsquelle* wird anscheinend unterschiedlich verwendet, was zum einen am jeweils zugrunde liegenden Rechtsbegriff liegen mag, zum anderen an der Bedeutung von *Quelle*, die im Sinne eines sachlichen Ausgangspunktes beziehungsweise Grundes, aber auch – wie im Sprachgebrauch der Historiker – im Sinne einer Erkenntnisquelle verstanden werden kann. In diesem letzteren Sinne wird der Ausdruck auch hier gebraucht (ebenso Schwerin 1950, S. 5–7 sowie Kaufmann 1986). Bader 1954/1984, S. 268–270 weist auf die Problematik einer Unterscheidung zwischen sogenannten *unmittelbaren* und *mittelbaren Rechtsquellen* hin, die von einem positivistischen Verständnis von *Recht* ausgeht und den historischen Verhältnissen nicht gerecht wird. Für den Hinweis auf diesen Aufsatz danke ich Herrn Dr. Heino Speer.

[105] Zwar ist bei Inkunabeln aufgrund der Angaben im Kolophon zum Teil eine recht präzise Datierung möglich, dies betrifft aber nur einen kleinen Teil des Korpus.

jedenfalls nicht unmittelbar erkennbar, worin diese Änderungen wohl bestehen. Ein umfassender Vergleich mag für eine Einzeledition durchführbar sein, er ist es aber nicht für ein Korpus der hier untersuchten Größenordnung bei begrenzten personellen Kapazitäten und ohne die Anwendung automatisierter Verfahren, die ja das Vorliegen maschinenlesbarer Texte voraussetzen würden.

Immerhin haben sich aber durch die Massendigitalisierung der letzten Jahre die Möglichkeiten sehr verbessert, da inzwischen von einem erheblichen Teil der betreffenden Ausgaben Faksimiles verfügbar sind. So lässt sich inzwischen oft zum Beispiel zumindest oberflächlich prüfen, ob sich Aufbau und Umfang zweier Ausgaben im Wesentlichen entsprechen. Da auch dies nur nach und nach bei der näheren Beschäftigung mit einem Text erfolgen kann, ist freilich damit zu rechnen, dass sich hierbei noch Erweiterungen für das Korpus ergeben.

Umgekehrt kann sich auch zeigen, dass ein zunächst als eigenständig betrachteter Druck tatsächlich nur die Neuauflage eines älteren Werkes ist. Da manche Texte unter recht verschiedenen Titeln erschienen sind, ist mit Fehleinschätzungen durchaus zu rechnen. Im Interesse einer stabilen Adressierbarkeit der digitalen Ressourcen bleiben solche Ausgaben allerdings im Korpus. Ein weiterer Grund für Textdubletten besteht darin, dass das Korpus eine Reihe von Textsammlungen und von Autorenwerken mit angehängten Texten enthält, in denen zum Beispiel bestimmte Normtexte immer wieder zu finden sind. Solche Sammlungen haben auch als solche einen eigenen Quellenwert, da sich in ihrer Zusammenstellung Interessen der Zeit widerspiegeln. Dementsprechend werden auch sie zumindest in ihrer jeweils ersten Form in DRQEdit berücksichtigt. Da die Zusammenstellung der enthaltenen Texte bei ihnen aber oft nicht in allen Auflagen gleich ist, stellen sich hier besondere Schwierigkeiten für die Entscheidung, was ins Korpus aufzunehmen ist und was nicht.

Die oben beschriebene Ausgrenzung bestimmter Textgruppen ist pragmatisch motiviert, da die Textmenge anders nicht zu bewältigen wäre und Kirchenordnungen sowie policeyrechtliche Sonderordnungen zudem überwiegend andere Sachbereiche regeln als die ins Korpus aufgenommenen Normtexte. Aus dieser Abgrenzung ergibt sich als thematischer Schwerpunkt das Zivil- und Strafrecht.

Die Vollrezeption des römischen Rechts im deutschen Sprachraum – also die zunehmende Umgestaltung der Rechtspraxis auf der Basis des *Corpus Iuris Civilis* entsprechend der universitären Lehre – lässt sich zeitlich wohl in etwa eingrenzen auf die zweite Hälfte des 15. und das 16. Jahrhundert,[107] fällt also mit dem im Projekt

[106] So in weiteren Auflagen zu SammelwerkSächsR.(Wolrab) 1541 (zum Beispiel VD16 Nr. K 1846, K 1847 und K 1848). Gieseke 1995, S. 27 erklärt die vielfachen Überarbeitungen als eine gegen Nachdrucke durch andere Drucker gerichtete Maßnahme.

[107] So in Eisenhardt 1995, S. 99. Wie etabliert genau diese Grenzziehung ist, soll hier nicht erörtert werden. Wie allgemein bei Epochenabgrenzungen ist natürlich mit Übergangsphasen zu rechnen.

behandelten Zeitraum zusammen – die ältesten zum Korpus gehörenden Drucke erschienen im Jahr 1472.[108] Der neu entstandene Buchdruck und die zunehmende Publikation von Werken in deutscher Sprache ermöglichten die Verbreitung von Kenntnissen über das römische Recht weit über den universitären Rahmen hinaus und insbesondere bei den Rechtspraktikern.[109] Zudem wurde in der Reichskammergerichtsordnung von 1495 das römische Recht als subsidiär im Heiligen Römischen Reich geltendes Recht verankert.[110]

In dieselbe Zeit fallen auch die Reichsreform – in die die eben erwähnte Reichskammergerichtsordnung einzuordnen ist – und der Ausbau der territorialen Administration und Gesetzgebung, woran juristisch gebildete Amtsträger wesentlichen Anteil hatten.[111] Die Reichsgesetzgebung, Landrechte und Landesordnungen sowie Stadtrechtsreformationen stellen wichtige Textgruppen innerhalb des Korpus dar.

Ein besonders wichtiges Anliegen war offenbar die Regelung der prozessualen Verfahrensweisen. Dieses Thema nimmt in den Land- und Stadtrechten oft einen breiten Raum ein. Zudem gibt es eine Vielzahl von Gerichtsordnungen zum Zivilprozess sowie einige zum Beispiel als *Peinliche Gerichtsordnung* bezeichnete Ordnungen zum Straf- und Strafverfahrensrecht.

Neben diesen verschiedenen Normtexten bilden privat erstellte, der Wissensvermittlung dienende Autorenwerke die zweite große Hauptgruppe des Korpus. Allerdings wäre eine scharfe Unterscheidung zwischen beiden Gruppen wohl nicht sachgerecht. So ist etwa der (im 13. Jahrhundert verfasste) *Sachsenspiegel* zwar das Werk Eike von Repgows, ihm wurde aber durchaus rechtliche Geltung zugeschrieben.[112] Und umgekehrt lässt sich die *Wormser Reformation* von 1498 trotz der offiziellen Verabschiedung von der Art des Textes her auch oder sogar vorrangig als Lehrbuch beschreiben.[113]

So wurde der *Klagspiegel*, das erste große Handbuch zum römischen Recht in deutscher Sprache, um 1436 verfasst (vgl. Deutsch 2004, S. 26–51 und Deutsch 2012, Sp. 1864 f.).

[108] Es handelt sich um BerthRechtssumme(dt.) 1472, Belial(dt.) 1472 (Zainer) und OrdoJudiciarius(dt.) 1472.

[109] Den Begriff *Rechtspraktiker* übernehme ich von Eva Schumann, die ihn erklärt als „alle funktional in die Rechtspflege eingebundenen Personen (Richter, Schöffen, Advokaten, Prokuratoren, Gerichtsschreiber, Notare usw." (Schumann 2013, S. 123 Anm. 1) und darauf hinweist, dass es sich dabei in der Frühen Neuzeit vielfach um Personen ohne ein Jurastudium handelte (ebd. S. 125 f.). Auch sie weist auf den Buchdruck als Voraussetzung für die weite Verbreitung der Praktikerliteratur und auf die Bedeutung dieser Werke für die Rezeption hin (ebd. S. 140 f.; zum Begriff *Praktikerliteratur* ebd. S. 136–138).

[110] Vgl. zum Beispiel Laufs 1984, S. 60 f.

[111] Vgl. zum Beispiel ebd. S. 57–59 und Eisenhardt 1995, S. 154 f. Auch für die Beisitzer des Reichskammergerichts wurde in zunehmendem Maße juristische Bildung vorgeschrieben (vgl. Laufs 1984, S. 62).

[112] Vgl. zum spätmittelalterlichen Verständnis vom *Sachsenspiegel* Kroeschell 1998, insbesondere S. 88 f.

[113] Vgl. Schumann 2013, S. 135 (mit Anm. 36).

Auch unter den Autorenwerken lassen sich einige Untergruppen von Texten ausmachen. Wie bei den Normtexten kommt dem Prozessrecht eine erhebliche Bedeutung zu. Zum einen wird es in einer Reihe von darauf fokussierten Werken behandelt,[114] zum anderen ist es verschiedentlich – wie in Land- oder Stadtrechten – einer der Hauptgegenstände umfassenderer Rechtsdarstellungen.[115] Auch andere Themen werden in Einzeldarstellungen behandelt, so zum Beispiel das Erbrecht mit beziehungsweise ohne Testament oder die Vormundschaft.[116]

Viele dieser Werke dienen der Vermittlung von Kenntnissen über das römische Recht. In diesem Zusammenhang sind auch Übersetzungen und Kommentare zu Teilen des *Corpus Iuris Civilis* anzuführen, insbesondere zum darin enthaltenen Anfängerlehrbuch, den *Institutiones*, die von Murner, Fuchsperger und Gobler übersetzt wurden. Aber auch Werke des sächsischen, magdeburgischen und lübischen Rechts wurden herausgegeben, und es finden sich Vergleiche zwischen den Bestimmungen der verschiedenen Rechtssysteme.[117]

Schließlich ist noch auf die Formularbücher, die Notariats- und Kanzleiliteratur hinzuweisen. Diese zum Teil sehr umfangreichen Werke enthalten insbesondere auch viele Formulare, also Textvorlagen. Diese Texte bereiten für die Untersuchung von Abhängigkeitsverhältnissen besondere Probleme. Einerseits ist bei ihnen in besonderem Maße mit Übernahmen zu rechnen, da die Formulare ja gerade dafür gedacht waren, abgeschrieben zu werden, und nur begrenzt Anlass bestand, in einer neuen Sammlung bewährte Mustertexte durch andere zu ersetzen. Andererseits finden sich darin vielfach feststehende Formulierungen, deren wiederholtes Vorkommen in der Regel kein Hinweis auf eine Beziehung zwischen den betreffenden Texten ist, sondern vielmehr einer auf formelhaften Sprachgebrauch insbesondere im Geschäftsschriftgut.

Während einige Texte des Korpus, insbesondere große Rechtsordnungen, schon verschiedentlich untersucht wurden und auch in rechtshistorischen Aufrissen Erwähnung finden, haben die Autorenwerke bisher insgesamt nur recht wenig Interesse auf sich gezogen. Den umfassendsten Überblick über diese Werke bie-

[114] Diese Werke tragen häufig den Titel *Process*, zum Teil mit einem attributiven Zusatz und entsprechend dem damaligen Gebrauch weiteren, oft ausführlichen Erläuterungen. Als Beispiele seien „GErichtlicher Proceß“ (Gobler,GerProz. 1536) und „TEutscher Process“ (Stumphart,Proz. 1541) genannt.

[115] Ein Beispiel hierfür ist Meurer,Liberey(Rücker) 1597 mit einer Gliederung in vier Hauptteile, die der von WürtLR. 1555 und WürtLR. 1567 entspricht – einer der zahlreichen Punkte, an denen Wechselwirkungen zwischen Normtexten und Autorenwerken zu erkennen sind.

[116] Aus dem ausgewerteten Korpus sind hierzu Bemel,TraktTestam. 1587, Schubeus,Erbschaft 1597 und Damhouder,Patrocinium 1576 zu nennen.

[117] So stellt Schwartzkopf,DiffIur. 1586 Unterschiede zwischen dem sächsischen und dem römischen Recht zusammen, und Kolle,LübR. 1586 verweist im Kommentar zu einzelnen Artikeln des lübischen Rechts immer wieder auf Bestimmungen zum Beispiel im *Sachsenspiegel* und im *Corpus Iuris Civilis* hin.

ten auch heute noch zwei Darstellungen des 19. Jahrhunderts, die freilich von einer insgesamt negativen Bewertung der geistigen Leistung dieser Praktikerliteratur[118] geprägt sind: zum einen die *Geschichte der deutschen Rechtsquellen* von Otto Stobbe, deren hier relevante zweite Abteilung 1864 erschien, zum anderen die 1867 publizierte *Geschichte der populären Literatur des römisch-kanonischen Rechts in Deutschland am Ende des fünfzehnten und im Anfang des sechszehnten Jahrhunderts* von Roderich Stintzing. So heißt es in Stobbes einleitenden Bemerkungen zu den im entsprechenden Abschnitt aufgeführten Schriften:

> Bei den Gelehrten stand diese Art von Schriftstellerei von vorne herein in Missachtung; sie erklärten sich gegen die Popularisirung der Rechtswissenschaft, da sie nur in den seltensten Fällen auf eine gesunde und brauchbare Weise die fremde Rechtskenntniss verbreite, regelmässig nur Oberflächlichkeit und Scheinwissen befördere.[119]

Auch wenn dieses Zitat den Eindruck erweckt, nur zeitgenössische Äußerungen zusammenzufassen, dürfte es sich doch primär um Stobbes eigene Sicht handeln.[120] Auch zu den einzelnen von ihm in diesem Zusammenhang vorgestellten Werken gibt er überwiegend kritische Kommentare, wobei er freilich differenziert.[121]

Stintzing betont in der Einleitung zu seinem Buch zwar die Bedeutung der Praktikerliteratur für die Rezeption des römischen Rechts, zugleich aber auch die „Mittelmäßigkeit“[122] ihrer Verfasser:

> So ward das römische Recht, nicht wenig verstümmelt und verunstaltet, von plumpen Händen in der zweiten Hälfte des fünfzehnten Jahrhunderts unter das Volk gebracht, um vielfach Aergerniß zu geben und dennoch sein dauerndes Besitzthum zu werden, dessen vollen Werth erst eine spätere Zeit herausbilden sollte.[123]

Trotz dieser kritischen Einschätzung erklärt er aber, dass diese Werke durch gute Strukturierung und einprägsame Darstellung die Verbreitung von gewissen Kenntnissen über das römische Recht – wenn auch ohne tiefere Durchdringung – ermöglicht hätten.[124]

---

[118] Diese Bezeichnung stammt von Eva Schumann, die darauf hinweist, dass das mangelnde Interesse der Rechtshistoriker an diesen Schriften wohl durch Stintzings Klassifikation als „populäre Literatur“ zu erklären sei (vgl. Schumann 2013, S. 138 f.).

[119] Stobbe 1864, S. 167.

[120] Der Nachweis für die allgemeine Behauptung beschränkt sich auf einen Hinweis auf eine Äußerung von Zasius.

[121] So bewertet er den *Laienspiegel* im Vergleich zum *Klagspiegel* recht positiv, und die Werke Andreas Perneders zählt er zu den „vergleichsweise [...] besten Schriften der popularisirenden Richtung“ (Stobbe 1864, S. 173). Immer wieder konstatiert er aber geistige Unselbständigkeit und fehlendes Durchdringen der dargestellten Materie sowie wörtliche Übernahmen. Vgl. zum Beispiel zum *Klagspiegel* ebd. S. 170, zu Noe Meurers *Liberey* ebd. S. 177 und besonders kritisch zu Justin Gobler ebd. S. 174–176.

[122] Stintzing 1867, S. XXIII.

[123] Ebd. S. XXIII.

[124] Ebd. S. XXXVIII.

Stintzing beschränkt seine Darstellung auf die Zeit bis zum Anfang des 16. Jahrhunderts. Den erstmals 1509 erschienenen *Laienspiegel* sieht er als „historischen Abschluß“ und „Becken, in welchem sich die vielen kleinen, von jener literarischen Fluth gebildeten, Rinnsale sammelten.“[125] Seiner Behauptung, dass es danach keine weiteren Schriften dieser Art mehr gegeben habe und auch die älteren bis auf den *Klagspiegel* kaum mehr wieder aufgelegt worden seien,[126] kann man allerdings wohl allenfalls hinsichtlich des zweiten Teils zustimmen. Tatsächlich sind von den erstmals vor 1509 erschienenen Autorenwerken des Korpus von DRQEdit nicht allzu viele später erneut gedruckt worden.[127] Die Zahl der der Praktikerliteratur zuzurechnenden älteren deutschsprachigen Werke ist allerdings ohnehin klein – Stintzings Darstellung bezieht sich zu erheblichen Teilen auf lateinische Werke –, und dass nicht alle von ihnen auf fortdauerndes Interesse gestoßen sind, sollte nicht weiter erklärungsbedürftig sein, zumal eben – anders als Stintzings Darstellung vermuten lässt – in der Folgezeit eine Vielzahl weiterer Schriften mit einer ähnlichen Ausrichtung erschien.[128]

Mit Hilfe der Darstellungen von Stobbe und Stintzing kann immerhin ein gewisser Gesamteindruck von der deutschsprachigen juristischen Praktikerliteratur des 15. und 16. Jahrhunderts gewonnen werden, wenn auch jedenfalls die von Stobbe vorgenommenen Bewertungen sicherlich vielfach fragwürdig sind. Untersuchungen, die über den damit erreichten Kenntnisstand hinausgehen, hat es dann wohl lange Zeit nicht mehr gegeben. Erst in den letzten Jahren finden die Schriften der Rechtspraktiker wieder vermehrtes Interesse. Im Hinblick auf den *Klag*- und den *Laienspiegel*, die beiden wichtigsten Werke der Frühzeit, sowie auf die *Rethorica* von Alexander Hugen ist insbesondere auf Publikationen von Andreas Deutsch hinzuweisen.[129] Eva Schumann strebt die Erarbeitung eines Gesamtüberblicks an und betont die Bedeutung der „Rechtsexperten“ und der von ihnen verfassten Werke für die Rezeption des römischen Rechts und die Entwicklung der deutschen

[125] Ebd. S. XLVII.

[126] Ebd. S. XLVIII.

[127] Zu nennen sind etwa Gessler,Form. 1493 mit weiteren Auflagen von 1502, 1507, 1511, 1514 und 1519, Riederer,Rhetorik 1493 mit weiteren Auflagen von 1505, 1509, 1517 und 1535 und OrdoJudiciarius(dt.) 1472 mit einer Reihe von Inkunabelausgaben sowie weiteren Auflagen von 1505, 1506, 1507, 1512, 1515, 1517, 1530 und 1531 (vgl. jeweils die Übersichten zu weiteren Ausgaben in DRQEdit).

[128] Stintzing modifiziert seine Aussage zwar im Folgenden noch im Hinblick auf Schriften „zur theoretischen Einleitung“ (Stintzing 1867, S. XLVIII). Aber auch wenn möglicherweise nach seinem Verständnis nicht alle Autorenwerke, die in DRQEdit Berücksichtigung finden, der „populären Literatur“ zuzurechnen sein sollten (vgl. zur Abgrenzungsproblematik ebd. S. LI), lässt sich doch zumindest thematisch wohl kein wesentlicher Unterschied feststellen etwa zwischen der Kanzlei- und Notariatsliteratur oder den Darstellungen zum Gerichtsprozess, die in späteren Jahren publiziert wurden, und den von ihm behandelten Schriften. Vgl. gegen Stintzings Darstellung des *Laienspiegels* als Endpunkt der Entwicklung der „populären Literatur“ auch Schumann 2013, S. 139.

[129] Vgl. Deutsch 2004, Deutsch 2008 und Deutsch (Hg.) 2011.

Rechtssprache.[130] In ähnlicher Weise ist auch das Projekt DRQEdit ausgerichtet, allerdings fokussiert auf die Präsentation und Erschließung der Texte, wobei, wie schon gesagt, nicht nur die Werke der Rechtspraktiker, sondern auch die Normtexte der Zeit berücksichtigt werden.

Das Angebot von DRQEdit umfasst drei Ebenen: Informationen zu den Texten, bis auf wenige Ausnahmen digitale Faksimiles (Imagedigitalisate) der zum Korpus gehörenden Drucke sowie nach Möglichkeit strukturierte und durchsuchbare Volltexte.

Das Quellenverzeichnis enthält natürlich die bibliographischen Daten zur jeweiligen Ausgabe. Durch eine Verlinkung zu den entsprechenden Einträgen im *Incunabula Short Title Catalogue (ISTC)* und zum *Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts (VD 16)* können auch die dort verzeichneten umfassenderen Informationen, etwa zu vorhandenen Exemplaren, leicht eingesehen werden; ebenso sind auch die den Autoren zugeordneten Datensätze in der *Gemeinsamen Normdatei (GND)*[131] entsprechend den Angaben im VD16 eingebunden.

Darüber hinaus enthält das Quellenverzeichnis von DRQEdit nach Möglichkeit noch weitere Informationen, die der Einordnung und inhaltlichen Erschließung der Texte dienen sollen. So werden gegebenenfalls in der wissenschaftlichen Literatur übliche Kurztitel für die Texte genannt, und es wird auf Editionen und wissenschaftliche Literatur hingewiesen.

Insbesondere werden auch weitere Ausgaben bis zum Ende des 16. Jahrhunderts – soweit sie anhand der Kataloginformationen als solche zu erkennen sind – mit Erscheinungsort und -jahr aufgelistet und mit den eben genannten Katalogen sowie gegebenenfalls mit online verfügbaren Digitalisaten verlinkt. Schon anhand dieser Auflistung lässt sich ein gewisser Eindruck von der Rezeption eines Textes gewinnen.

Ein besonderes Anliegen des Projekts ist die Verdeutlichung von Zusammenhängen zwischen den Texten. Deshalb gehört zu den Informationen im Quellenverzeichnis des Öfteren auch ein Hinweis auf Abhängigkeitsverhältnisse, insbesondere wenn diese recht eng und umfassend sind. Häufig beruht diese Beziehung darauf, dass eine Rechtsordnung einer Revision unterzogen wurde oder dass ein Autor ein Werk überarbeitete. Hierbei ergeben sich aber natürlich Abgrenzungsproble-

[130] Vgl. Schumann 2007 (zum Begriff „Rechtsexperten“ ebd. S. 448) und Schumann 2013, außerdem Kramer 2010. In diesem Zusammenhang ist auch auf mehrere von Schumann betreute Dissertationen und Dissertationsprojekte hinzuweisen, die sich mit den Rechtspraktikern und Rechtsexperten der Frühen Neuzeit befassen, vgl. *http://www.deutsche-rechtsgeschichte.uni-goettingen.de/promotionen/abgeschlossene.html* und *http://www.deutsche-rechtsgeschichte.uni-goettingen.de/promotionen/laufende.html.*

[131] Vgl. *http://www.dnb.de/gnd.*

me zu weiteren Ausgaben eines Textes – auf die Problematik wurde oben schon hingewiesen.

Mit der Präsentation der Drucke des Korpus als Faksimiles bietet DRQEdit eine virtuelle Bibliothek. Das Projekt nutzt dabei die von verschiedenen wissenschaftlichen Bibliotheken erstellten Digitalisate, erschließt sie aber zusätzlich durch ein Adressierungsschema, das nach Möglichkeit auf den originalen Seitenbezeichnungen aufbaut und die Drucke über Siglen bezeichnet, die nach inhaltlichen Kriterien gebildet sind. So beginnen die Siglen von Autorenwerken in der Regel mit dem Nachnamen des Autors (oder des Übersetzers beziehungsweise Bearbeiters) und die Siglen von Normtexten mit einer (möglicherweise abgekürzten) Bezeichnung des betreffenden Territoriums. Für einen Kenner sollte damit im Normalfall recht unmittelbar erkennbar sein, worum es sich jeweils handelt.

Die zentrale Aufgabe des Projekts DRQEdit ist die Erschließung der Texte durch die Erstellung und Aufbereitung durchsuchbarer Volltexte. Dabei wird eine weitestgehend buchstabengetreue Repräsentation angestrebt, die vielfachen Kürzungszeichen (bis auf den Abkürzungspunkt) werden allerdings aufgelöst und (nach dem derzeitigen Stand) durch geschweifte Klammern gekennzeichnet.[132]

Wie die Faksimiles können auch die transkribierten Volltexte entsprechend der originalen Seitengliederung angezeigt werden. Auch eine Parallelanzeige ist möglich, in der die transkribierte Fassung mit originalem Zeilenumbruch dargestellt wird und daneben das Faksimile, so dass gegebenenfalls leicht überprüft werden kann, ob die Transkription korrekt ist.

Darüber hinaus gibt es für die meisten Texte auch eine Anzeige nach Abschnitten entsprechend ihrer originalen Gliederung. Dadurch können zusammenhängende Textstücke als Einheit präsentiert werden, und zudem ist eine Adressierung entsprechend der Textstruktur möglich, die im Idealfall auf einer originalen Zählung basiert.

In manchen Fällen fehlt eine solche Zählung in der Vorlage, aber es gibt eine etablierte systematische Zitierweise. Ein Beispiel hierfür sind die Reichsabschiede, bei denen auch in neueren Editionen üblicherweise die Paragraphenzählung der *Neuen und vollständigeren Sammlung der Reichs-Abschiede* von 1747[133] zugrunde gelegt wird.[134] Allerdings stimmt die dabei vorgenommene Strukturierung nicht immer mit der der Erstausgabe überein. Da in DRQEdit diese Erstausgabe Grundlage für die Edition ist, basiert die Gliederung auf der Absatzeinteilung der Vorlage;

[132] Es werden keine eckigen Klammern verwendet, da es sich nicht im eigentlichen Sinne um Zusätze handelt – im Original sind ja stattdessen die Kürzungszeichen zu finden. Kursivdruck dient in DRQEdit zur Wiedergabe von Antiqua und kann deshalb nicht zur Kennzeichnung der aufgelösten Abkürzungen verwendet werden.

[133] NSRA.

[134] Vgl. etwa RTA.JR Bd. 5/6 (2011), S. 198 und Bd. 16 (2003), S. 1657.

zugleich wird aber nach Möglichkeit auch die etablierte Zählung als Basis für eine Bezeichnung von Textstücken angeboten. Damit ist es zum Beispiel leicht möglich, Textstellen, die in der wissenschaftlichen Literatur über eine solche Zählung bezeichnet werden, in DRQEdit zu finden.[135]

Zwar kann jedenfalls im Moment nicht jeder Text so strukturiert werden, dass eine dauerhaft stabile Adressierung auf dieser Basis gewährleistet werden kann, und in den übrigen Fällen wird keine Anzeige nach Abschnitten angeboten. Auch dann steht aber so wie sonst auch – sofern Überschriften vorhanden sind – eine Gliederungsübersicht zur Verfügung, in der Links von den Überschriften zu den betreffenden Textstücken beziehungsweise zur jeweiligen ersten zugehörigen Druckseite führen.

Die Anzeige erfolgt in der Regel entsprechend der originalen Absatzgliederung.[136] Als Alternative besteht aber die Möglichkeit, jeden einzelnen Satz beziehungsweise längeren Teilsatz als eigenen Absatz darzustellen. Dies ist als Lesehilfe gedacht, da die frühneuhochdeutschen Satzkonstruktionen teilweise nur schwer zu überblicken sind. Die Unterteilung beruht freilich weitgehend auf der originalen Zeichensetzung und konnte nur in geringem Umfang nachbearbeitet werden. Sie dient gleichzeitig auch als Basis für die Ergebnisanzeige bei der Suche nach bestimmten Wortformen oder Wortverbindungen.

Im vorliegenden Zusammenhang ist insbesondere darauf hinzuweisen, dass DRQEdit für einige Textpaare eine synoptische Anzeige bietet. Dies betrifft vor allem Fälle, in denen es sehr weitgehende wörtliche Übereinstimmungen gibt, wie etwa zwischen der Bambergischen und der Brandenburgischen Halsgerichtsordnung sowie der *Peinlichen Gerichtsordnung* Karls V. Darüber hinaus steht eine synoptische Anzeige auch für die drei zum Korpus von DRQEdit gehörenden Übersetzungen der *Institutiones* zur Verfügung. Nach dem derzeitigen Stand beschränkt sich die Datenaufbereitung für die Synopse auf eine Zuordnung von Abschnitten, die jeweils auch als Anzeigeeinheiten in der Einzeldarstellung dienen. Es wird also keine Feinparallelisierung auf einer tieferen Ebene vorgenommen,[137] und insbesondere erfolgt keine Hervorhebung von Übereinstimmungen beziehungsweise Unterschieden. Auch auf dieser Basis sollte es aber schon recht gut möglich sein, einen Vergleich zwischen solchen Textpaaren mit hoher Ähnlichkeit durchzuführen.

[135] Ein Beispiel ist *http://drw-www.adw.uni-heidelberg.de/drqedit-cgi/zeige?index=sonstige_zaehlung&sigle=Fuchsperger%2CInst.+1536&zaehlung_nach=CICiv.&term=I.3.7* – die URL basiert auf der heutigen Zählung des *Corpus Iuris Civilis*, Fuchsperger hingegen legt eine auch hinsichtlich der Titeleinteilung etwas abweichende Gliederung zugrunde.

[136] Eine möglichst exakte Wiedergabe des Druckbildes wird dabei nicht angestrebt. So wird nicht zwischen den verschiedenen Formen der Absatzgliederung, etwa durch Einzug der ersten Zeile oder durch einen vertikalen Abstand, unterschieden.

[137] Bei den Übersetzungen der *Institutiones* ist allerdings jeweils zusätzlich die Paragrapheneinteilung nach der heute als Standard zu betrachtenden Ausgabe CICiv.(Krüger/Mommsen) vermerkt, so dass sich bei ihnen auch für diese kleineren Texteinheiten recht gut erkennen lässt, was in den nebeneinander gezeigten Texten einander zuzuordnen ist.

1.2.2 Das ausgewertete Korpus

Das Korpus von DRQEdit umfasst nach dem derzeitigen Stand etwa 450 Drucke mit etwa 90.000 Druckseiten.[138] Für die vorliegende Untersuchung konnten fast alle bisher transkribierten Texte, insgesamt 176 Drucke mit etwas weniger als 27.000 Druckseiten und etwa 6.680.000 laufenden Wortformen, berücksichtigt werden.[139]

Anders als der Titel dieser Untersuchung vermuten lässt, sind darunter auch einige mittelniederdeutsche Texte, die allerdings gegenüber den frühneuhochdeutschen Texten kaum ins Gewicht fallen. Sie sind bei der Untersuchung nicht ausgeklammert worden, da die beiden Sprachen einander so nahe stehen, dass das hier entwickelte Verfahren zumindest ansatzweise geeignet ist, Übereinstimmungen auch bei einem Sprachtransfer vom Mittelniederdeutschen ins Hochdeutsche zu ermitteln. Allerdings ist festzuhalten, dass die zugrunde gelegten Verarbeitungsregeln für die Varianz des Frühneuhochdeutschen konstruiert wurden und die lautlichen Unterschiede zum Mittelniederdeutschen nicht hinreichend berücksichtigt sind. Dementsprechend konnte mit dem Verfahren eine Reihe von Übereinstimmungen zwischen dem mittelniederdeutschen Eiderstedter Landrecht von 1572 und der revidierten hochdeutschen Fassung von 1591 ermittelt werden, es muss hier aber offen bleiben, wie stark die Erkennung durch den Sprachwechsel beeinträchtigt ist.

Innerhalb des ausgewerteten Korpus lassen sich verschiedene Untergruppen ausmachen, deren Glieder strukturelle und/oder thematische Ähnlichkeiten aufweisen und in nicht wenigen Fällen auch zumindest teilweise durch wörtliche Übernahmen erkennbar miteinander verbunden sind. Im vorigen Unterkapitel wurde schon ein erster Überblick über Textgruppen innerhalb des Gesamtkorpus von DRQEdit gegeben. Hier soll das Bild zum Beispiel im Hinblick auf zugehörige wichtige Texte etwas konkreter werden. Zugleich geht es dabei auch um wörtliche Übereinstimmungen und Abhängigkeitsverhältnisse innerhalb der jeweils betrachteten Gruppen.

Zunächst sollen die Normtexte des Untersuchungskorpus betrachtet werden.[140] Insbesondere die Gerichtsordnungen, die im Korpus von DRQEdit sehr zahlreich vertreten sind, bilden auch im hier untersuchten Teilkorpus eine große Gruppe.[141] Je nach Art des Gerichts, um das es jeweils geht, ist die Bezeichnung *Gerichtsordnung* häufig durch ein weiteres Bestimmungswort näher spezifiziert. So gibt es zum

[138] Wie oben schon festgestellt wurde, ist das Korpus durch inhaltliche Kriterien bestimmt, und es kann im Laufe der Bearbeitung noch zu gewissen Erweiterungen kommen.

[139] Das ausgewertete Korpus ist unten im Quellenverzeichnis dokumentiert (vgl. dazu die Hinweise auf S. 291).

[140] Bei der bisherigen Texterfassung für DRQEdit wurden die Normtexte bevorzugt berücksichtigt, um für diese rechtshistorisch oft bedeutsamen Quellen frühzeitig eine Volltextsuche sowie eine Adressierung entsprechend der Textgliederung anbieten zu können.

[141] Die Angabe einer genauen Zahl ist hier wie auch bei der folgenden Beschreibung der Textgruppen des Korpus wohl nicht sinnvoll, da es durchaus nicht immer einfach ist, einen Text einer bestimmten Kategorie zuzuordnen.

Beispiel *Unter-*, *Hof-* und *Oberhofgerichtsordnungen* entsprechend der Zuständigkeit für bestimmte Personengruppen beziehungsweise der Instanz sowie *Hals-* oder *peinliche Gerichtsordnungen* für das Straf- und Strafverfahrensrecht.

Als zentrale Texte für letztere Gruppe sind die Bambergische Halsgerichtsordnung von 1507 (die *Constitutio Criminalis Bambergensis*) sowie die *Peinliche Gerichtsordnung* Karls V. aus dem Jahr 1532 (die *Constitutio Criminalis Carolina*) zu nennen. Bekanntlich steht die *Carolina* in einem recht engen Abhängigkeitsverhältnis zur *Bambergensis*; beide Texte liegen in einer synoptischen Edition[142] vor. Schon vor dieser Rezeption im Reichsrecht wurde der Text der Bambergischen Halsgerichtsordnung im Jahr 1516 fast unverändert als sogenannte *Brandenburgische Halsgerichtsordnung* für die ehemals zur Burggrafschaft Nürnberg gehörenden Gebiete „vnter vnd oberhalb des Gebůrgs“[143] übernommen.[144] Wesentlich stärker sind aber sicherlich die über die *Carolina* vermittelten textuellen Traditionslinien. Die *Peinliche Gerichtsordnung* wurde übernommen als Teil der *Reformation* des Erzstifts Köln von 1538 (mit abweichender Artikelzählung) und weitgehend unverändert als Hessische Halsgerichtsordnung von 1535.[145] Außerdem zeigen verschiedene weitere Texte des Korpus nach Aussagen in der wissenschaftlichen Literatur eine Abhängigkeit von der *Carolina*; im Hinblick auf das Untersuchungskorpus ist insbesondere die steirische *Landt vnd Peindlich Gerichts Ordnung* aus dem Jahr 1574 zu erwähnen.[146]

Für die wesentlich zahlreicheren Gerichtsordnungen zur Regelung des Zivilprozesses ergibt sich kein so klares Bild der Abhängigkeitsverhältnisse, allerdings lassen sich auch bei ihnen Texte mit stärkerem Einfluss ausmachen. So ist bekannt, dass die Mainzer Hofgerichtsordnung von 1516 und die Mainzer Untergerichtsordnung von 1534 vielfach rezipiert wurden.[147] Außerdem finden sich bei den Gerichtsordnungen einige revidierte Fassungen mit vielfach starker Übereinstimmung mit der Vorgängerversion sowie kleinere Textgruppen mit ausgeprägten Übereinstimmungen, die durch territoriale Zusammenhänge erklärlich sind. So sind

[142] Zoepfl (Hg.) 1883.

[143] BrandenbAnsbHalsGO. 1516, Art. 1 (Bl. 1 v). Vgl. zur historischen Entwicklung dieser Territorien zum Beispiel *https://de.wikipedia.org/wiki/Burggrafschaft_N%C3%BCrnberg*.

[144] Vgl. außerdem zur Nutzung der *Bambergensis* durch Gerichte anderer Territorien Schmidt 1965, S. 130 f.

[145] Letzterer Text gehört nicht zum hier ausgewerteten Korpus. Laut Schmidt 1965, S. 142 handelt es sich dabei um „eine fast wörtliche Wiedergabe“ der *Carolina*.

[146] Vgl. Byloff 1907, S. 89–99 sowie 105–108, Schmidt 1965, S. 142 f. und Pauser 2004, S. 227. Dass Schmidt dabei die Pommersche Hofgerichtsordnung zusammen mit der *Kölner Reformation* im Hinblick auf wörtliche Anlehnungen nennt, scheint allerdings nicht korrekt: Der entsprechende Teil der *Kölner Reformation* ist vielmehr eine umfassende Übernahme, die Pommersche Hofgerichtsordnung hingegen ist keine Halsgerichtsordnung, so dass wohl wenig Anlass bestanden haben dürfte, dafür die *Peinliche Gerichtsordnung* als Vorlage zu verwenden, und tatsächlich wird für dieses Textpaar jedenfalls mit dem hier verwendeten Verfahren keine Übereinstimmung gefunden, die als Anzeichen für eine Beziehung zwischen diesen Texten gedeutet werden könnte.

[147] Vgl. Schwartz 1898, S. 37, Marquordt 1938, S. 11–15 und Otte 1964, S. 77–96.

die Hofgerichtsordnungen von Braunschweig-Wolfenbüttel und Braunschweig-Lüneburg eng miteinander verbunden[148] und in ähnlicher Weise auch die sächsischen Gerichtsordnungen[149]. Auch für den Zivilprozess kommt dem Reichsrecht, in diesem Fall den Reichskammergerichtsordnungen, wohl eine prägende Rolle zu,[150] als Textvorlage haben sie aber nach den Auswertungsergebnissen dieser Untersuchung anscheinend nur punktuell gedient.

Das Prozessrecht bildet auch in den Stadtrechtsreformationen und den Landrechten sowie teilweise auch den Landesordnungen einen inhaltlichen Schwerpunkt. Für erstere ist auf die Reformationen von Nürnberg (1479, 1503, 1522 und 1564), Worms (1498) und Frankfurt (1509 und 1578) sowie auf das Freiburger Stadtrecht von 1520 hinzuweisen.[151] Unter den wenigen zum Untersuchungskorpus gehörenden Inkunabeln sind damit – als umfangreichste Texte – zwei Stadtrechtsreformationen vertreten. Die großen Landrechte stammen aus dem 16. Jahrhundert. Hier sollen als Beispiele die Bayrische Landrechtsreformation von 1518, die Württembergischen Landrechte von 1555 und 1567 sowie das Pfälzische Landrecht von 1582 genannt werden.

Landesordnungen umfassen zwar vielfach Themen, die auch in Landrechten behandelt werden,[152] sind aber oft insgesamt thematisch weiter angelegt und behandeln auch policeyrechtliche Themen, also Fragen der öffentlichen Ordnung, und teilweise auch das Strafrecht, während Landrechte in der Regel das bürgerliche Recht und den Zivilprozess zum Gegenstand haben. Landesordnungen mit thematischen Überschneidungen zu Landrechten beziehungsweise Gerichtsordnungen sind zum Beispiel die Tiroler Landesordnungen von 1526, 1532 und 1573 sowie die Henneberger Landesordnung von 1539.

Die Stadtrechtsreformationen sind rechtshistorisch vielfach untersucht worden, und auch zu den großen Landrechten und Landesordnungen gibt es teilweise umfangreiche Studien.[153] Stark ausgeprägte Abhängigkeitsverhältnisse zwischen den hier betrachteten Texten – abgesehen von revidierten Fassungen und ihren Basistexten – werden darin anscheinend nur selten verzeichnet,[154] und auch die in

[148] Vgl. Theuerkauf 1968, S. 241.

[149] Vgl. Lück 1997, S. 139–142.

[150] Vgl. zum Beispiel Otte 1964, S. 84 über den Einfluss der Reichskammergerichtsordnung von 1548/1555 auf die Ordnungen für höhere Gerichte.

[151] Die Bezeichnung als *Reformation* entstammt (teilweise in anderer Schreibung) den genannten Texten; das in der wissenschaftlichen Literatur gängige Kompositum *Stadtrechtsreformation* dient der Vermeidung von Missverständnissen.

[152] Andere Landesordnungen klammern diese Themen offenbar aus und stehen teilweise neben einem Landrecht für dasselbe Territorium (so zum Beispiel PfalzLO. 1582 neben PfalzLR. 1582).

[153] Vgl. die Literaturhinweise im Quellenverzeichnis von DRQEdit.

[154] Als Ausnahme ist die Beziehung zwischen der Tiroler Landesordnung von 1532 (TirolLO. 1532) und der Henneberger Landesordnung von 1539 (HennebergLO. 1539) zu nennen, die aber in der Literatur offenbar übertrieben wird – vgl. dazu unten Unterkapitel 4.3.2.

der vorliegenden Untersuchung angewandten Methoden zeigen wohl keine Übernahmen ganzer Texte. Dies mag darauf zurückzuführen sein, dass das materielle Recht stärker von lokalen Rechtstraditionen geprägt blieb als das stark römischrechtlich umgestaltete Prozessrecht und dass dementsprechend die Ordnungen anderer Territorien oder Städte nicht einfach mehr oder weniger vollständig kopiert werden konnten. Allerdings gibt es durchaus Abhängigkeiten zwischen bestimmten Textpassagen – Kapitel 4.3 bietet dafür Beispiele.[155]

Im Hinblick auf die Autorenwerke des Untersuchungskorpus ist zunächst einmal darauf hinzuweisen, dass ihre Zahl zwar wesentlich kleiner ist als die der Normtexte, dass darunter aber die umfangreichsten Texte des Korpus sind und die gesamte hier ausgewertete Textmenge in etwa der der Normtexte entspricht.

Übersetzungen beziehungsweise Editionen älterer Texte, teilweise mit Kommentar, sind für die eben angestellte Berechnung als Autorenwerke eingeordnet worden, da sie nicht der zeitgenössischen Gesetzgebung zuzurechnen sind, sondern vielmehr ihre Erstellung beziehungsweise Publikation in den Zusammenhang der Wissensvermittlung und der Beschäftigung mit den überkommenen Rechtsquellen einzuordnen ist.

Wörtliche Übernahmen aus den Übersetzungen sind in Texten, wie sie hier untersucht werden, wohl unwahrscheinlich, da als Referenztext das lateinische Original diente und bei Verweisen darauf in aller Regel nur eine Stelle angegeben, nicht aber das betreffende Textstück zitiert wurde.[156] Auch zwischen den drei im Korpus enthaltenen Übersetzungen der *Institutiones* lassen sich mit dem hier angewandten Verfahren nur wenige Übereinstimmungen finden, und diese sind sicherlich nicht auf eine Abhängigkeit zwischen den Übersetzungen zurückzuführen, sondern vielmehr darauf, dass bei Mehrfachübersetzung eines Textes wohl fast zwangsläufig immer wieder auch ähnliche oder in begrenztem Umfang sogar gleiche Formulierungen gewählt werden.

[155] Die Übernahmen können auch größere Textteile betreffen – ein Beispiel hierfür ist die (trotz zahlreicher Unterschiede) ziemlich ausgeprägte Abhängigkeit des zweiten Teils von WürtLR. 1555 von FreiburgStR. 1520.

[156] Die zahlreichen Ausgaben des *Sachsenspiegels* und oft auch weiterer dem sächsischen Recht zuzuordnender Texte – die größte Gruppe von Ausgaben mittelalterlicher deutscher Rechtsquellen im Korpus von DRQEdit – sind im hier untersuchten Korpus nicht vertreten. Ein Vergleich der Texte dieses Korpus mit einer modernen Edition des mittelniederdeutschen *Sachsenspiegel*-Textes ergab allerdings nur eine relativ kleine Zahl von Übereinstimmungen. Inwieweit dies auf die sprachlichen Unterschiede zwischen Mittelniederdeutsch und Frühneuhochdeutsch zurückzuführen ist, inwieweit auf die Zusammenstellung des Korpus (das die dem *Sachsenspiegel* sehr nahestehenden mittelalterlichen Texte nicht enthält und auch nur wenige Texte des 16. Jahrhunderts, die sich auf das sächsische Recht beziehen) und inwieweit darauf, dass im Untersuchungszeitraum auch auf den *Sachsenspiegel* in einer ähnlichen Weise wie auf das *Corpus Iuris Civilis* vielfach über Stellenangaben Bezug genommen wurde, ohne den Text zu zitieren, muss hier offen bleiben.

Neben den Übersetzungen und Editionen, die primär der Zugänglichmachung von Texten dienen und allenfalls begrenzt an den praktischen Bedürfnissen der Leser ausgerichtet sind, stehen Werke, in denen es gerade um die Wissensvermittlung für die Rechtspraxis beziehungsweise für den Schriftverkehr geht.

Besonders ausgeprägt ist die Aufbereitung für die unmittelbare Anwendung in den Formularbüchern, die Textvorlagen für eine Vielzahl von Anlässen – zum Beispiel für Briefe, Verträge oder Schriftstücke im Rahmen des Gerichtsverfahrens – enthalten, teilweise ergänzt um Erläuterungen. Neben etwas kleineren Werken dieser Art enthält das Untersuchungskorpus mit der 1528 erstmals publizierten *Rethorica* von Alexander Hugen und dem 1568 erschienenen *Groß Formular* von Johann Peter Zwengel zwei sehr umfangreiche Formularbücher. In ähnlicher Weise finden sich aber auch in anderen Autorenwerken und in Normtexten Mustertexte. So besteht der Justin Gobler zugeschriebene *Gerichtliche Proceß* von 1536 neben Erläuterungen zum Prozessrecht zu erheblichen Teilen aus Formularen für das Gerichtsverfahren, und die Mainzer Untergerichtsordnung von 1534 enthält nach der inhaltlichen Darstellung eine Reihe von Formularen für Klagen und Urteile. Formularbücher wiederum (beziehungsweise die Texte, die man heute in dieser Weise zusammenfasst) sind nicht notwendigerweise nur auf den Abdruck von Mustertexten beschränkt. So enthalten die genannten Werke von Hugen und Zwengel jeweils zu Beginn auch Erläuterungen zur Rhetorik.

Wie schon erwähnt, ist die Sprache der Formulare stark von feststehenden Wendungen bestimmt, so dass sich zwischen den Formularbüchern eine enorme Zahl kürzerer Übereinstimmungen feststellen lässt, die aber meist nicht auf einigermaßen direkte Beziehungen zwischen den Werken zurückzuführen sind.

Auch wenn sich zwischen den Formularbüchern des hier ausgewertete Korpus wohl allenfalls wenige im Hinblick auf solche Beziehungen aussagekräftige Übereinstimmungen ermitteln lassen, ist davon auszugehen, dass der Vergleich mit dem Gesamtbestand der Texte von DRQEdit im Hinblick auf die Ermittlung von Abhängigkeiten zwischen den Formularbüchern recht ergiebig wäre. So ist für einzelne Texte durchaus bekannt, dass sie teilweise auf der – möglicherweise adaptierenden – Übernahme von Formularen und Erläuterungen aus älteren derartigen Schriften beruhen,[157] die Erfahrung im Projekt DRQEdit hat aber gezeigt, dass ein Vergleich des Inhalts solcher Sammlungen von zahlreichen Einzeltexten – den enthaltenen Formularen – insbesondere bei Umstellungen ohne technische Unterstützung ausgesprochen mühselig ist, so dass etwa eine Untersuchung, welche Formulare langfristig Verwendung fanden und welche Formulierungen beziehungsweise Fallkonstellationen dabei ausgetauscht wurden, in größerem Umfang erst

[157] Vgl. Deutsch 2008, S. 48–65 zu Quellen und Vorlagen von Hugen,Rhetor. 1528 sowie ebd. S. 68–72 beziehungsweise Stobbe 1864, S. 161 f. zum Verhältnis dieses Textes zu Notariatbuch 1534 und zu weiteren Ausgaben des letzteren Werks unter verschiedenen Titeln und mit Umstellungen und Erweiterungen.

aufgrund von Vergleichsmöglichkeiten, wie sie hier vorgestellt werden, realistisch erscheint.

Während sich für Formularbücher vermuten lässt, dass sie auch ohne tiefergehende Beschäftigung mit ihren Inhalten praktisch verwendbar waren, da die Mustertexte im Idealfall einfach übernommen werden konnten und eine Anpassung nur im Hinblick auf Namen, Daten und Ähnliches erforderlich war, lassen sich andere Werke der juristischen Praktikerliteratur wohl als Lehr- und Handbücher beschreiben, die für die Vermittlung von zumindest gewissen Kenntnissen zum Beispiel im Hinblick auf das römische Recht geeignet waren.

In dieser Weise lassen sich Werke mit einem weiten thematischen Spektrum wie der *Klagspiegel* von Conrad Heyden[158] sowie *Der Rechten Spiegel* von Justin Gobler einordnen, aber auch Darstellungen zum gerichtlichen Verfahren und zu anderen Rechtsgebieten. Auch hier lässt sich wie bei den Normtexten feststellen, dass das Verfahrensrecht sowohl in den übergreifenden Darstellungen als auch bei den thematisch engeren Werken besonders häufig und teilweise sehr ausführlich behandelt wird.

Im Hinblick auf Werke Justin Goblers (beziehungsweise Werke, die ihm zugeschrieben werden) findet sich in der wissenschaftlichen Literatur verschiedentlich der Vorwurf des Plagiats; dieses Thema wird unten in Kapitel 4.1 näher untersucht. Daneben lassen sich auch in Werken anderer Autoren sehr umfangreiche Übernahmen feststellen, die bei Anwendung heutiger Kriterien als Plagiate zu bezeichnen wären.[159] Auch Normtexte konnten in ähnlicher Weise wie private Arbeiten als Quelle für solche Übernahmen dienen, ohne als Vorlage kenntlich gemacht zu sein.[160] Daneben gibt es auch viele Fälle deutlich kenntlich gemachter Zitate und Auszüge bis hin zur vollständigen Wiedergabe von Normtexten in den privaten Publikationen.[161]

Auf einer anderen Ebene lassen sich die privaten Arbeiten auch nach Verfassern gruppieren. Einige Autoren sind mit mehreren Werken vertreten, zum Beispiel Andreas Perneder mit sämtlichen postum publizierten Schriften in der ersten Druckfassung.[162] Das ermöglicht werkübergreifende Untersuchungen zu den einzelnen

[158] Vgl. Deutsch 2004 (zur Verfasserfrage insbesondere S. 79–198).

[159] Hier ist etwa auf ausgeprägte Übereinstimmungen zwischen Gobler,GerProz. 1536 und Stumphart, Proz. 1541 hinzuweisen.

[160] So nennt Lettscher,Notariat 1576, Bl. 6 v zwar zwei Stellen der Reichsnotariatsordnung (RNotariatsO. 1512) mit Seitenangabe, es ist aber nicht zu erkennen, dass im Folgenden bis Bl. 15 r fast der gesamte Text dieser Ordnung weitgehend wörtlich wiedergegeben wird.

[161] So zitiert Andreas Perneder wiederholt zum Beispiel bayrische Rechtsquellen (etwa Perneder, Inst. 1544, Bl. 83 r), Meurer,Liberey(Rücker) 1597 besteht zu erheblichen Teilen aus Auszügen aus verschiedenen Normtexten, und Saur,Fasc. I (1589) und Saur,Fasc. II (1589) (beide nicht zum Untersuchungskorpus gehörig) drucken zahlreiche Gerichts- und sonstige Ordnungen zumeist vollständig ab.

Autoren und insbesondere auch einen Vergleich zwischen den Texten eines Verfassers, bei denen sich teilweise recht erhebliche Übernahmen zeigen – Kapitel 4.2 bietet dafür ein Beispiel.

1.3 Frühneuhochdeutsche Schreibung und Lautung

Nach einer in der gegenwärtigen Germanistik verbreiteten Gliederung steht das Frühneuhochdeutsche als eigene Sprachstufe zwischen dem Mittelhochdeutschen und dem Neuhochdeutschen. Mit diesem Terminus wird dann meist die Sprache des ober- und mitteldeutschen Sprachraums in der Zeit von etwa 1350 bis 1650 bezeichnet.[163]

Das Frühneuhochdeutsche ist vom Mittelhochdeutschen und auch vom Neuhochdeutschen durch eine Reihe von Phänomenen und Charakteristika abzugrenzen. Als übergreifendes Merkmal ist wohl insbesondere das Fehlen beziehungsweise die erst allmählich erfolgende Neuetablierung einer allgemein gültigen Sprachnorm zu nennen. Stattdessen ist das Frühneuhochdeutsche geprägt durch eine Vielzahl nebeneinander stehender Varietäten, im Hinblick auf die geschriebene Sprache namentlich durch Schreibdialekte, Geschäftssprachen und Druckersprachen.[164]

Für die hier untersuchte Fragestellung ist vor allem von Interesse, welche Regeln und welche Varianz in der Schreibweise von Wörtern festzustellen sind. Die aus den verschiedenen Sprachräumen des Deutschen stammenden Texte lassen in ihren Schreibungen häufig eine Prägung durch die jeweilige regionale Aussprache vermuten; allerdings konnten sich zum Teil bestimmte Schreibweisen auch dort etablieren, wo sie nicht mit der Aussprache übereinstimmten, und in vielen Fällen lässt sich aus einer Schreibung nicht sicher auf die Lautung schließen.[165] Neben den durch den jeweiligen Dialekt beeinflussten Unterschieden in der Schreibweise stehen solche, bei denen weniger oder gar nicht von einer lautlichen Differenz aus-

[162] Diese Texte wurden im 16. Jahrhundert noch von zwei weiteren Herausgebern in wohl teilweise nicht unerheblich geänderter Form erneut veröffentlicht. Vgl. die Übersicht in DRQEdit unter *http://drw-www.adw.uni-heidelberg.de/drqedit-cgi/zeige?index=siglen&term=perneder**.

[163] Vgl. zu den Sprachstufengliederungen Roelcke 1998. Ebd. S. 804–811 sind die verschiedenen Gliederungsvorschläge zusammengestellt. Die genannten Eckdaten können natürlich nur der ungefähren zeitlichen Einordnung dienen, wie es bei Epochenabgrenzungen generell gilt und insbesondere bei der Sprachentwicklung, da die allgemein als regelkonform akzeptierte Sprache sich nicht von einem Tag auf den anderen ändert, sondern Neuerungen sich erst durchsetzen müssen, was in verschiedenen Regionen und Bevölkerungsgruppen zu unterschiedlichen Zeiten der Fall sein kann. Vgl. dazu auch ebd. S. 798 f. Auf diese Abgrenzungsproblematik braucht hier nicht weiter eingegangen zu werden, da die für die Untersuchung zugrunde gelegten Texte aus dem 15. und 16. Jahrhundert und damit der Kernzeit des Frühneuhochdeutschen stammen.

[164] Vgl. Reichmann 1989, S. 33 f.

[165] Vgl. zum Verhältnis von Schreibung und Lautung Reichmann/Wegera 1993, S. 13–25. Die Schwierigkeiten, den Schreibungen bestimmte Lautwerte zuzuordnen, werden ebd. S. 32–151 in der Darstellung der einzelnen Vokale und Konsonanten immer wieder betont.

zugehen ist, sondern vielmehr von verschiedenen Schreibtraditionen, die allerdings oft ebenfalls regional verankert sind und die entsprechende Verortung eines Textes ermöglichen.[166]

Im Folgenden sollen einige Punkte zur Charakterisierung von Lautung und Schreibung des Frühneuhochdeutschen angeführt werden, insbesondere auch im Hinblick auf die Entwicklung gegenüber dem Mittelhochdeutschen. Für den Vokalismus sind wohl insbesondere folgende übergreifende Phänomene anzuführen (dabei sind bei Bedarf Lautzeichen mit */.../* und normalisierte Schreibungen mit *<...>* gekennzeichnet; bei den Lautzeichen werden die Langvokale durch einen nachgestellten Doppelpunkt von den Kurzvokalen unterschieden):

- *Neuhochdeutsche Diphthongierung*: Die mittelhochdeutschen Langvokale */i:/*, */u:/* und */ü:/* (*<iu>*) wurden durch die Diphthonge */ae/* (*<ei>*, *<ai>*), */ao/* (*<au>*) und */oe/* (*<eu>*, *<äu>*) ersetzt. Erste entsprechende Schreibungen lassen sich schon um 1100 in Südtirol feststellen. Im Bairischen erreichten die Digraphien schon um 1350, also zu Beginn der frühneuhochdeutschen Periode, einen Anteil von ca. 90 %, in vielen anderen Regionen trat eine ähnliche Entwicklung aber erst deutlich später ein.[167]
- *Mitteldeutsche Monophthongierung*: Die mittelhochdeutschen Diphthonge */ie/*, */uo/* und */üe/* wurden durch die Langvokale */i:/*, */u:/* und */ü:/* ersetzt. Der Lautwandel wird im Wesentlichen in die mittelhochdeutsche Zeit eingeordnet, die Anpassung der Schreibung dagegen erst in die frühneuhochdeutsche Periode, wobei beim */i:/* die alte Schreibung *<ie>* erhalten blieb. Wie der Name schon sagt, lässt sich die lautliche Entwicklung vor allem im mitteldeutschen Sprachraum feststellen, aber im Oberdeutschen wurden die entsprechenden Schreibungen übernommen.[168]
- *Dehnung* und *Kürzung*: Die Vokallänge änderte sich in vielen Fällen. Insbesondere wurden Kurzvokale in offenen Silben (also ohne ein konsonantisches Ende) schon vor Beginn der frühneuhochdeutschen Periode in der Regel gedehnt und (mit weniger Konsequenz) Langvokale in geschlossenen Silben gekürzt.[169]
- *Senkung*: In bestimmten Positionen wurden die Kurzvokale */i/*, */u/* und */ü/* zu */e/*, */o/* und */ö/* verändert. Entsprechende Schreibungen breiteten sich nach frühen Vorläufern im 14. und 15. Jahrhundert vor allem im mitteldeutschen Raum aus und wurden im 16. Jahrhundert zum Teil auch in anderen Regionen übernommen.[170]

[166] Vgl. verstreute Hinweise ebd. S. 38–63 und 84–151 zu den einzelnen Lauten sowie ebd. S. 32–35 zu den regional unterschiedlich verlaufenden Entwicklungen hinsichtlich der Bezeichnung von Vokallänge und Umlaut.
[167] Vgl. ebd. S. 64–67.
[168] Vgl. ebd. S. 67–70.
[169] Vgl. ebd. S. 71–75.
[170] Vgl. ebd. S. 70 f.

- *Entrundung* und *Rundung*: Die Lang- und Kurzvokale *ü* und *ö* sowie die mittelhochdeutschen Diphthonge *öu* (normal weiterentwickelt zu */oe/*, also *<eu>*/*<äu>*) und *üe* wurden vielfach durch *i*, *e*, */ae/* (*<ei>*/*<ai>*) beziehungsweise *ie* ersetzt, und die entstehenden Lautgleichheiten konnten dazu führen, dass umgekehrt Wörter mit originalem *i*, *e*, *ei* beziehungsweise *ie* stattdessen aufgrund einer Hyperkorrektur mit *ü*, *ö* oder *eu* (beziehungsweise den anderen diesen Lauten entsprechenden Schreibweisen) geschrieben wurden. Außerdem kam es insbesondere im Alemannischen und Ostfränkischen oft zu einer tatsächlichen Rundung von */i/*, */i:/*, */e/* und */e:/*.[171]
- *Apokope* und *Synkope*: *e* in finaler Position fehlt relativ oft, wobei die Häufigkeit je nach Dialekt und Funktion des *e* schwankt. Auch in den Präfixen *ge-* und *be-*, innerhalb des Worts und in den Flexionsendungen lässt sich – in Abhängigkeit von verschiedenen Faktoren – recht häufig eine Tilgung des *e* feststellen.[172]
- *Sprossvokal* und *epithetisches e*: Zwischen Vokal und *r* und in Konsonantenclustern kann ein *e* eingeschoben sein, und am Wortende kann ein *e* angehängt sein.[173]
- Die Nebensilben sind teilweise mit anderen Vokalen als *e* gebildet, wobei sich je nach Sprachraum unterschiedliche Tendenzen feststellen lassen.[174] In den Texten des hier untersuchten Korpus ist insbesondere die mitteldeutsche Schreibung mit dem Präfix *vor-* anstelle von neuhochdeutsch *ver-* bemerkenswert.[175]

Für den Konsonantismus sollen folgende Punkte erwähnt werden:

- Die alte Opposition zwischen einfachen und langen (verdoppelten) Konsonanten entfiel, stattdessen dienen verdoppelte Konsonanten im Frühneuhochdeutschen dazu, die Kürze des vorangehenden Vokals zu bezeichnen[176] und stehen häufig auch nach Langvokalen beziehungsweise anderen Konsonanten ohne erkennbare Funktion (*Konsonantenhäufung*).[177]
- *Spirantisierung*: Die Verschlusslaute *b* und *g* wurden (beziehungsweise blieben) in bestimmten Positionen im Mitteldeutschen sowie in großen Teilen des oberdeutschen Raums durch Reibelaute ersetzt (*b* durch ein bilabiales oder labiodentales *w*, *g* durch *j* oder *ch*).
- *Binnendeutsche Konsonantenschwächung* (*Lenisierung*): In weiten Teilen des mittel- und oberdeutschen Raums wurde die Druckstärke der Fortes (*p*, *t* und *k*) in vielen Fällen reduziert. Zudem entfiel bei den Lenes (*b*, *d* und *g*) der Stimmton,

[171] Vgl. ebd. S. 75–77.
[172] Vgl. ebd. S. 79–81.
[173] Vgl. ebd. S. 82 f.
[174] Vgl. ebd. S. 78.
[175] Nach Reichmann/Wegera 1993, S. 78 waren im Mitteldeutschen insbesondere *i*-Schreibungen verbreitet, zum Beispiel *vir-*.
[176] Vgl. ebd. S. 21.
[177] Vgl. insbesondere ebd. S. 95 f., 101, 107 f., 131 f., 147 und 149 zu *tt*, *ck*, *ff*, *tz*, *ll* und *rr*.

so dass die beiden Reihen vor allem im Anlaut lautlich zusammenfallen konnten, soweit die Lenes nicht spirantisiert wurden.[178]
- Die weiterhin bestehende Auslautverhärtung wurde im Schriftbild schon im 14. Jahrhundert oft nicht mehr wiedergegeben, sondern die verschiedenen Formen eines Wortes in ihrem Konsonantenbestand einander angeglichen (*morphologisches Schreibprinzip*).[179]
- In einem je nach Text unterschiedlichen, von verschiedenen Faktoren abhängigen Maße sind die Schreibungen von dialektalen Lautungen beeinflusst.[180]

Für die vorliegende Untersuchung sind fünf Phänomene bei den Schreibungen besonders erwähnenswert:
- Sowohl zwischen <*i*> und <*j*> (und auch <*y*>) als auch zwischen <*u*> und <*v*> wurde sehr lange nicht nach dem Lautwert unterschieden, sondern die Schreibung richtete sich nach davon unabhängigen Konventionen. Bei <*u*> und <*v*> ist in den entsprechend geschriebenen Texten die Position ausschlaggebend: Initial wird <*v*> verwendet, medial und final <*u*>. Bei Kompositen und präfigierten Wörtern kann zu Beginn eines Stammmorphems <*v*> auch im Wortinneren erscheinen, dafür gibt es allerdings keine allgemein gültige Regel.[181] Bei der Verwendung von <*i*> und <*j*> sind keine so klaren Grundregeln zu erkennen, allerdings gewisse Präferenzen, die zum Teil auch vom jeweiligen Wort abhängen.[182]
- Einzelne Laute werden recht häufig durch Buchstabenkombinationen wiedergegeben. Die dabei entwickelten Konventionen entsprechen teilweise den heutigen – so zum Beispiel die Verwendung von <*h*> als Zeichen zur Kennzeichnung der Länge des vorangehenden Vokals oder die von verdoppelten Konsonanten teilweise zur Kennzeichnung von dessen Kürze –, daneben sind aber auch zahlreiche heute nicht mehr verwendete Kombinationen belegt. Zudem können manche Laute auch durch verschiedene Einzelbuchstaben repräsentiert werden. Als (allerdings extremes) Beispiel seien die Schreibungen angeführt, die in der *Frühneuhochdeutschen Grammatik* für die Affrikata */z/* zusammengestellt sind: „<*z, zz, zc, zcz, zt, ztc, zts, zh, zch, c, cc, cz, czc, ccz, czh, czt, czz, ctz, czcz, ch, t, tc, ts, tz, tcz, tzc, ttz, tzz, tzt, tztz, tsch, sq, sz, scz, htc*>“.[183] Bei den Vokalen wird der Variantenreichtum noch durch die Verwendung von übergeschriebenen Buchstaben und anderen diakritischen Zeichen erweitert.

[178] Vgl. ebd. S. 162 f.
[179] Vgl. ebd. S. 22 f.
[180] Ein Überblick über die wichtigsten mundartlichen Einflüsse findet sich ebd. S. 160.
[181] Im für diese Arbeit zugrunde gelegten Quellenkorpus korreliert die Schreibung sehr klar mit der initialen beziehungsweise nichtinitialen Stellung. In Reichmann/Wegera 1993, S. 46 und 108 werden aber auch andere Fälle aufgeführt.
[182] Vgl. ebd. S. 43 und 119.
[183] Ebd. S. 130.

- Die Getrennt- beziehungsweise Zusammenschreibung von Wörtern entspricht in vielen Fällen nicht dem heutigen Gebrauch. Auch wenn sich bei getrennt geschriebenen Wortverbindungen zum Teil nicht sicher entscheiden lässt, ob sie überhaupt als Worteinheit zu betrachten sind,[184] gibt es doch viele Fälle, in denen sich zum Beispiel der Artikel auf das Grundwort bezieht und damit recht deutlich hervorgeht, dass es sich um Kompositen handelt, aber trotzdem ein Leerraum zwischen den Wortbestandteilen klar erkennbar ist. Auch bei Wortbildungen mit trennbaren Präfixen weisen die Texte zum Teil eine deutliche Tendenz auf, diese Präfixe auch dort vom Rest des Wortes abzugrenzen, wo sie direkt davor stehen. Und oft ist einfach keine klare Entscheidung zu treffen, ob überhaupt eine Abgrenzung intendiert ist oder nicht – der Abstand zwischen Buchstaben ist teilweise (wenn er nicht sehr deutlich ist) allem Anschein nach von den jeweiligen Drucktypen abhängig und deshalb nicht immer aussagekräftig, so dass sich hier vielfach editorische Zweifelsfälle ergeben. In diesem Zusammenhang ist auch darauf hinzuweisen, dass sich die Kennzeichnung der Worttrennung beim Zeilenumbruch erst allmählich durchsetzte[185] und vor allem in älteren Texten ein fehlendes Trennzeichen kein klarer Hinweis auf eine getrennte Schreibung ist.
- Neben dem Abkürzungspunkt gibt es verschiedene weitere Kürzungszeichen, denen (anders als beim Punkt) mit hoher Wahrscheinlichkeit ein einigermaßen klarer Lautwert zugeordnet werden kann. Insbesondere beim Nasalstrich gibt es aber mehrere Möglichkeiten, ihn aufzulösen, nämlich als *n* oder *m*, aber auch (im Wort *vnd*) als *d* oder (nach einem *m*) als *b*. Vor allem die Entscheidung, ob am Wortende die Wiedergabe als *n* oder *m* der Textintention entspricht, ist nicht immer sicher möglich.
- Die Interpunktion entspricht vor allem im älteren Frühneuhochdeutsch, aber zu einem nicht unerheblichen Teil auch in jüngeren Texten nicht den heutigen Regeln.[186] Für die hier untersuchte Fragestellung ist insbesondere von Interesse, dass sich die Zeichensetzung auch zwischen parallelen Textfassungen unterscheiden kann und dass sich aus ihr in vielen Fällen eine Untergliederung in Sätze – die für die Untersuchung heutiger Texte gerne als Analyseeinheiten zugrunde gelegt werden – nicht eindeutig ableiten lässt.

Um einen etwas klareren Eindruck von zumindest einem Teil der beschriebenen Phänomene zu verschaffen, sollen im Folgenden einige Zitate aus verschiedenen

[184] Vgl. ebd. S. 32 (mit Literaturhinweisen).

[185] Als Trennzeichen findet sich im Korpus von DRQEdit in Textpassagen in gotischen Drucktypen meist der doppelte Bindestrich, zwei übereinander stehende Striche in leichter Schrägstellung (⸗), aber auch ein Schrägstrich, der vom Interpunktionszeichen / (der Virgel) oft äußerlich nicht zu unterscheiden ist.

[186] Vgl. Reichmann/Wegera 1993, S. 46 f.

Quellen des in dieser Arbeit untersuchten Textkorpus angeführt und kurz erläutert werden. Sie entstammen einigen Texten, die recht deutlich vom jeweiligen Dialekt beeinflusst sind. Das trifft nicht auf alle Texte dieser Zeit zu, es sind allerdings auch keine Einzelfälle.[187] Die Zitate entstammen dem Westoberdeutschen, dem Ostoberdeutschen, dem Westmitteldeutschen und dem Ostmitteldeutschen und damit den nach einem Gliederungsschema angesetzten vier Großräumen des Hochdeutschen.[188]

Zunächst ein Ausschnitt aus dem Freiburger Stadtrecht von 1520 als Beispiel für das Westoberdeutsche:

> Ein yede person die vff yeglich fürpott nit für gericht kumpt / so die selb von sinem gegẽteil als vngehorsam anzogen / vnd ir vom schultheissen nach dem alten bruch gerůfft würt / es syg der kleger oder antwůrter / so sy nit zů gegen were so man im Münster zů dem fronampt zůsamen gelütet hat / oder vngeferlich vmb dieselben zit / sol sy zů yeder vngehorsami dry schilling pfennig zů pene dem Schultheissen verfallen sin / dar zů irem gegenteil nach gelegenheit der sach / vmb kosten vnd schaden der vngehorsami halb erlittẽ / nach des gerichts erkantnuß vñ můtmassung abtrag thůn. Vnd es mo̊cht sich yemand so geuerlicher wise vff fürpott vngehorsam haltẽ / er würd ho̊her gestrafft.[189]

In mehreren Wörtern lässt sich feststellen, dass hier die neuhochdeutsche Diphthongierung nicht erfolgt ist[190]: *vff* (= *auf*), *bruch* (= *Brauch*), *syg* (= *sei*), *gelütet* (= *geläutet*), *wise* (= *Weise*). (Die Zeichenfolge *ei* in *Schultheissen*, *gegenteil* und *gelegenheit* entspricht den mittelhochdeutschen Formen[191], ist also nicht der neuhochdeutschen Diphthongierung zuzuordnen.) Die Schreibungen mit *ů* entsprechen dem mittelhochdeutschen Diphthong *uo*. Das *i* in der zweimal vorkommenden Form *vngehorsami* lässt sich als Variation eines Endungs-*e* erklären. Dem *g* in *syg* ist wohl kein Lautwert zuzuordnen.[192] Die Buchstaben *b* in *vmb* (= *um*) und *p* in *kumpt* (= *kommt*) und *fronampt* entsprechen der mittelhochdeutschen Schreibtradition, vermutlich ohne dass dem noch ein Lautwert zuzuordnen wäre.[193] Die Formen *vngeferlich* und *geuerlicher* zeigen, dass graphische Variation auch beim selben Wortstamm und in unmittelbarer Textnachbarschaft auftreten kann. Die Schreibung mit *u* ist für den damit nicht vertrauten heutigen Leser vermutlich erst

[187] Insgesamt ist die Varianz in früheren und handschriftlichen frühneuhochdeutschen Texten höher als in den hier behandelten Quellen, da sich im 16. Jahrhundert schon eine gewisse Vereinheitlichung erkennen lässt, insbesondere in den Drucken.

[188] Daneben wird eine Fünfteilung vertreten, die das Nordoberdeutsche als eigenen Raum betrachtet, der durch Einflüsse auch vom Mitteldeutschen geprägt sei. Vgl. Hartweg/Wegera 2005, S. 30 f.

[189] FreiburgStR. 1520, I 2 Art. 7 (Bl. VII v).

[190] Es wäre auch eine Konservierung alter Schreibgewohnheiten trotz veränderten Lautstands denkbar, allerdings ist der Text dem Alemannischen zuzuordnen und damit gerade einem Sprachraum, in dem diese Lautänderung nur in bestimmten Positionen erfolgte (vgl. Moser 1929, S. 161).

[191] Vgl. Lexer s. v. *schult-heiʒe*, *gegen-teil* und *ge-lëgen-heit*.

[192] Vgl. Reichmann/Wegera 1993, S. 99.

[193] Vgl. ebd. S. 87.

einmal irreführend (insbesondere im Zusammenhang mit dem vorangehenden *e*), sie ist aber im Frühneuhochdeutschen recht verbreitet. Eine Neigung zur Konsonantenverdopplung lässt sich in diesem Zitat vor allem bei den Schreibungen mit *ff* feststellen. Schließlich ist noch auf die Verwendung von *für* im Sinne des heutigen *vor* hinzuweisen.[194]

Ein Zitat aus der Bayrischen Gerichtsordnung von 1520 soll einen Eindruck von einem ostoberdeutschen Text vermitteln:

> Wo sich aber der Clager / oder Anntwurter / darnach abwesennlich ennthallten / vnnd khainen anwalld hinder jne verlassen / vnd dem Rechten nit mer nachkomen würden. Alßdann sȯllen sȯllich ladüng vnd verkündung / so offt die durch das Gericht außgeen / an der aussenbeleibenden parthey gewȯnndlichen behausungen / oder anndern jrn wonungen / oder vor den kirchmenigen / vnnd in den Stetten vnd Märckten / an den Ratheüsern / angeslagen werden / wie dann hyeuor jm̃ dritten vnd vierdtem gesatz diss Tittels begriffen ist.[195]

Die neuhochdeutsche Diphthongierung ist hier durchgeführt (*aussenbeleibenden*, *parthey*, *behausungen*, *Ratheüsern*). Die Setzung von Umlauten entspricht verschiedentlich nicht dem neuhochdeutschen Gebrauch (*Clager*, *sȯllen*, *sȯllich*); *ladüng* ist sehr ungewöhnlich und wohl am ehesten durch einen Druckfehler zu erklären. Initiales *kh* und *ai* (beides in *khainen*) sind typisch für mittel- und südostoberdeutsche Texte.[196] Eine Neigung zu Konsonantenhäufungen zeigt sich hier bei verschiedenen Buchstaben, zum Beispiel beim *n*, das in *abwesennlich* sogar in einer Nebensilbe verdoppelt ist. Die Schreibung *sl* in *angeslagen* ist für die Entstehungszeit wohl als konservativ zu bezeichnen;[197] im selben Text gibt es auch zahlreiche Schreibungen mit *schl*.[198] *hinder* (= *hinter*) entspricht der mittelhochdeutschen Form.[199]

Als Beispiel für das Westmitteldeutsche soll ein Ausschnitt aus der 1538 im Druck erschienenen *Reformation* des Kölner Erzstifts, also aus einem mittelfränkischen Text, dienen:

> Vnser gnedigster Herr Ertzbischoff zů Collen / Chůrfurst 7c ist vß teglichem anbringẽ Chůrfursten / Fursten vñ andern Stendẽ / Stetten vnd vnderthanen des heiligẽ Rȯmischen Reichs bericht / wie die Friegreuen vil parthien vñ sachẽ so an die Frienstoill vñ heimliche gerichte nit gehȯren vff gesynnen dere ancleger mit jren Citation ziehen / vnd vff vergebliche kost vñ schȧdẽ bringẽ. In den sachẽ sich parthielichẽ vermerckẽ laissen / den anklegern jrẽ

[194] Vgl. zu Austauschverhältnissen zwischen diesen beiden Wörtern DWB Bd. 4, Abt. 1, 1. Hälfte, passim s. v. „FÜR“ (zum Beispiel Sp. 639, 641 u. 643) und Bd. 26, Sp. 776 f.
[195] BayrGO. 1520, Tit. 2, 6 (Bl. 14 r/v).
[196] Vgl. Reichmann/Wegera 1993, S. 58 u. 102.
[197] Laut Reichmann/Wegera 1993, S. 116 wurde das Ostoberdeutsche in der 2. Hälfte des 15. Jahrhunderts vom Wechsel zur Schreibung mit *sch* „erfaßt“.
[198] Zum Beispiel Tit. 2, 3: *anschlagen*.
[199] Vgl. zur Entwicklung von *nd* zu *nt* Reichmann/Wegera 1993, S. 94.

> rait vñ vnd'weisunge / jn allermaiß als Anwelde vnd raitgeber / mitteilen. Auch die gerichts hendele selbs persoinlich schrieben.[200]

Hier lassen sich wie im Freiburger Stadtrecht verschiedene Beispiele für das Ausbleiben der neuhochdeutschen Diphthongierung finden, unter anderem *vß*, *vff* und *Friegreuen* (= *Freigrafen*). Das *i* in *ai* und *oi* (*Frienstoill*, *laissen*, *rait* etc.) ist nicht als Bestandteil eines Diphthongs zu lesen, sondern dient vielmehr als Dehnungszeichen – eine für das Mittelfränkische typische Schreibweise.[201] *vnd'* (zu lesen als *under*) entspricht wie das oben angeführte *hinder* noch der mittelhochdeutschen Form.[202]

Als vierter sprachlicher Großraum ist noch das Ostmitteldeutsche anzuführen. Hierzu ein Zitat aus der Leipziger Oberhofgerichtsordnung von 1529:

> Ich .N. schwere / als mich mein gnedigist vnd gnedige herren / an yhrer gnaden Oberhoffgericht zusitzen vorordent haben / das ich doselbst zu rechte / nach meinem hoͤchsten vorstentnus / sprechen / thuen vnd handeln will / vnd das nicht lassen / vmb Liebe / Naydt / Gabe / Freuntschafft / nach kaynerley sach willen / auch dorumb von partheyen ynn sonderhayt nichts nehmen / ader wissentlich wartende sein / will mich allenthalben ynn weltlichen sachen tzwuschen meiner gnedigist / vnd gnedigen herren vnderthan / dieweil ich dem Gerichte vorwandt bin / ausserhalben der suhne wissentlich tzurathen / adder tzuschreiben / wann die vor diss Oberhoffgerichte kommen sein / enthalten / getrewlich vnd ane geuerde / Als mir Gott helffe etc.[203]

In diesem Textstück sind insbesondere die Vokalunterschiede zum Neuhochdeutschen auffällig. Typisch wohl insbesondere für ostmitteldeutsche Texte[204] ist, dass anstelle der Vorsilbe *ver-* hier *vor-* steht (*vorordent*, *vorstentnus*, *vorwandt*). Die Verwendung von *ad(d)er* statt *oder* deutet darauf hin, dass es sich um einen mitteldeutschen Text handelt.[205] *ane* anstelle von *ohne* ist die ältere Form,[206] *do* statt *da* ist vor allem im älteren Frühneuhochdeutsch häufig.[207] *tzwuschen* entspricht neuhochdeutsch *zwischen*. Dass *suhne* ohne Umlaut geschrieben ist, mag dadurch

[200] KölnErzstiftRef. 1538, Blatt F iiij v.
[201] Vgl. Reichmann/Wegera 1993, S. 33.
[202] Vgl. oben Anm. 199.
[203] LeipzOHofGO. 1529, Bl. A ij r/v.
[204] In Reichmann/Wegera 1993, S. 78 heißt es: „Im Md. konkurrieren bis ins 16. Jh. bes. *i* mit *e* (*in-*, *int-*, *ir-*, *vir-*), seltener *u* (*zur-*, *unt-*) und *o* (*vor-*, *ont-*)." Im hier untersuchten Korpus, das freilich insbesondere im Hinblick auf den Zeitraum und die Beschränkung auf Drucke die Bandbreite des Frühneuhochdeutschen nur teilweise abdeckt, lässt sich davon nur *vor-* gehäuft finden, und dies allem Anschein nach deutlich überwiegend in ostmitteldeutschen Texten. Da *vor* am Wortanfang aber durchaus auch einem neuhochdeutschen *vor* (oder auch *for*, zum Beispiel im Wort *vort*) entsprechen kann und über 20.000 Funde für eine qualifizierte Aussage entsprechend zu prüfen wären, soll dies als Vermutung stehen bleiben.
[205] Vgl. ebd. S. 38.
[206] Vgl. DWB Bd. 13, Sp. 1210 s. v. „ohne".
[207] Vgl. Moser 1929, S. 147 (§ 75 Anm. 6).

zu erklären sein, dass sich die Umlautkennzeichnung bei *ü* im mitteldeutschen Raum teilweise erst in der ersten Hälfte des 16. Jahrhunderts durchsetzte.[208].

Wie aus dem bisher Dargestellten deutlich geworden sein sollte, ist eine Zuordnung frühneuhochdeutscher Schreibformen zu neuhochdeutschen Wörtern (oder auch Wortstämmen) recht häufig nicht ohne Weiteres zu leisten, es gibt sogar nicht wenige Fälle, in denen eine frühneuhochdeutsche Schreibung einem neuhochdeutschen Wort exakt entspricht (eventuell abgesehen von der Groß-/Kleinschreibung), aber trotzdem einem anderen Wort zuzuordnen ist. Aus den vorgestellten Zitaten sind hier anzuführen: *für* (= *vor*), *bruch* (= *Brauch*), *hinder* (= *hinter*), *schwere* (= *schwöre*) und *ader* (= *oder*). Die Verwendung des Präfixes *vor-* anstelle des neuhochdeutschen *ver-* führt im zuletzt angeführten Zitat zwar nicht zu solchen Entsprechungen, wohl aber sonst des Öfteren (als Beispiele seien die neuhochdeutschen Wortpaare *vergeblich – vorgeblich*, *verführen – vorführen* und *verkommen – vorkommen* genannt). Und einige Schreibungen weichen nicht nur von der heutigen Orthographie ab, sondern legen dem nicht entsprechend vorgebildeten heutigen Leser auch eine Aussprache nahe, die bei der Wortzuordnung in die Irre führt. Das gilt insbesondere für die Verwendung von *u* für den Lautwert *f* beziehungsweise *w* (etwa in dem in den angeführten Zitaten nicht vorkommenden, im Korpus aber sehr häufig belegten Wort *Grauen*) sowie auch für die Verwendung von *i* (oder *e*) als Dehnungszeichen nach einem Vokal. Überhaupt muss bei der Verarbeitung der Buchstabenfolgen frühneuhochdeutscher Texte mit Interpretationsunsicherheiten auf mehreren Ebenen gerechnet werden:

1. Zunächst einmal ist zu entscheiden, welche aufeinander folgenden Buchstaben jeweils als Einheit zu lesen sind. Dieses Problem stellt sich in einem gewissen Maße auch bei der Verarbeitung von deutschen Texten in heutiger Orthographie, da auch sie eine Reihe von aus mehreren Buchstaben zusammengesetzten Schreibungen für einzelne Laute aufweist.[209] Bei der Verarbeitung frühneuhochdeutscher Texte stellt es sich verstärkt, da im Frühneuhochdeutschen eine viel größere Zahl von unterschiedlichen Buchstabenfolgen zur Repräsentation von einzelnen Lauten belegt ist. Viele dieser Kombinationen lassen sich allerdings dadurch recht gut als Einheit erkennen, dass die Deutung der enthaltenen Buchstaben als Repräsentation von Einzellauten nach den im Deutschen geltenden Lautfolgemustern gar nicht oder nur an Morphemgrenzen (vor allem bei Kompositen) plausibel ist: Zum Beispiel kann *dt* am Silbenende nur für einen einzigen Laut stehen. Entsprechendes gilt für andere Paare von Lenis- und Fortis-Schreibungen und natürlich auch für verdoppelte Konsonanten. Besondere Schwierigkeiten bereiten hingegen Buchstabenfolgen wie *aue* und *eue*, da hier

[208] Vgl. Reichmann/Wegera 1993, S. 48.

[209] Als Beispiele, bei denen die Zusammenfassung solcher Buchstabenfolgen in die Irre führt, seien *sch* in *Häuschen* und *th* in *Rathaus* genannt.

ohne weitere Kenntnis der Schreibgewohnheiten im jeweiligen Text beziehungsweise ohne Berücksichtigung des gesamten Wortes nicht zu entscheiden ist, ob das *u* hier vokalischen Charakter hat und damit Teil eines Diphthongs ist oder den Laut *f* beziehungsweise *w* repräsentieren soll.

2. Den Schreibungen können häufig verschiedene Lautungen zugeordnet werden. Auch hierfür lassen sich entsprechende Fälle in der gegenwärtigen Orthographie finden; so ist die Vokallänge auch heute aus der Schreibweise nicht immer sicher zu ermitteln, und die Auslautverhärtung von *b*, *d* und *g* wird nicht abgebildet. Das Frühneuhochdeutsche ist hiervon freilich ungleich stärker betroffen. Auch hier ist die lange Zeit nicht (beziehungsweise nur begrenzt) am Lautwert orientierte Verwendung der Buchstaben *u*/*v* sowie *i*/*j*/*y* ein herausragendes Beispiel, da bei ihnen ohne Berücksichtigung des jeweiligen Buchstabenumfelds noch nicht einmal eine sichere Zuordnung zur Gruppe der Vokale oder der der Konsonanten möglich ist. Weitere Beispiele sind nicht als solche gekennzeichnete Umlaute[210], Wechsel zwischen Fortis- und Lenis-Schreibungen insbesondere im Anlaut sowie Wechsel zwischen *b* und *v*/*w*/*u* sowie zwischen *g* und *j*. Grundsätzlich ist davon auszugehen, dass die Texte – in sehr unterschiedlichem Maße – einerseits von Schreibtraditionen, andererseits von der Aussprache beeinflusst sind. Das erklärt – neben den Schreibweisen, die trotz anderer Lautung im jeweiligen Dialekt der modernen Hochlautung weitgehend entsprechen – nicht nur Schreibungen entsprechend der tatsächlichen Aussprache, sondern auch Hyperkorrekturen.
3. Die Schreibung kann eine nicht dem Neuhochdeutschen entsprechende Lautung abbilden, der dann noch das neuhochdeutsche Äquivalent zuzuordnen ist. Dies betrifft im Wesentlichen die eben schon genannten Punkte (bis auf die Verwendung von *i*, *j*, *y*, *u* und *v*).[211]

Die Interpretation der Buchstabenfolgen im Hinblick auf die tatsächliche oder intendierte Lautung sowie die Einordnung in die Entwicklung von Lautung und Schreibung sowie in die Dialektlandschaft ist sicherlich von philologischem Interesse, allerdings für die Ermittlung von Textübernahmen nicht erforderlich, so dass die eben angeführten Punkte 2 und 3 auch zusammengefasst werden können. Tatsächlich ist dafür, wie noch dargelegt werden soll, auch keine Transformation der Wortformen in neuhochdeutsche Entsprechungen erforderlich. Es sollte aber

[210] Vgl. Reichmann/Wegera 1993, S. 34 f.

[211] Bei konsequenter Verfolgung dieses Wegs wäre schließlich noch – soweit man als Ziel der Transformation Buchstaben und nicht Laute sieht – die Umwandlung neuhochdeutscher Lautfolgen in die heutige Orthographie erforderlich. Die damit verbundenen Schwierigkeiten brauchen hier nicht weiter erörtert zu werden, da die angeführten Interpretationsebenen nicht als Vorschlag für einen tatsächlich durchzuführenden mehrstufigen Analyseprozess vorgestellt wurden. Generell ist bei Schreibungen mit einem starken Einfluss von Schreibtraditionen zu rechnen, und viele unserer orthographischen Regeln entsprechen Schreibgewohnheiten, die sich schon (wenn auch mit geringerer Verbreitung) im Frühneuhochdeutschen finden.

deutlich geworden sein, dass bei der Varianz der Schreibungen die Lautverhältnisse und -entwicklungen eine Rolle spielen, und zugleich, dass im Folgenden bei Aussagen über einzelne „Laute“ keine Behauptungen über den tatsächlichen jeweiligen Lautwert beabsichtigt sind.

2 String- und Textvergleich. Techniken und Anwendungen

Teil 2 bietet einen Überblick über verschiedene Techniken, die für den Vergleich von Zeichenketten und Texten entwickelt wurden, und über verschiedene Arbeitsgebiete der *Digital Humanities*, in denen es um Textvergleich geht.

Kapitel 2.1 stellt grundlegende Probleme und Lösungen aus dem Bereich der Stringalgorithmik vor, Kapitel 2.2 Verfahren und Programme, die für den Stringvergleich im Rahmen der Bioinformatik entwickelt wurden, und Kapitel 2.3 sprach- und textorientierte Ansätze. Kapitel 2.4 beschäftigt sich mit der Plagiatserkennung, Kapitel 2.5 mit der *Text-Reuse*-Forschung und Kapitel 2.6 mit dem Thema der automatischen Kollationierung.

2.1 Algorithmen für den Vergleich von Zeichenketten

Die Ermittlung von Textübereinstimmungen gehört in technischer Hinsicht in den Bereich der Untersuchung von Zeichenketten (im Folgenden meist aus Gründen der sprachlichen Vereinfachung mit dem aus dem Englischen übernommenen Terminus *Strings* bezeichnet). In diesem Kapitel geht es vor allem um zwei algorithmische Probleme beim Stringvergleich und um ihre klassischen Lösungen.

Diese beiden schon seit Jahrzehnten erörterten Probleme[212] beim Vergleich zweier Strings sind zum einen die Ermittlung der längsten gemeinsamen (nicht unterbrochenen) Zeichenkette, des längsten gemeinsamen Teilstrings (*longest common substring*, auch als *longest common factor* bezeichnet[213]), zum anderen die der längsten gemeinsamen Teilsequenz (*longest common subsequence*)[214] – mit Letzterem ist eine Folge von Zeichen (oder auch Zeichenketten, zum Beispiel Wörtern) gemeint, die in beiden untersuchten Strings (oder anderen geordneten Folgen von Elementen, zum Beispiel Texten) in der gleichen Reihenfolge vorkommt, wobei anders als beim längsten gemeinsamen Teilstring zwischen den zur längsten gemeinsamen Teilsequenz gehörenden Elementen auch noch anderes stehen darf. Die ähnliche Benennung und die für beide Probleme verwendete Abkürzung *LCS* (oder

[212] Eine nach Erscheinungszeitraum eingegrenzte Recherche in der *Google Buchsuche* ergab als älteste Stelle für „longest common substring“ eine von 1966 (*https://books.google.de/books?id=v0s1AQAAIAAJ&q=%22longest+common+substring%22*); für „longest common subsequence“ ließen sich mit *Google* Treffer ab 1972 finden (*https://www.google.de/search?q=%22longest+common+subsequence%22&tbs=cdr%3A1%2Ccd_min%3A1900%2Ccd_max%3A1974*).

[213] So etwa in Crochemore/Rytter 1994/2009, zum Beispiel S. 109.

[214] Die ebenfalls anzutreffende Übersetzung mit *längste gemeinsame Teilfolge* wird hier nicht verwendet, da unter einer *Zeichenfolge* meines Erachtens üblicherweise eine nicht unterbrochene Abfolge von Zeichen verstanden wird, was in diesem Zusammenhang verwirren könnte.

auch *LCSS*)[215] führen leicht zu einer Verwechslung, es handelt sich aber im Hinblick auf Aufgabe und Komplexität um klar voneinander abzugrenzende Fragestellungen.

2.1.1 Suffixbäume und längster gemeinsamer Teilstring

Ein bekanntes Verfahren zur Ermittlung des längsten gemeinsamen Teilstrings basiert auf einem Suffixbaum, das heißt auf einer bestimmten Regeln unterworfenen Baum-Datenstruktur, die alle Suffixe eines Strings verzeichnet. Dabei ist *Suffix* nicht im linguistischen Sinne zu verstehen, sondern steht für jede im Gesamtstring enthaltene zusammenhängende Teilzeichenkette, die bis zu dessen Ende reicht (einschließlich der vollständigen Zeichenkette)[216] – einem String mit der Länge *n* lassen sich also *n* Suffixe zuordnen. Die Bezeichnung *Suffixbaum* legt vielleicht die Vermutung nahe, dass damit irgendwie vom Ende her, also rückwärts, gesucht werde. Tatsächlich ist es aber umgekehrt: Von der Wurzel des Baums ausgehend sind die Zeichen jedes Suffixes in der originalen Reihenfolge verzeichnet. Ein Suffix liegt erst dann vor, wenn ein Endknoten erreicht ist. Der Pfad zu diesem Endknoten führt aber über die Präfixe des Suffixes, also über die zum Suffix gehörigen Teilstrings der gesamten Zeichenkette.

Ein Suffixbaum weist einige Charakteristika[217] auf, die dazu führen, dass zum Beispiel sehr effizient nach einem beliebigen Teilstring gesucht werden kann:

- Die Kanten des Baums entsprechen Teilstrings, innere Knoten den Stellen, an denen es unterschiedliche Fortsetzungen zu einem Teilstring gibt, Endknoten (Blätter) dem Ende des jeweiligen Suffixes, das sich durch die Verkettung der Strings ergibt, die den Kanten von der Wurzel bis zu einem Endknoten zugeordnet sind.
- Jeder innere Knoten hat mindestens zwei Kindknoten, und alle von ihm ausgehenden Kanten beginnen mit unterschiedlichen Anfangsbuchstaben.
- Es gibt eine 1:1-Beziehung zwischen Suffixen und Endknoten. Um dies auch dann zu erreichen, wenn ein Suffix Präfix eines anderen Suffixes ist, wird üblicherweise am Ende des in den Suffixbaum überführten Strings ein Zeichen angehängt, das im eigentlichen String nicht vorkommt und nur als Anzeiger für das String-Ende fungiert.

Die Beschreibung soll anhand eines Beispiels veranschaulicht werden. Abbildung 2.1 stellt den Suffixbaum zum String *ababcabd* dar. Die Baumwurzel steht um der

[215] Vgl. die Ergebnisse entsprechender Recherchen mit *Google*, zum Beispiel *https://www.google.de/search?q=%22longest+common+substring%22+%22lcss%22*, sowie *http://de.wikipedia.org/wiki/LCS*.

[216] Nach der Definition in Gusfield 1997, S. 4 (in Abgrenzung zu einem „*proper* [...] suffix“, das kürzer ist als die gesamte Zeichenkette, aber aus mindestens einem Zeichen besteht) sollte auch der leere String ein Suffix darstellen, er wird aber in Suffixbäumen allenfalls insofern repräsentiert, als darin meist ein Zeichen zur Markierung des String-Endes aufgenommen wird.

[217] Vgl. ebd. S. 90 f.

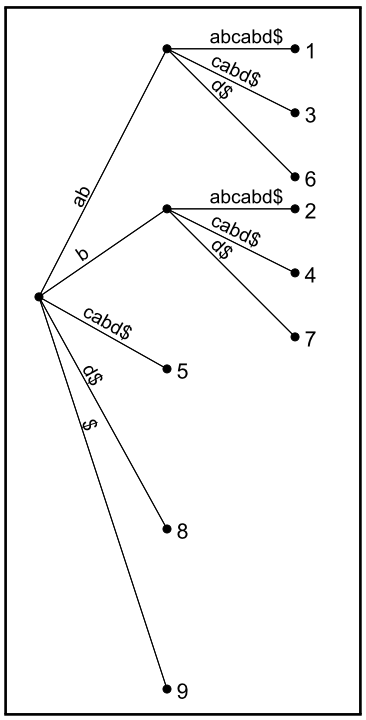


Abb. 2.1: Suffixbaum des Strings *ababcabd*

besseren Lesbarkeit willen links. Als abschließendes Zeichen, das im String selbst nicht vorkommt, wird (wie in anderen Darstellungen zu Suffixbäumen[218]) *$* verwendet. Neben den Endknoten steht jeweils eine Zahl, die angibt, beim wievielten Zeichen das entsprechende Suffix beginnt.

So enthält der oberste Pfad, der von der Wurzel über die Kanten *ab* und *abcabd$* zum Endknoten mit der Zahl 1 führt, den gesamten String (er beginnt entsprechend der Zählung beim ersten Zeichen und reicht, da er zu einem Endknoten führt, bis zum Ende der untersuchten Zeichenkette). Der Pfad, der mit derselben Kante *ab* beginnt, dann aber mit *cabd$* fortgesetzt wird, gibt hingegen den Teilstring wieder, der mit dem dritten Zeichen beginnt – die Kante *ab* steht hier also für ein anderes Stück des Gesamtstrings als im Pfad zum Endknoten 1.

Schon in diesem recht kleinen Beispiel zeigt sich, dass im Prinzip gleiche Informationen in Suffixbäumen häufig wiederholt verzeichnet werden. Zum einen betrifft das die Zeichenketten, die den Kanten zugeordnet sind – in diesem Fall handelt es sich einschließlich des jeweiligen Abschlusszeichens *$* um 39 Zeichen, während der zugrunde liegende String (ebenfalls einschließlich des angehängten *$*) nur neun Zeichen umfasst. Das Verhältnis divergiert umso stärker, je länger der String ist, dessen Suffixe verzeichnet werden, und je weniger Wiederholungen darin enthalten sind. Das muss technisch allerdings nicht ins Gewicht fallen, da es

[218] Zum Beispiel Gusfield 1997, S. 91 u. ö.

nicht erforderlich ist, die Kantenbeschriftungen im Klartext zu speichern – es reicht vielmehr prinzipiell aus, den jeweiligen Anfangs- und Endpunkt im Gesamtstring festzuhalten, was bei Kanten, denen eine große Zahl von Zeichen zugeordnet ist, den Speicherbedarf erheblich reduziert. Zum anderen stehen zum Teil aber auch mehrere Kanten für die gleichen Stücke im Gesamtstring. Das trifft im Beispiel auf die beiden Dreiergruppen von Kanten zu, die von den beiden inneren Knoten ausgehen (also von den Knoten, die weder Wurzel noch Blatt sind). Jede Kante einer dieser Dreiergruppen hat eine genaue Entsprechung in der anderen Dreiergruppe, und die Zahlen der sich daraus ergebenden Endknotenpaare unterscheiden sich jeweils um 1. Das lässt sich dadurch erklären, dass sie jeweils auf eine Kante folgen, deren Beschriftung mit *b* endet, wobei im ersten Fall zusätzlich noch ein *a* vorangeht; die Endknoten der unteren Dreiergruppe entsprechen also jeweils den Suffixen, die entstehen, wenn man in den Suffixen der oberen Dreiergruppe das erste Zeichen streicht. Die mögliche Zahl innerer Knoten (und damit der Speicherbedarf) ist aber begrenzt: Da jeder innere Knoten entsprechend den Konstruktionsregeln mindestens zwei Kindknoten hat und da es jeweils nur einen Pfad von der Wurzel zu einem Endknoten gibt, ist die Zahl der Endknoten um mindestens 1 größer als die Zahl der inneren Knoten.[219]

Suffixbäume ermöglichen die effiziente Bearbeitung einer Vielzahl von Aufgaben. Wenn ein String in dieser Form aufbereitet ist, lässt er sich zum Beispiel mit sehr wenig Verarbeitungsschritten nach Zeichenketten durchsuchen, da nur vom Wurzelknoten aus geprüft werden muss, ob es einen Pfad gibt, der dem jeweils gesuchten String entspricht. Der maximale Aufwand dafür ist nicht abhängig von der Länge des Strings im Suffixbaum, sondern nur von der der gesuchten Zeichenkette, bietet also gerade für häufige Suchen in einem konstanten Text erhebliche Vorteile.[220]

Für die hier untersuchte Fragestellung ist insbesondere von Interesse, dass sich über einen Suffixbaum mit einem vergleichsweise geringen Aufwand feststellen lässt, was der längste gemeinsame Teilstring zweier Zeichenketten ist: Zusätzlich zur Konstruktion eines Suffixbaums für einen aus den beiden Zeichenketten zusammengefügten Gesamtstring ist es nur erforderlich, die inneren Knoten mit Markierungen zu versehen, aus denen die Zuordnung zu einem der beiden Strings oder zu beiden hervorgeht, und beim Traversieren des Baums festzuhalten, was die bis dahin gefundene längste Zeichenfolge ist, die einem Knoten mit doppelter Markierung zugeordnet werden kann.[221]

Obwohl sich Suffixbäume so konstruieren lassen, dass der Zeit- und der Speicherbedarf bei gleichbleibendem Alphabet nur linear von der verarbeiteten Stringlänge

[219] Schon Weiner 1973, S. 6 weist auf diese obere Grenze hin.
[220] Weitere Anwendungsbeispiele werden in Gusfield 1997, S. 122–168 sowie in den ebd. auf S. 168 aufgelisteten Abschnitten zusammengestellt.
[221] Vgl. ebd. S. 125.

abhängen,[222] kann der Speicherbedarf bei der Anwendung auf etwas größere Textmengen recht schnell so groß werden, dass sich praktische Probleme ergeben. Der Speicherbedarf und Verfahrensweisen zur Effizienzsteigerung werden deshalb in der Literatur immer wieder erörtert,[223] und es wurden auch andere Datenstrukturen entwickelt, die in ähnlicher Weise wie Suffixbäume genutzt werden können, aber einen geringeren Ressourcenbedarf haben. Unterkapitel 2.2.1 soll dazu im Zusammenhang mit der Beschreibung einiger Programme aus dem Bereich der Bioinformatik einen kurzen Überblick geben.

2.1.2 Editierdistanz, Alinierung und längste gemeinsame Teilsequenz

Während es beim längsten gemeinsamen Teilstring um vollständige Gleichheit geht, gehört die längste gemeinsame Teilsequenz in den Bereich der Übereinstimmungen unter Zulassung gewisser Abweichungen. Das eröffnet einerseits andere Verwendungsmöglichkeiten, macht aber andererseits die Ermittlung im Vergleich zu der des längsten gemeinsamen Teilstrings wesentlich aufwendiger.[224]

Eine längste gemeinsame Teilsequenz lässt sich aus einer Alinierung zweier Elementfolgen entnehmen, wenn diese mit dem Ziel einer Maximierung der Entsprechungen erfolgt ist. Tabelle 2.1 zeigt in jeweils einer Tabellenspalte drei diesem Kriterium genügende Alinierungen zu den Strings *abacabdcb* und *bcaacdaa*. Darin

1. alinierter String:	ab-acabdcb	ab-acabdcb--	abacab-dcb
2. alinierter String:	-bca-acdaa	-bca-a--cdaa	-b-caacdaa
Editierskript:	-u+u-ucucc	-u+u-u--uc++	-u-uuc+ucc

Tab. 2.1: drei alinierte längste gemeinsamen Teilsequenzen zweier Strings

sind die beiden Zeichenketten jeweils in einer Zeile so dargestellt, dass möglichst viele der übereinander stehenden Zeichenpaare gleich sind. Zur Verdeutlichung sind diese Zeichenpaare unterstrichen. Die dieser Alinierung entsprechende längste gemeinsame Teilsequenz besteht also aus der Folge der unterstrichenen Zeichen in einer der beiden Zeilen. „-“ ist als Code eingeschoben, wenn das Zeichen, das darüber oder darunter im anderen String steht, hier keine Entsprechung hat. In einer dritten Zeile ist ein Editierskript[225] angegeben, das die mindestens erforderlichen Schritte bei der Transformation des ersten Strings in den zweiten codiert, die zu dieser Alinierung führen. Dabei wird vorausgesetzt, dass ein solcher Schritt in der Löschung, Einfügung oder dem Austausch eines Zeichens besteht. In der

[222] Vgl. ebd. S. 94–119.

[223] Vgl. zum Beispiel ebd. S. 116–119, wo verschiedene Möglichkeiten der Implementierung mit unterschiedlichen Auswirkungen auf Laufzeit und Speicherbedarf betrachtet werden.

[224] Vgl. ebd. S. 210 und 228 zu den grundlegenden Unterschieden zwischen beiden Problemen.

[225] Die englische Formulierung „edit transcript“ (Gusfield 1997, S. 215) ist so im Deutschen anscheinend nicht gebräuchlich, es handelt sich auch nicht um ein Transkript im Sinne einer Kopie, sondern um ein Skript im Sinne einer Anweisungsfolge.

Tabelle wird die Eliminierung eines Zeichens mit „-“ codiert, eine Einfügung mit „+“, eine Ersetzung mit „c“ (für *changed*) und eine unveränderte Übernahme mit „u“ (für *unchanged*).[226]

Wie man hier sieht, kann es mehrere Möglichkeiten der Alinierung mit längster gemeinsamer Teilsequenz geben – im Beispiel wird davon nur ein Teil dargestellt. Die drei hier gezeigten längsten gemeinsamen Teilsequenzen bestehen aus unterschiedlichen Zeichenketten – *baad*, *baac* und *bcad*. Daneben kann es – abhängig von Elementwiederholungen – aber auch unterschiedliche Möglichkeiten geben, den zu einer solchen Teilsequenz gehörenden Elementen Positionen in den untersuchten Gesamtfolgen zuzuordnen.[227] Es reicht also nicht unbedingt aus, eine längste gemeinsame Teilsequenz zu verzeichnen, um später aus ihr eine bestimmte Alinierung rekonstruieren zu können. Wenn es jedoch nur darum geht, überhaupt eine Alinierung zu ermitteln, die dieser Teilsequenz entspricht, lässt sich dies dadurch erreichen, dass jede Elementfolge so weit durchgeschaut wird, bis eine Position für das erste Element der Teilsequenz gefunden wird, dann im Rest nach dem ersten Vorkommen des zweiten Elements gesucht wird usw.

Wie schon 1974 von Robert A. Wagner und Michael J. Fischer dargelegt wurde,[228] lässt sich ein nach bestimmten Kriterien optimales Editierskript (mit den genannten Operationen Löschung, Einfügung und Ersetzung sowie der unveränderten Übernahme jeweils einzelner Zeichen) zur Transformation eines Strings durch Aufbau eines zweidimensionalen Arrays ermitteln, also einer Datenstruktur, die sich in Tabellenform darstellen lässt.[229] Darin stehen die Zeilen für einzelne Zeichen des einen und die Spalten für einzelne Zeichen des anderen Strings (wobei am Anfang eine Zeile und eine Spalte für „kein Zeichen“ stehen), und in den Tabellenzellen wird sukzessive verzeichnet, wie groß die Editierdistanz ist, das heißt – nach dem hier zugrunde gelegten Verständnis[230] – die Bewertung der mindestens erforderlichen Editieroperationen, um das Stringpräfix[231], das bis zu dieser Zeile reicht, in das Stringpräfix, das bis zu dieser Spalte reicht, zu transformieren. Voraussetzung ist,

[226] Das ist das Codierungsschema, das im *Perl*-Modul *Algorithm::Diff* (*http://search.cpan.org/~tyemq/Algorithm-Diff-1.1902/lib/Algorithm/Diff.pm*) verwendet wird.

[227] Greenberg 2003 enthält Formeln zur Berechnung oberer Schranken für die Zahl unterschiedlicher längster gemeinsamer Teilsequenzen und unterschiedlicher Positionszuordnungen in Abhängigkeit von der Länge der Elementfolgen.

[228] Wagner/Fischer 1974. Das Verfahren wurde seitdem vielfach beschrieben, zum Beispiel in Gusfield 1997, S. 217–223 und in Crochemore/Rytter 1994/2009, S. 259–262. Der unten genannte, in Needleman/Wunsch 1970 vorgestellte Ansatz ist zwar ähnlich und etwas älter, zielt allerdings nicht auf die Berechnung der Editierdistanz.

[229] Um der besseren Beschreibbarkeit willen wird im Folgenden davon ausgegangen, dass die Werte tatsächlich in einer Tabelle eingetragen werden.

[230] Der Terminus wird hier nicht (wie zum Beispiel in Gusfield 1997, S. 216) synonym mit *Levenshtein-Distanz* gebraucht (vgl. dazu den folgenden Absatz), sondern so wie in Wagner/Fischer 1974, S. 169, wo ausdrücklich darauf hingewiesen wird, dass die Bewertung der verschiedenen Editieroperationen unterschiedlich sein und auch von den betroffenen Zeichen abhängen kann.

		b	**c**	**a**	**a**	**c**	**d**	**a**	**a**
	0	1	2	3	4	5	6	7	8
a	1	1	2	2	3	4	5	6	7
b	2	1	2	3	3	4	5	6	7
a	3	2	2	2	3	4	5	5	6
c	4	3	2	3	3	3	4	5	6
a	5	4	3	2	3	4	4	4	5
b	6	5	4	3	3	4	5	5	5
d	7	6	5	4	4	4	4	5	6
c	8	7	6	5	5	4	5	5	6
b	9	8	7	6	6	5	5	6	6

Tab. 2.2: Levenshtein-Distanz-Tabelle zur Transformation von *abacabdcb* in *bcaacdaa*

dass den Editieroperationen jeweils eine Bewertung zugeordnet ist, die aber nicht notwendigerweise bei allen gleich sein muss.

Als Beispiel enthält Tabelle 2.2 Editierdistanzen zwischen den Präfixen von *abacabdcb* und *bcaacdaa*, also von den beiden Strings, die schon als Basis für das Alinierungsbeispiel gedient haben. Grau unterlegt sind die Tabellenzellen, die der ersten Alinierung in Tabelle 2.1 entsprechen, also dem Editierskript *-u+u-ucucc*[232]. Die hier gezeigten Editierdistanzen entsprechen dabei der Levenshtein-Distanz[233], die die mindestens benötigte Gesamtzahl an Einfüge-, Lösch- und Ersetzungsoperationen angibt.

Im Folgenden wird dargestellt, wie sich die in der Tabelle verzeichneten Werte berechnen lassen. In dieser Beschreibung wird angenommen, dass – wie in Tabelle 2.2 – den Zeilen der Tabelle die Zeichen des Ausgangsstrings zugeordnet sind und den Spalten die Zeichen des Resultats der Transformation.

Zunächst einmal werden die erste Zeile und die erste Spalte gefüllt. Die erste Zeile enthält die Editierdistanzen für die Transformation des leeren Strings in eine von links nach rechts wachsende Zahl von Zeichen; dementsprechend ergibt sich als Wert jeweils das Produkt aus der Zeichenzahl und der Bewertung der Einfügeoperation. Umgekehrt enthält die erste Spalte die Transformation einer von oben nach unten wachsenden Zahl von Zeichen in den leeren String, und

[231] Wie beim *Suffix* ist dies nicht im morphologischen Sinne zu verstehen, sondern es ist ein Teilstring gemeint, der am Anfang des Gesamtstrings beginnt.

[232] Vgl. zu den dabei verwendeten Codes oben S. 58.

[233] Das Konzept einer Editierdistanz in diesem Sinne wurde wohl erstmals in den 1960er Jahren von Vladimir Levenshtein vorgestellt, deshalb ist diese Bezeichnung verbreitet, vgl. Gusfield 1997, S. 216 und die deutsche Wikipedia s. v. „Levenshtein-Distanz“ (*http://de.wikipedia.org/wiki/Levenshtein-Distanz*).

die Editierdistanz beträgt hier jeweils das Produkt aus der Zeichenzahl und der Bewertung der Löschoperation. Die übrigen Tabellenzellen lassen sich jeweils auf der Basis schon ermittelter Werte berechnen. Denn wenn man nur das Ende eines einer Zelle zuzuordnenden Editierskripts betrachtet, gibt es für ein optimales Skript drei mögliche Fälle (beziehungsweise vier, da der dritte zwei Unterfälle enthält):

1. Das letzte Zeichen des ersten Stringpräfixes muss gelöscht werden. Dann ergibt sich die Editierdistanz aus der Summe der Editierdistanz, die in der unmittelbar darüber stehenden Zelle vermerkt ist (da darin die Bewertung der Transformation eines um ein Zeichen kürzeren Präfixes ins gleiche Resultat steht), und der Bewertung der Löschoperation.
2. Das letzte Zeichen des zweiten Stringpräfixes muss eingefügt werden. Dann ergibt sich die Editierdistanz aus der Summe der Editierdistanz, die in der unmittelbar links davor stehenden Zelle vermerkt ist (also aus der Bewertung der Transformation desselben ersten Stringpräfixes in ein um ein Zeichen kürzeres Resultat) und der Bewertung der Einfügeoperation.
3. Das letzte Zeichen des ersten Stringpräfixes ist dem letzten Zeichen des zweiten Stringpräfixes zuzuordnen. Wenn beide Zeichen übereinstimmen, ist die Editierdistanz der unveränderte Wert der Zelle, die schräg darüber auf der linken Seite steht, wenn sie sich unterscheiden, die Summe dieses Werts und der Bewertung der Ersetzungsoperation.

Um die optimale Möglichkeit zu finden, muss man diese drei Werte berechnen und den besten davon auswählen. Umgekehrt kann man eines der möglichen Editierskripte finden, indem man von der jeweils betrachteten Zelle aus die Zelle darüber, die links davor und die schräg darüber dahingehend prüft, ob ihr Wert als Vorgängerwert in Frage kommt, und damit Schritt für Schritt rückwärts die Einträge im Editierskript ermittelt – oder man kann schon bei der Erstellung der Tabelle verzeichnen, welche Zellen diesem Kriterium entsprechen.[234]

Wenn das Ziel darin besteht, eine längste gemeinsame Teilsequenz zu ermitteln, ist es sinnvoll, die Ersetzung so zu bewerten wie eine Kombination von Löschung und Einfügung, also mit 2. Dann ergibt sich die Länge der gesuchten Teilsequenz aus der Summe der Längen von Ausgangs- und Resultatstring abzüglich der Editierdistanz, geteilt durch 2.[235] Dass die Ersetzung – jedenfalls bei diesem Ziel – stärker ins Gewicht fällt als die Einfügung oder Löschung, lässt sich leicht am Beispiel eines vertauschten Buchstabenpaars veranschaulichen: Die Transformation von *ab* in *ba* lässt sich erreichen durch eine Löschung und eine Einfügung (dann gibt es ein gemeinsames Zeichen *b*), durch eine Einfügung und eine Löschung (dann gibt es ein gemeinsames Zeichen *a*) oder auch durch zwei Ersetzungen – dann gibt es

[234] Vgl. Gusfield 1997, S. 221.
[235] Vgl. Wagner/Fischer 1974, S. 173 oder auch Crochemore/Rytter 1994/2009, S. 263.

zwar die gleiche Gesamtzahl an Editieroperationen, aber es wird kein gemeinsames Zeichen ermittelt.[236]

Das Konzept der Editierdistanz wurde für den Vergleich von Zeichenketten entwickelt, die einander sehr ähnlich sind – als Anwendungsmöglichkeit wird zum Beispiel von Wagner und Fischer die Korrektur von Schreibfehlern genannt[237]. Unter einem anderen Blickwinkel kann man nach dem gleichen Grundschema anstelle der Editierdistanzen Ähnlichkeitswerte ermitteln. Das bietet sich insbesondere dann an, wenn Lücken bei der Zuordnung (also im Sinne der Editierdistanz Löschungen und Einfügungen) keine oder nur eine geringe Rolle spielen oder wenn nicht unbedingt komplette Elementfolgen miteinander verglichen werden müssen, sondern gegebenenfalls auch kürzere Folgen mit hoher Ähnlichkeit von Interesse sind, also gerade auch dann, wenn vor dem Vergleich noch gar nicht bekannt ist, ob es relevante Entsprechungen gibt.[238]

Während bei der Editierdistanz ein Wert um so günstiger ist, je niedriger er ist, ist bei einer Ähnlichkeitsmessung ein hoher Wert besser. Um die längste gemeinsame Teilsequenz zu ermitteln, kann man nach einem 1970 von Saul B. Needleman und Christian D. Wunsch vorgestellten Verfahren jede Übereinstimmung mit 1 bewerten und jede Nichtübereinstimmung, Löschung oder Einfügung mit 0.[239] Um Bereiche mit hoher Ähnlichkeit zu ermitteln, lässt sich festlegen, dass Einfügungen und Löschungen in den Randstücken der verglichenen Sequenzen irrelevant sind, nicht aber innerhalb der Grenzen der Bereiche.[240] Bei der Bewertung von Einfügungen und Löschungen kann es sinnvoll sein, nicht jede entsprechende Operation pro Element der Sequenz für sich zu bewerten, sondern zum Beispiel alle aufeinander folgenden Löschungen zusammenzufassen und auf diese Weise nur die Zahl der Unterbrechungen zu zählen.[241] Und eine Bewertung kann davon abhängig gemacht werden, welche Elemente jeweils betroffen sind – zum Beispiel ist im Hinblick auf den Vergleich von über eine Schreibmaschine oder Tastatur erfassten Texten leicht ersichtlich, dass manche Buchstabenpaare aufgrund der Tastenanordnung häufiger

[236] Im in Tabelle 2.2 gezeigten Beispiel entspricht die Folge der grau markierten Zellen einem Editierskript, das sowohl nach der dort verzeichneten Bewertung als auch entsprechend dem eben genannten Bewertungsschema optimal ist. Nach der Tabelle wäre aber zum Beispiel auch das Editierskript *ccu-ucucc* optimal, das nur drei unveränderte Zeichen vorsieht.

[237] Wagner/Fischer 1974, S. 168.

[238] Vgl. Gusfield 1997, S. 225–235, insbesondere S. 230 ff. über das „local alignment", also die lokale Alinierung.

[239] Needleman/Wunsch 1970, S. 444 f. Dieses Verfahren aus dem Bereich der Bioinformatik weist große Ähnlichkeiten zur Berechnung der Editierdistanz nach dem Algorithmus von Wagner/Fischer auf. Eine direkte Entsprechung zur Formel in Wagner/Fischer 1974 für die Berechnung der Werte der Tabellenzellen findet sich in einem Aufsatz von D. Sankoff, der auf dem Ansatz von Needleman/Wunsch aufbaut (Sankoff 1972).

[240] Vgl. Gusfield 1997, S. 228 f.

[241] Vgl. ebd. S. 235 f.

ausgetauscht werden als andere, und im Hinblick auf Textdateien, die per OCR erstellt wurden, dass mit typischen Verwechslungen zwischen bestimmten Buchstaben zu rechnen ist. Die angeführten Optionen sollen natürlich keine vollständige Zusammenstellung sein, aber erkennen lassen, dass die Art des Bewertungsschemas sinnvollerweise auf den jeweiligen Verwendungszweck abgestimmt werden sollte, um eine dafür optimale Alinierung zu erzielen.

Ein praktisches Problem bei der Ermittlung einer längsten gemeinsamen Teilsequenz (oder einer Alinierung nach anderen Kriterien) liegt in dem schnell steigenden Aufwand beim Vergleich längerer Elementfolgen. Der Zeitbedarf für das vorgestellte Verfahren ergibt sich aus der Anzahl der Tabellenzellen, also aus dem Produkt der Längen von Ausgangs- und Ergebnisstring.[242] In ähnlicher Weise hängt davon bei unveränderter Umsetzung des vorgestellten Ansatzes auch der Speicherbedarf ab.[243] Wenn es allerdings ausreicht, die Editierdistanz zu ermitteln, ist es nicht erforderlich, die komplette Tabelle zu speichern. Für die Berechnung der einer Zelle zuzuordnenden Editierdistanz sind alle Zeilen oberhalb der unmittelbar darüber stehenden nicht relevant, und auch in der Zeile darüber werden die Zellen vor der Spalte links von der vor der aktuellen Zelle nicht mehr benötigt, so dass die Zahl der gespeicherten Zellenwerte die Anzahl der Spalten (oder bei entsprechend anderer Bearbeitungsreihenfolge die der Zeilen) nur um 1 übersteigen muss.[244] Auch der Speicherbedarf für die Ermittlung einer entsprechenden Alinierung kann zum Beispiel nach einem von D. S. Hirschberg entwickelten Algorithmus mithilfe einer rekursiven Zerlegung in Teilprobleme so reduziert werden, dass er nur linear mit der Länge des kürzeren Strings wächst, wobei sich der Zeitbedarf gegenüber dem vorgestellten Verfahren, das die Speicherung der kompletten Tabelle erfordert, schlimmstenfalls verdoppelt.[245]

Das reduziert die Schwierigkeit zwar enorm, das Verfahren überhaupt für den Vergleich längerer Strings einzusetzen,[246] in der Praxis zeigt sich aber, dass der

[242] Vgl. Wagner/Fischer 1974, S. 172.

[243] Die Zahl der Tabellenzeilen und -spalten ist jeweils um 1 größer als die Länge der verzeichneten Strings; das muss auch bei der Berechnung des Speicherbedarfs berücksichtigt werden, fällt allerdings für den Zeitbedarf nicht ins Gewicht, da sich die Werte der ersten Zeile und Spalte schneller berechnen lassen als die übrigen (zum vorangehenden Wert muss nur jeweils der einer Einfügung beziehungsweise Löschung zugeordnete Betrag addiert werden).

[244] Gusfield 1997, S. 254 f.

[245] Vgl. ebd. S. 255–259.

[246] Gusfield weist ebd. S. 254 darauf hin, dass bei der Stringalinierung häufig der Speicherbedarf die Grenzen der Verarbeitbarkeit bestimmt. Dass das vorgestellte Verfahren der Alinierung mithilfe einer vollständigen Tabelle nicht für größere Textmengen geeignet ist, lässt sich leicht an einem Rechenbeispiel erläutern: Der Vergleich von zwei Texten mit jeweils einer Million Zeichen (bei einer Zahl von 2000–3000 Zeichen pro Seite also zwischen 333 und 500 Seiten, was ja kein extremer Wert ist), ergäbe eine Billion Tabellenzellen – selbst wenn jeder Tabellenwert in einem einzigen Byte gespeichert werden könnte (was schon dann nicht geht, wenn die Editierdistanz größer als 255 werden kann), entspräche das einem Speicherbedarf von einem Terabyte.

Zeitbedarf bei der Ermittlung längster gemeinsamer Teilsequenzen schon beim Vergleich relativ kurzer Elementfolgen so groß werden kann, dass das Verfahren für den systematischen Vergleich großer Textmengen kaum geeignet ist.[247] Das Problem lässt sich deutlich verkleinern, wenn der Vergleich nicht auf Zeichen-, sondern auf Wortebene erfolgt,[248] aber auch damit lässt sich kein Laufzeitverhalten erreichen, das den Vergleich größerer Textmengen wirklich praktikabel machen würde.[249]

[247] Eine Implementierung des Hirschberg-Algorithmus lag mir leider nicht vor, wohl aber das *Perl*-Modul *Text::WagnerFischer* (*http://search.cpan.org/~davidebe/Text-WagnerFischer-0.04/WagnerFischer.pm*) – und der tatsächliche Zeitbedarf des Wagner-Fischer-Algorithmus ist niedriger als der des Hirschberg-Algorithmus, wenn der Aufwand für die Speicherallozierung nicht überhandnimmt. Mit diesem Modul wurden auf einem etwas älteren Rechner für den Vergleich von zwei Strings mit jeweils 1.000 Zeichen über 8 Sekunden benötigt, bei jeweils 2.000 Zeichen waren es über 33 Sekunden. Ein von James W. Hunt und anderen für das Programm *diff* entwickelte Algorithmus (vgl. Hunt/McIlroy 1976/2012 sowie Hunt/Szymanski 1977) zur Ermittlung der längsten gemeinsamen Teilsequenz weist im Regelfall einen nur linear steigenden Speicherbedarf und ein auch im schlechtesten Fall nur wenig ungünstigeres Laufzeitverhalten als der Wagner-Fischer-Algorithmus auf. Das *Perl*-Modul *Algorithm::Diff* (*http://search.cpan.org/~tyemq/Algorithm-Diff-1.1902/lib/Algorithm/Diff.pm*), das darauf basiert, benötigte für die Ermittlung der längsten gemeinsamen Teilsequenz für zwei Strings bei einer jeweiligen Länge von 1.000 Zeichen etwas mehr als 1 Sekunde, bei jeweils 5.000 Zeichen ca. 40 Sekunden und bei jeweils 10.000 Zeichen ca. 250 Sekunden. Dabei wurden allerdings Zeichenfolgen verglichen, die auf einem kleinen Zeicheninventar beruhen, so dass die Einzelzeichen und auch viele Zeichenkombinationen sehr häufig vorkommen.

[248] Eine noch stärkere Vereinfachung durch den Vergleich auf Zeilenebene, wie er etwa für die Ermittlung der Unterschiede zwischen zwei Quellcode-Versionen eines Computerprogramms üblich und – bei nicht zu großen Änderungen – auch sinnvoll ist, scheidet für die Suche nach vorher unbekannten Übereinstimmungen zwischen Texten aus, da eine exakte Übereinstimmung des Zeilenumbruchs auch bei prinzipieller Textgleichheit einen besonderen Zufall darstellt, soweit sie nicht durch eine Versstruktur motiviert ist oder darauf beruht, dass der eine Text auch hinsichtlich des Seitenlayouts als Vorlage für den anderen gedient hat (wie es teilweise bei Neuauflagen der Fall ist). Eher in Betracht käme ein Vergleich von Sätzen, aber in vielen Fällen lassen sich relevante Übereinstimmungen zwischen kürzeren Textfragmenten feststellen, die damit nicht gefunden würden. Abgesehen davon ist eine Satzgliederung jedenfalls für das hier untersuchte Textkorpus in vielen Fällen nicht sicher vorzunehmen, so dass schon durch eine unterschiedliche Einteilung Übereinstimmungen übersehen werden können.

[249] Es handelt sich selbstverständlich nicht um ein „unhandhabbares“ Problem im Sinne der Komplexitätstheorie (vgl. dazu zum Beispiel Harel/Feldman 2006, S. 191 ff.), da sich der Aufwand aus dem Produkt der Anzahl der zu vergleichenden Einheiten ergibt, also nicht exponentiell steigt, sondern allenfalls (bei gleich großer Anzahl in beiden Texten) quadratisch. Deshalb lässt sich bei einem Vergleich auf Wortebene die längste gemeinsame Teilsequenz von Texten im Umfang von einigen tausend Wörtern noch relativ schnell ermitteln (vgl. die Angaben oben in Anm. 247), wenn es aber um Texte im Umfang von mehreren hunderttausend Wörtern geht, wie sie zum Beispiel im hier untersuchten Korpus vorkommen, dürfte die Anwendung des Verfahrens derzeit wohl allenfalls für einzelne Textpaare, bei denen das Ergebnis von besonderem Interesse ist, in Betracht kommen.

2.2 Verfahren und Programme der Bioinformatik

Während in Kapitel 2.1 grundlegende Verfahren zum Vergleich von Strings vor allem mit Blick auf die theoretische Konzeption vorgestellt wurden, soll es hier insbesondere um praktische Verfahren und Lösungen gehen, die im Zusammenhang mit dem Vergleich von DNA- und Aminosäuresequenzen entwickelt wurden.

Die Probleme, um die es dabei geht, lassen sich teilweise sehr gut auf den Vergleich von Texten übertragen. Allerdings sind viele Programme darauf zugeschnitten, dass die zu vergleichenden Strings nur ein bestimmtes Zeicheninventar enthalten. Deshalb geht es in Unterkapitel 2.2.1, das die Ermittlung von sogenannten *maximal exact matches* vorstellt, insbesondere auch um die Frage, inwieweit die untersuchten Programme für die Verarbeitung von Textdateien geeignet sind.[250] Anschließend bietet Unterkapitel 2.2.2 eine Einführung in das *Dotplot*-Visualisierungsverfahren.

2.2.1 *Maximal exact matches (MEMs)*

Seit 1999 stehen in zunehmenden Maße vollständig sequenzierte Genome zur Verfügung, die so nahe miteinander verwandt sind, dass ihr Vergleich sinnvoll erscheint.[251] Die Größenordnung der dabei untersuchten Datenmengen – teilweise geht es um Milliarden Basenpaare – stellt die Bioinformatik vor besondere Herausforderungen.

Für den Genomvergleich wegweisend war der im Programmpaket *MUMmer* gewählte Ansatz, sogenannte *maximal unique matches (MUMs)*, ab Version 2 *MUM-candidates* oder ab Version 3 wahlweise generell *maximal exact matches (MEMs)* zu ermitteln und als Ankerpunkte für den weiteren Vergleich zu verwenden. *Maximal exact matches* sind Übereinstimmungen, die nicht weiter ausgedehnt werden können (da die Zeichen davor und danach in den beiden verglichenen Sequenzen nicht übereinstimmen). *Maximal unique matches* sind *maximal exact matches*, die in beiden Sequenzen nur ein einziges Mal vorkommen. *MUM-candidates* sind zwar in der sogenannten Referenzsequenz nur einmal vorhanden, können aber in der mit ihr verglichenen Abfragesequenz möglicherweise mehrfach vorkommen. Die eben genannten Bezeichnungen implizieren es zwar nicht, es werden aber nur Matches ermittelt, die eine bestimmte Mindestlänge aufweisen. Diese Länge kann vom Benutzer festgelegt werden, andernfalls beträgt sie in *MUMmer* 20.[252] Die folgende Darstellung ist auf *maximal exact matches* fokussiert, da es auch im Hauptteil der

[250] Die *Open-Source*-Programme wurden mit GCC (vgl. *https://gcc.gnu.org/*) in der *Cygwin*-Umgebung (vgl. *https://www.cygwin.com/*) kompiliert.
[251] Vgl. Kurtz u. a. 2004, S. R12.1.
[252] Vgl. *http://mummer.sourceforge.net/manual/#mummer*.

vorliegenden Untersuchung um die Erkennung und weitere Auswertung von MEMs geht.

Die Ermittlung der MEMs erfolgt in *MUMmer* über einen Suffixbaum der Referenzsequenz,[253] der in der Version 3 des Programms sehr speichereffizient implementiert ist.[254] Da auch Zeichen akzeptiert werden, die weder für ein Nukleotid noch für eine Aminosäure stehen,[255] lässt sich das Programm prinzipiell auch für den Vergleich von Textdaten in einem 8-Bit-Zeichensatz nutzen. Allerdings setzt *MUMmer* – wie auch die meisten weiteren in diesem Unterkapitel beschriebenen Programme – eigentlich voraus, dass die einzulesenden Daten im *multi-FASTA*-Format vorliegen. Das bedeutet, dass eine Eingabedatei mehrere Sequenzen enthalten kann, wobei für jede Sequenz zunächst in einer mit „>“ beginnenden Zeile ein Identifikator anzugeben ist und in den Folgezeilen die Sequenz selbst. Ansonsten sind *Whitespace*-Zeichen (also Leerzeichen, Tabulatoren und Zeilenumbrüche) bedeutungslos, und zwischen Groß- und Kleinbuchstaben wird nicht unterschieden.[256] Auf mögliche Anpassungsmaßnahmen soll unten in Unterkapitel 3.2.1 eingegangen werden.

[253] In Version 1 wurden beide verglichenen Sequenzen in einen gemeinsamen Suffixbaum eingelesen, dies wurde aber ab Version 2 geändert, um den Speicherbedarf zu reduzieren. Wenn es mehrere Referenzsequenzen gibt, müssen sie alle in einer einzigen Datei enthalten sein, wobei eine Kennzeichnung nach den Regeln des *multiFASTA*-Formats vorausgesetzt wird. Mehrere Abfragesequenzen können auf bis zu 32 Dateien verteilt werden (was den Speicherbedarf reduziert), und auch die Abfragedateien können mehrere Sequenzen enthalten (vgl. *http://mummer.sourceforge.net/manual/#mummer*).

[254] Vgl. Kurtz u. a. 2004, S. R12.3. Die im Folgenden gegebene Beschreibung basiert auf der Programmversion 3.23 (*https://sourceforge.net/projects/mummer/files/mummer/3.23/MUMmer3.23.tar.gz/download*). Inzwischen wird außerdem die Version 4 als *Beta-Release* angeboten. Die Änderungen in dieser Version betreffen laut Beschreibung auf *http://mummer.sourceforge.net/* insbesondere das zum Programmpaket gehörige *NUCmer*, das für den Vergleich von sehr ähnlichen Nukleotidsequenzen gedacht ist (vgl. *http://mummer.sourceforge.net/manual/#running*). Hier soll es aber nur um die Erkennung der MUMs, MUM-candidates beziehungsweise MEMs gehen, die vom zu *MUMmer* gehörenden Einzelprogramm *mummer* geleistet wird.

[255] Die Präprozessordirektive „#ifdef WARNINGIFNONUCLEOTIDES“ in der Datei *maxmatinp.c* ist als Konfigurationsmöglichkeit vorgesehen, um bei solchen Zeichen gegebenenfalls eine Warnung auszugeben und sie jeweils durch „n“ zu ersetzen, „WARNINGIFNONUCLEOTIDES“ ist aber im originalen Quellcode jedenfalls in der hier zugrunde gelegten Programmversion nicht definiert. Nukleotide sind die wohl bekannten Bestandteile von DNA-Sequenzen. Sie (beziehungsweise die für die genetische Information entscheidenden zugehörigen Basen) werden durch die vier Buchstaben A, C, G und T repräsentiert. Für die Beschreibung der (entsprechend den Informationen in der DNA gebildeten) Proteine als Polypeptidketten dient ein Alphabet mit zwanzig Buchstaben, die für die zwanzig Aminosäuren stehen, die die Hauptbausteine von Proteinen bilden (vgl. zum Beispiel Böckenhauer/Bongartz 2003, S. 16–19).

[256] Eine detailliertere Beschreibung des Formats findet sich zum Beispiel in der deutschen *Wikipedia* (*https://de.wikipedia.org/wiki/FASTA-Format*). Nach der Darstellung dort sieht das Format auch vor, dass auf die mit „>“ beginnenden Kopfzeilen Kommentarzeilen folgen können, die durch „;“ eingeleitet werden. Das ist in den hier untersuchten Programmen aber anscheinend nicht (oder nicht in jedem Fall) vorgesehen und laut *https://en.wikipedia.org/wiki/FASTA_format* auch nicht mehr üblich. Die im Format vorgesehene Begrenzung der Zeilenlänge ist für die Verarbeitung durch die hier betrachteten Programme wohl nicht (oder nicht bei allen) relevant.

Der *MUMmer* zugrunde liegende Ansatz, einen Vergleich auf der Basis von exakten Übereinstimmungen mit einer bestimmten Mindestlänge durchzuführen, hat sich offenbar für die Bioinformatik sehr bewährt. Allerdings kann der Speicherbedarf des Suffixbaums auch in dieser sehr effizienten Implementierung[257] Probleme bereiten.[258] Deshalb wurden für die Ermittlung solcher Übereinstimmungen eine Reihe weiterer Programme entwickelt, die verschiedene Ansätze nutzen, um gleiche Resultate mit einem geringeren Aufwand zu erreichen. Das Ausgabeformat ist dabei oft weitgehend das gleiche wie bei *MUMmer*.[259]

Als wohl älteste im Hinblick auf den Speicherbedarf günstigere Alternative zum Suffixbaum wurde das sogenannte Suffixarray konzipiert, das den Zugriff auf die Suffixe eines Strings über ein alphabetisch sortiertes Array bietet.[260] Die unter der Bezeichnung *enhanced suffix array (ESA)* bekannte Kombination eines Suffixarrays mit verschiedenen Zusatztabellen ermöglicht es, die gleiche Zeitkomplexität wie bei Suffixbäumen zu erreichen, wobei der Speicherbedarf jedenfalls für Strings auf der Basis von kleinen Alphabeten geringer ist.[261]

Das Programmpaket *Vmatch*[262] nutzt unter anderem diesen Ansatz und arbeitet mit einem in verschiedenen Dateien gespeicherten Index. Es hat nach der Erstellung des Index nur einen sehr geringen Speicherbedarf, benötigte aber in einem Testlauf mit dafür aufbereiteten Textdateien – anscheinend anders als bei der Verarbeitung von DNA-Sequenzen[263] – im Vergleich zu anderen in diesem Unterkapitel vorgestellten Programmen viel mehr Zeit für die Ermittlung exakter Übereinstimmungen. Über *Vmatch* können auch Übereinstimmungen mit einer begrenzten Zahl von Abweichungen ermittelt werden, allerdings wird im Handbuch des Programms darauf hingewiesen, dass dies – abhängig von der Zahl der zugelassenen Unterschiede – zu einer starken Erhöhung der Laufzeit führen kann.[264] Aus unbekannten Gründen ließen sich im Test entgegen der Beschreibung im Handbuch[265] nur 7-Bit-ASCII-Zeichen einlesen, also zum Beispiel keine Umlaute. Leerzeichen kön-

[257] Je nach Größe der Sequenz werden laut Kurtz u. a. 2004, S. R12.3 ungefähr 12,5 beziehungsweise 15,4 Bytes je Basenpaar benötigt.

[258] Vgl. zum Beispiel die Empfehlungen zum möglichst speichereffizienten Vergleich im Handbuch des Programms (*http://mummer.sourceforge.net/manual/#mummer*).

[259] Die Reihenfolge der ausgegebenen Match-Informationen stimmt nicht unbedingt überein, und die Formatierung der Spalten durch *Whitespace* kann abweichen.

[260] Vgl. Manber/Myers 1993.

[261] Vgl. Abouelhoda/Kurtz/Ohlebusch 2004. Ebd. S. 84 zeigt eine Übersicht den Speicherbedarf der verschiedenen Tabellen und gibt an, welche Tabellen für welche der untersuchten Aufgaben benötigt werden.

[262] *Vmatch* wird in kompilierter Form zur Verfügung gestellt (*http://www.vmatch.de/download.html*). Für die folgenden Informationen zu *Vmatch* wurde die Programmversion 2.3.0 zugrunde gelegt.

[263] Vgl. die Angaben in Khan u. a. 2009, S. 1614 dokumentierten Messungen. Danach hat das Programm eine ähnliche Verarbeitungsgeschwindigkeit wie das im Folgenden noch beschriebene *sparseMEM*. Die Gründe für diese Abweichung können hier nicht geklärt werden.

[264] Kurtz 2016, S. 35 f.

[265] Ebd. S. 1.

nen zwar enthalten sein, werden aber bei den Positionsangaben offenbar nicht berücksichtigt.[266]

Eine weitere Möglichkeit, den Speicherbedarf zu reduzieren, besteht darin, nicht alle Suffixe zu verzeichnen, sondern nur eine nach bestimmten Kriterien gebildete Auswahl. Die Idee, in einem sogenannten *sparse suffix tree* nur jedes *K*te Suffix aufzuführen, wurde 1996 vorgestellt.[267] Später wurde sie auch auf Suffixarrays übertragen und wird in den Programmen *sparseMEM* und *essaMEM* für die Ermittlung von MEMs eingesetzt.[268] Einer der Grundgedanken ist dabei offenbar, dass dann, wenn ein Match der geforderten Mindestlänge L an einer Stelle beginnt, deren Suffix im *sparse suffix tree* nicht verzeichnet ist, das nächste verzeichnete Suffix immerhin mindestens $L - K + 1$ Zeichen des Matchs enthält (wenn K nicht größer als L ist).[269] Dementsprechend lassen sich über die verzeichneten Suffixe zunächst Matches ermitteln, die diese Länge aufweisen, und in einem zweiten Schritt kann das Umfeld überprüft werden.[270] Im Vergleich zu einem vollständigen Suffixbaum oder -array sind also zusätzliche Verarbeitungsschritte für die MEM-Ermittlung erforderlich, bei $K = 1$ ist der in Testläufen ermittelte Zeitbedarf allerdings ebenso wie der Speicherbedarf deutlich niedriger als bei *MUMmer*, und bei etwas größeren Werten für K steigt zwar die Verarbeitungszeit, aber der Speicherbedarf erreicht noch niedrigere Werte.[271]

Das Programm *essaMEM*[272] baut auf *sparseMEM* auf und fügt mit dem *sparse child table* eine mit gewissen Anpassungen vom *enhanced suffix array* übernommene Tabelle hinzu.[273] Wie bei *Vmatch* werden die Tabellen, aus denen sich der Index zusammensetzt, in mehreren Dateien gespeichert. Welche Tabellen für die MEM-Ermittlung verwendet werden, kann teilweise über zusätzliche Parameter beim Programmaufruf gesteuert werden; wenn diese Parameter fehlen, hängt die Auswahl von der Größe von K ab. Außerdem sieht *essaMEM* auch eine Ausdünnung bei der Verarbeitung des Abfragestrings vor.[274] *essaMEM* erreicht durch diese

[266] Dies gilt nach den Beobachtungen im Testlauf auch bei Verwendung des *plain*-Formats (vgl. aber Kurtz 2016, S. 6).

[267] Vgl. Kärkkäinen/Ukkonen 1996. Dort wird k (in Kleinbuchstaben) verwendet. Aus Darstellungsgründen erfolgt hier eine Angleichung an die Form in Khan u. a. 2009.

[268] Vgl. Khan u. a. 2009 sowie Vyverman u. a. 2013a. *sparseMEM* und *essaMEM* stehen auf *Github* zum Download zur Verfügung (*https://github.com/zia1138/sparseMEM* und *https://github.com/readmapping/essaMEM*).

[269] Vgl. Khan u. a. 2009, S. 1611. Dort wird der äquivalente Term $L - (K - 1)$ verwendet, aber wohl nicht begründet.

[270] Ebd. S. 1611–1613 wird gezeigt, wie dies sehr effizient durchgeführt werden kann.

[271] Vgl. ebd. S. 1614.

[272] Vgl. Vyverman u. a. 2013a und Vyverman u. a. 2013b.

[273] Vgl. Vyverman u. a. 2013b, S. 2 f.

[274] Auch dafür kann der automatisch festgelegte Wert gegebenenfalls über einen Parameter verändert werden. Eine Kurzübersicht über die Parameter wird beim Programmaufruf mit dem Parameter

Änderungen jedenfalls für den Vergleich von DNA-Daten, dass bei einer Erhöhung von *K* der Zeitaufwand viel weniger ansteigt als bei *sparseMEM*.[275]

sparseMEM und *essaMEM* setzen voraus, dass Leerzeichen in Sequenzdaten allenfalls am Anfang oder Ende von Zeilen vorkommen und bedeutungslos sind und dass zwischen Groß- und Kleinbuchstaben nicht zu unterscheiden ist. Darüber hinaus gibt es weitere Annahmen im Quellcode hinsichtlich des zu verarbeitenden Zeichenbestands. Auch hier gibt Unterkapitel 3.2.1 Hinweise zu Anpassungsmöglichkeiten für die Verarbeitung von Textdaten. Schon hier sei angemerkt, dass sich in den dort dokumentierten Tests in gewissen Konstellationen Probleme zeigten, die im Rahmen dieser Untersuchung nicht behoben werden konnten.

Auch das Programm *backwardMEM*[276] legt eine gegenüber einem vollständigen Suffixarray ausgedünnte Datenstruktur zugrunde, wobei die Auswahl auf einem anderen Auswahlkriterium beruht und der Ausdünnungsfaktor mit *k* (in Kleinschreibung) bezeichnet wird.[277] Im *Makefile* sind für *k* bestimmte Werte vorgesehen, für die jeweils eine eigene *.exe*-Datei kompiliert wird. Auf die bei der Datenindizierung und der MEM-Ermittlung eingesetzten Techniken soll hier nicht weiter eingegangen werden.[278] Wie bei *sparseMEM* und *essaMEM* werden bei Verwendung des unveränderten Quellcodes Leerzeichen am Anfang und Ende von Zeilen entfernt, und auch hier erfolgt automatisch eine Umwandlung in Kleinbuchstaben.[279]

Die MEM-Erkennung in *slaMEM*[280] baut auf dem Algorithmus auf, der *backwardMEM* zugrunde liegt.[281] *slaMEM* hat sich in Tests der Entwickler als ähnlich effizient wie *essaMEM* und deutlich vorteilhaft gegenüber den anderen hier vorgestellten Programmen erwiesen.[282] Allerdings ist der Programmcode auf die Anwendung auf DNA-Sequenzen zugeschnitten,[283] und der Versuch, das Programm

-h gezeigt. In Vyverman u. a. 2013b werden unter anderem nähere Informationen dazu und Vergleichsdaten für verschiedene Aufrufvarianten gezeigt.

[275] Vgl. Vyverman u. a. 2013a, S. 803 f. sowie Vyverman u. a. 2013b, S. 11–16, 18, 20 und 22 mit Beispieldaten für die *K*-Werte 1, 2, 4, 8, 16 und 32.

[276] Es kann von *http://www.uni-ulm.de/in/theo/research/seqana/* aus heruntergeladen werden. Beim Kompilieren auf dem Testsystem waren entsprechend den Compiler-Fehlermeldungen kleinere Anpassungen erforderlich (Ergänzung von „#include <stdio.h>“ in *sparseMEM/sparseSA.cpp* sowie Ergänzung von „=NULL“ in der Deklaration in Zeile 41 und Entfernung von „=NULL“ in der Definition in Zeile 70 in *backwardMEM/sdsl-0.7.3/src/sdsl/select_support_bs.hpp*). Vgl. außerdem die Hinweise unter der angegebenen URL.

[277] Vgl. Ohlebusch/Gog/Kügel 2010, S. 356 f.

[278] Vgl. dazu Ohlebusch/Gog/Kügel 2010, S. 347–355.

[279] Wie bei den übrigen hier beschriebenen Programmen gilt das nur für die Grundbuchstaben des lateinischen Alphabets.

[280] *slaMEM* steht auf *GitHub* zur Verfügung (*http://github.com/fjdf/slaMEM/*).

[281] Vgl. Fernandes/Freitas 2013/14, S. 468 f.

[282] Vgl. ebd. S. 469 f.

[283] Vgl. zum Beispiel die Belegung der Konstante *ALPHABETSIZE* mit dem Wert 6 in der Datei bwtindex.c. Dabei ist neben A, C, G und T sowie dem Abschlusszeichen $ (und dem gleich

für die MEM-Ermittlung in Textdateien zu nutzen, brachte keine verwendbaren Ergebnisse.

Schließlich soll noch *E-MEM* genannt werden.[284] Auch *E-MEM* basiert auf dem Grundgedanken, zunächst etwas kürzere Matches über eine Zugriffsstruktur zu ermitteln und in einem darauf aufbauenden Schritt die tatsächliche Matchlänge festzustellen. Es legt dabei aber keinen Suffixbaum und kein Suffixarray zugrunde, sondern eine Hashtabelle, in der sogenannte *k-mers* verzeichnet werden, das heißt Zeichenfolgen der Länge k.[285] Dabei muss aber nicht jedes *k-mer* eingetragen werden, vielmehr reicht es, die Positionen auszuwählen, die ein Vielfaches von $L - k + 1$ sind, wobei L wie schon oben in der Beschreibung der *sparse suffix arrays* die geforderte Mindestmatchlänge bezeichnet.[286] Auch der eben genannte Term entspricht dem oben angeführten, wobei man allerdings wohl von einer umgekehrten Verwendung sprechen kann: Während bei *sparse suffix arrays* K der Ausdünnungsfaktor ist und $L - K + 1$ die sich daraus ergebende Länge, die im *sparse suffix array* für jedes MEM mindestens verzeichnet sein muss, ist hier $L - k + 1$ der Ausdünnungsfaktor und k die Länge, für die zunächst die Übereinstimmungen ermittelt werden.

Das Verfahren, ausgehend von Übereinstimmungen mit Einträgen in der Hashtabelle zu ermitteln, welche davon Teile von MEMs der geforderten Mindestlänge sind, ist hocheffizient gestaltet, und insgesamt ergeben sich im Hinblick auf Laufzeit und Speicherbedarf nach den Messungen der Entwickler deutliche Vorteile gegenüber den übrigen hier vorgestellten Programmen.[287] Allerdings mag dabei eine Rolle spielen, dass nur die Ermittlung von MEMs mit der Mindestlänge 100 untersucht wurde. Wie unten in Unterkapitel 3.2.3 noch gezeigt werden wird, nimmt die Zahl der MEMs bei allmählicher Reduzierung der Mindestlänge immer stärker zu. Dementsprechend übersteigt die Zahl der über die Hashtabelle ermittelten kürzeren Übereinstimmungen die Zahl der MEMs bei kurzen Mindestlängen mit Sicherheit in einem viel stärkeren Maße.

In welchem Maße sich das auswirkt und inwieweit es dabei Unterschiede zu anderen hier vorgestellten Verfahren gibt, kann hier nicht untersucht werden.[288]

---

behandelten Nullzeichen) auch der Code N vorgesehen, der für eine beliebige Base steht (vgl. *https://en.wikipedia.org/wiki/Nucleic_acid_sequence*).

[284] Der Quellcode kann von *http://www.csd.uwo.ca/~ilie/E-MEM/* aus heruntergeladen werden.

[285] Es handelt sich also um Zeichen-N-Gramme. Vgl. dazu unten Unterkapitel 2.3.3.

[286] Khiste/Ilie 2014/15, S. 510.

[287] Vgl. zu den eingesetzten Methoden ebd. S. 510 f., zum Vergleich mit anderen Programmen zur MEM-Ermittlung ebd. S. 511–513.

[288] In einem unabhängig von *E-MEM* entwickelten eigenen Verfahren zur MEM-Ermittlung, das teilweise auf ähnlichen Grundansätzen beruht, aber weniger elaboriert und deshalb nicht unbedingt vergleichbar ist, ist der Aufwand stark von der angesetzten Mindestlänge abhängig. Da mit einigen der hier beschriebenen Programme nach den unten in Unterkapitel 3.2.1 beschriebenen Anpassungsmaßnahmen auch für den Vergleich von Textdaten bessere Alternativen zur Verfü-

E-MEM ist für die Verarbeitung von DNA-Sequenzen konzipiert, nutzt die Tatsache, dass für die Speicherung eines Zeichens des genetischen Codes nur 2 Bit benötigt werden, und arbeitet auch beim Vergleich mit Bit-Operationen. Dementsprechend lässt es sich für Textdaten nicht sinnvoll verwenden.

Insgesamt kann festgestellt werden, dass in den letzten Jahren eine Reihe von Programmen für die Ermittlung von *maximal exact matches* (beziehungsweise *maximal unique matches* oder *MUM-candidates*) geschaffen wurden, von denen jedenfalls *MUMmer* und *backwardMEM* sehr gut geeignet sind, auch auf Textdaten in einem 8-Bit-Zeichensatz angewandt zu werden, wenn kleinere Anpassungen vorgenommen werden und wenn diese Daten bestimmte Zeichen nicht enthalten, denen bei der Verarbeitung eine Sonderbedeutung zukommt.

Für diese beiden Programme sowie für *sparseMEM* und *essaMEM* wird unten im Rahmen der technischen Beschreibung in Unterkapitel 3.2.1 im Detail dargestellt, wie diese Veränderungen durchzuführen sind und an welchen Stellen sich bei der Verarbeitung des Untersuchungskorpus Probleme zeigten.

2.2.2 Dotplots zur Ermittlung von ähnlichen Bereichen

Die Ermittlung einer nach den jeweils angesetzten Kriterien optimalen globalen Sequenzalinierung ist zwar, wie in Unterkapitel 2.1.2 beschrieben, aufgrund des quadratisch von der Textmenge abhängigen Zeitaufwands für umfangreichere Texte nur schwer einsetzbar, soweit es aber primär darum geht, die Häufung von Übereinstimmungen in Teilbereichen und insbesondere Teilstücke mit exakter Gleichheit zu erkennen, bietet sich eine Visualisierungsform an, die in der Bioinformatik entwickelt wurde und in diesem Fachgebiet sowie in der *Text-Reuse*-Forschung unter dem Namen *Dotplot* bekannt ist.[289]

---

gung stehen und da mit einer näheren Beschreibung keine neuen Ideen vorgebracht würden, soll es hier nicht vorgestellt werden. Unten sind in den Tabellen 3.4 und 3.5 auf S. 139 und 140 Daten zu mehreren der in diesem Unterkapitel beschriebenen Programme zusammengestellt. Daraus kann entnommen werden, dass der Zeitaufwand je nach Programm und Konfiguration für sehr kurze Mindestlängen teilweise viel höher ist als für größere Mindestlängen.

[289] Die Bezeichnung wird auch in anderen Zusammenhängen verwendet und besagt als solche nichts darüber, was in Punktform (oder auch mit anderen graphischen Formen) dargestellt wird, ist aber in der Bioinformatik anscheinend nur für den Sequenzvergleich üblich. Vgl. zum Beispiel Schulz/Leese/Held 2008 oder die Artikel in der deutschen und der englischen *Wikipedia* (*http://de.wikipedia.org/wiki/Dotplot* und *http://en.wikipedia.org/wiki/Dot_plot_%28bioinformatics%29*). Gibbs/McIntyre 1970 als wohl erste Publikation dazu verwendet den Terminus noch nicht; dort ist von der „diagram method“ die Rede. Früher war wohl auch die Bezeichung *dot matrix* gängig (so in Vihinen 1988 und Vingron/Argos 1991), die die damit verbundene mathematische Struktur stärker zum Ausdruck bringt. Lee 2007, S. 474 Abb. 1 zeigt einen Dotplot für den Vergleich des Markus- und des Lukas-Evangeliums. Büchler 2013, S. 35 f. weist auf die Nutzung des Verfahrens für den Textvergleich hin.

Auch dieses Verfahren basiert darauf, den Vergleich über eine Matrix durchzuführen, in der die für den Vergleich zugrunde gelegten Teilstücke des einen Textes durch Zeilen repräsentiert werden und die des anderen durch Spalten.[290] Wenn jede einzelne Übereinstimmung (oder als hinreichend betrachtete Ähnlichkeit) der Teilstücke einer Zeile und Spalte durch eine Hervorhebung in der entsprechenden Zelle verzeichnet wird, ergibt sich für ein mehrere Teilstücke umfassendes Match in der Matrix eine entsprechende Folge von Hervorhebungen, die jeweils um eine Zeile und Spalte versetzt sind, bei entsprechender graphischer Präsentation also eine diagonale Linie. Solche Linien springen unmittelbar ins Auge, insbesondere wenn sie etwas länger sind und die Zahl kurzer Übereinstimmungen nicht extrem hoch ist, und auch wenn die Linien aufgrund von abweichenden Zwischenstücken unterbrochen oder aufgrund von Einschüben beziehungsweise Auslassungen versetzt zueinander angeordnet sind, werden sie – jedenfalls bei nicht zu großem Abstand oder zu vielen nicht zugehörigen Punkten im Umfeld – als irgendwie zusammengehörig wahrgenommen. Neben einer einfachen Form, die nur unterscheidet, ob an einer bestimmten Stelle eine Übereinstimmung vorliegt oder nicht, sind graduelle Verzeichnungsverfahren üblich, bei denen zu jeder Einzelposition festgehalten wird, in welchem Maße ein Umfeld von bestimmtem Umfang Übereinstimmungen aufweist.[291]

Mathematisch lässt sich die Zugehörigkeit eines Matchs zu einer Diagonale im Diagramm durch eine Abbildung des jeweiligen Positionspaars auf die Differenz zwischen den Positionen fassen, die bei jedem Punkt auf dieser Linie gleich ist. Zudem kann über den Vergleich mit den auf der Diagonale unmittelbar benachbarten Zellen für jede einzelne Übereinstimmung festgestellt werden, ob sie unmittelbar mit anderen übereinstimmenden Teilstücken verbunden ist, und bei Fortsetzung dieser Prüfung bis zu einer Abweichung lässt sich die Gesamtlänge eines zusammenhängenden Matchs ermitteln. Schon in der wohl ersten Publikation zum Dotplot-Verfahren von Adrian J. Gibbs und George A. McIntyre aus dem Jahr 1970 werden auf der Basis der Diagonalenzuordnung und der Länge der einzelnen Matches Bewertungsverfahren für die Einschätzung der Ähnlichkeit zweier Nukleotid- oder Aminosäuresequenzen vorgestellt.[292]

Auch wenn eine Dotplotdarstellung oft eine intuitive Aussage über Textbereiche mit hoher Ähnlichkeit ermöglicht, ist das Bild doch keineswegs immer klar, da es wesentlich davon abhängt, welches Ausmaß irrelevante Übereinstimmungen

[290] Das Verfahren wird in Gibbs/McIntyre 1970 vorgestellt unter Hinweis auf die Ähnlichkeit zu einer in Fitch 1969 präsentierten, auf den oben genannten Saul B. Needleman zurückgehenden Methode, die offenbar in engem Zusammenhang zum Needleman-Wunsch- beziehungsweise zum Wagner-Fischer-Algorithmus steht.

[291] Ein Beispiel für den Effekt einer solchen Kontextberücksichtigung bietet Schulz/Leese/Held 2008 in Abbildung 2.

[292] Gibbs/McIntyre 1970. Vgl. Schulz/Leese/Held 2008 zu neueren Analyseverfahren und weiterer Literatur.

haben und inwieweit als signifikant einzuschätzende Matches durch Unterbrechungen oder auch Verschiebungen aufgrund von Auslassungen beziehungsweise Einschüben voneinander getrennt sind. Entsprechend ist natürlich auch eine rechnerische Ermittlung solcher Bereiche nur begrenzt möglich und von Annahmen beziehungsweise Festlegungen abhängig, zum Beispiel, ob die Zugehörigkeit zu einer gemeinsamen Diagonale auch dann relevant ist, wenn der Abstand zwischen zwei Matches groß ist. Und insgesamt ist natürlich festzuhalten, dass auch bei diesem Verfahren der Aufwand vom Produkt der Längen der zu vergleichenden Strings abhängt und deshalb beim Vergleich längerer Texte leicht ein problematisches Ausmaß erreichen kann. Auch die Darstellung wird hierbei schwierig, da die Zahl der Matrixzellen schnell zum Beispiel die Zahl der Pixel eines Computerdisplays überschreitet und Übereinstimmungen, wenn sie nicht übergroß dargestellt werden, bei entsprechender Verkleinerung des Gesamtbildes kaum zu sehen sind. Insbesondere wenn Übereinstimmungen insgesamt eher selten und schon bekannt sind, kann dieses Problem aber dadurch reduziert werden, dass nicht jede kleinste Einheit durch eine Matrixzelle repräsentiert wird, sondern größere Stücke zusammengefasst werden und nur jeweils festgehalten wird, dass ein solches Teilstück einen übereinstimmenden Bereich enthält. Entsprechende Visualisierungen von Übereinstimmungen werden unten in Unterkapitel 3.4.3 vorgestellt.[293]

Auf der Basis von Matrizen entsprechend dem Dotplotprinzip wurden in der Bioinformatik verschiedene Verfahren entwickelt, um auch den Vergleich mehrerer Sequenzen als Gruppe zu ermöglichen. Mauno Vihinen[294] stellt ein Verfahren für den Vergleich einer Basissequenz mit einer Reihe weiterer Sequenzen vor, bei dem zunächst jeweils eine Alinierung durchgeführt wird und die Matrizen für jedes Sequenzpaar nur dort Übereinstimmungen verzeichnen, wo auch das Umfeld der einzelnen Matches eine bestimmte Mindestübereinstimmung erreicht. Anschließend kann für jedes Zeichen der Basissequenz festgestellt werden, inwieweit in den verschiedenen Matrizen insgesamt eine bestimmte Mindestzahl von Übereinstimmungen festzustellen ist, was sich ebenfalls in Form eines Dotplots darstellen lässt. Daraus lässt sich dann zwar nicht entnehmen, wo die Übereinstimmungen in den Vergleichssequenzen zu finden sind – es werden ja nicht die eigentlichen Sequenzen, sondern die Resultate der Alinierung zugrunde gelegt –, aber es ist zu erkennen, welche Bereiche der Basissequenz in einem hohen Maß mit den übrigen Sequenzen übereinstimmen.

Martin Vingron und Patrick Argos[295] zielen mit ihrem Verfahren wohl primär auf die Erkennung von übereinstimmenden kurzen Segmenten in nur entfernt

[293] Schematisierte Beispiele für Muster, wie sie im Bereich der Bioinformatik von Interesse sind, bietet Schulz/Leese/Held 2008 in Abbildung 6.
[294] Vihinen 1988.
[295] Vingron/Argos 1991.

verwandten Sequenzgruppen. Sie weisen auf die Schwierigkeiten hin, die mit der Alinierung verbunden sind, da sie vor allem bei nur schwacher Übereinstimmung stark von den gewählten Parametern, insbesondere der Bewertung von Unterbrechungen, abhängt. Stattdessen gehen sie davon aus, dass sich kurze Muster erkennen lassen, wenn sie in einer Vielzahl von Sequenzen vorhanden sind. Um die entsprechenden Stellen zu ermitteln, lassen sie die festgestellten Übereinstimmungen in mehreren Stufen über eine jeweils paarweise erfolgende Zusammenführung zweier Dotplotmatrizen auf der Basis einer Matrixmultiplikation filtern.

Claudine Landès, Alain Hénaut und Jean-Loup Risler[296] nennen als Ziel eines Gruppenvergleichs die Ermittlung evolutionärer Verwandtschaftsverhältnisse zwischen Proteingruppen oder auch einfach eine Klassifikation auf der Basis ihrer Ähnlichkeit. Sie erwähnen ebenfalls die Problematik einer Alinierung von Sequenzen mit einem geringen Übereinstimmungsgrad oder großen Abständen zwischen den einander zuzuordnenden Bereichen. Für den Vergleich mehrerer Sequenzen projizieren sie nach der Erstellung der (das jeweilige Umfeld der einzelnen Positionen berücksichtigenden) Dotplots zu allen Sequenzpaaren die in den einzelnen Dotplots verzeichneten Übereinstimmungsgrade, soweit eine bestimmte Mindestlänge erreicht wird, jeweils auf eine Achse. Aus einer solchen Projektion lassen sich zwar nicht die konkreten Matches rekonstruieren, es ist aber zu erkennen, für welche Bereiche der durch die Projektionsachse repräsentierten Sequenz sich in der anderen Sequenz Übereinstimmungen im entsprechenden Ausmaß finden lassen (wobei bei der Überlagerung mehrerer Projektionslinien jeweils die höheren Werte übernommen werden). Die Projektionen, die einer bestimmten Sequenz zuzuordnen sind, werden jeweils in einen Vektor überführt und die Vektoren der verschiedenen Sequenzen dann nach Erstellung einer Distanzmatrix über eine Clusteranalyse gruppiert.

2.3 Sprachstatistik und Inhaltsanalyse

In diesem Kapitel sollen einige Erkenntnisse und Techniken aus dem Bereich von Computerlinguistik und *Information Retrieval* vorgestellt werden, die bei der Untersuchung der Ähnlichkeit von Texten eine Rolle spielen oder spielen können.

2.3.1 Zipfsches Gesetz, Stoppwörter und Termgewichtung

Wörter werden mit sehr stark unterschiedlicher Frequenz verwendet – nach der von George Kingsley Zipf aufgestellten Formel verhält sich die Häufigkeit (in etwa) umgekehrt proportional zum Rang bei einer Sortierung nach der Häufigkeit.[297] Ein

[296] Landès/Hénaut/Risler 1993.
[297] Stock 2007, S. 320.

recht großer Teil der laufenden Wortformen eines einsprachigen Textkorpus lässt sich also wenigen unterschiedlichen Wörtern beziehungsweise Wortformen zuordnen, während die meisten unterschiedlichen Formen auch in einem umfangreichen Textkorpus selten oder sogar nur einmal belegt sind.

Unter den am häufigsten vorkommenden Wörtern bilden solche, die weniger als eigenständige semantische Einheiten als durch ihre Funktion im Satzzusammenhang zu beschreiben sind, einen hohen Anteil. So sind nach Auswertungen von 1898 und von 2001 unter den 38 beziehungsweise 62 häufigsten Wortformen des Deutschen bis auf Hilfsverben weder Verben noch Substantive oder Adjektive vertreten; die ersten Ränge nehmen dabei die Formen *der*, *die* und *und* ein.[298] Im *Information Retrieval* ist es üblich, solche Funktionswörter – oder jedenfalls die häufiger belegten – als *Stoppwörter* zu bezeichnen und sie bei der Textindizierung und der Recherche nicht zu berücksichtigen.

Im hier betrachteten Zusammenhang spielen sie wohl aus zwei Gründen eine Rolle. Zum einen reduziert sich durch die Eliminierung von Stoppwörtern die Textmenge und damit der für einen Vergleich erforderliche Aufwand ganz erheblich, ohne dass die für den Inhalt und die Ähnlichkeit der Formulierung entscheidenden Wörter verloren gehen – und wie sich immer wieder zeigt, ist der Bedarf an Zeit und Speicherplatz für Textvergleiche schon bei eher kleinen Textmengen so groß, dass das durchaus von praktischem Interesse ist. Zum anderen – und inhaltlich natürlich gewichtiger – tritt bei einem Textvergleich nach einer Stoppwortentfernung stärker zutage, welche Textpaare einen hohen Anteil an übereinstimmenden inhaltstragenden Wörtern haben. Auch wenn man bei einem Vergleich vollständiger Texte berücksichtigt, dass ein Großteil der Übereinstimmungen auf Stoppwörter zurückzuführen ist, kann man nicht einfach einen festen Wert von einem ermittelten Ähnlichkeitsmaß abziehen, da gerade auch der Gebrauch von Funktionswörtern in Texten einer Sprache nicht einheitlich ist, sondern zum Beispiel als Kriterium für die Autorenidentifikation verwendet werden kann.[299] Grundannahme bei einer Stoppwortentfernung ist freilich, dass es beim Vergleich weniger um übereinstimmende Formulierungen als vielmehr um eine gleiche oder ähnliche Wortwahl geht – oder im Falle einer Synonymenersetzung noch nicht einmal um Wortübereinstimmungen, sondern um gleiche Themen.

Bei der Erstellung einer Stoppwortliste kann man sich zunächst einmal an den am häufigsten belegten Formen orientieren und diese einer manuellen Filterung

[298] Kaeding (Hg.) 1898, S. 53 f. bietet die absoluten Vorkommenshäufigkeiten der im untersuchten Korpus mit insgesamt knapp 11 Millionen laufenden Wortformen mindestens 5000 mal belegten Wörter und Kompositabestandteile sowie Angaben zu den Anteilen, die die drei, vier, fünfzehn, 66 und 320 häufigsten dieser Formen bilden. Unter *http://wortschatz.uni-leipzig.de/Papers/top100de.txt* war früher eine nach Häufigkeit sortierte Liste der Auswertung von 2001 zu finden. Vgl. auch die seit 2007 publizierten DeReWo-Grund-/Wortformenlisten (*http://www.ids-mannheim.de/derewo*) des *Instituts für Deutsche Sprache*.

[299] Vgl. etwa Argamon/Levitan 2005 oder Swinson/Reyna 2013.

unterziehen, um inhaltstragende Wörter zu entfernen.[300] Der Begriff *Stoppwort* impliziert allerdings nicht notwendigerweise, dass es sich um ein Funktionswort handelt, sondern bringt zum Ausdruck, dass das betreffende Wort bei einer Recherche (oder im hier untersuchten Zusammenhang bei einem Vergleich) nicht berücksichtigt werden soll. So können zum Beispiel in Fachdatenbanken auf das jeweilige Fach bezogene Stoppwortlisten verwendet werden.[301]

Eine Stoppwortliste funktioniert nach einem binären Prinzip – ein Wort beziehungsweise eine Wortform wird entweder berücksichtigt oder nicht. Eine differenziertere Verarbeitung – anstelle oder in Ergänzung zu einer solchen Liste – ist über eine Termgewichtung möglich. Für die Ermittlung von Textähnlichkeiten ist wohl insbesondere das *tf-idf*-Maß von Interesse. Dabei wird jedem unterschiedlichen Term (also – wenn die Texte nicht weiter aufbereitet werden – jeder unterschiedlichen Wortform) in jedem Text ein Wert zugeordnet, der sich aus zwei Faktoren (im mathematischen Sinne) ergibt: zum einen aus einem Wert, der auf der Vorkommenshäufigkeit im Text basiert (*tf = term frequency* – es ist etabliert, aber nicht zwingend, dabei den Logarithmus zu verwenden), zum anderen aus dem Logarithmus des Verhältnisses der Gesamtzahl der untersuchten Texte zur Anzahl derer, in denen er vertreten ist (*idf = inverse document frequency* – der Wert wird als invers bezeichnet, da nicht der Anteil der betreffenden Dokumente am Gesamtkorpus zugrunde gelegt wird, sondern der Kehrwert dazu, so dass er umso höher ist, je weniger Texte den betreffenden Term enthalten).[302]

2.3.2 Lemmatisierung, *Stemming* und Synonymenersetzung

Im vorangehenden Unterkapitel ist recht unbestimmt von Wortformen, Wörtern und Termen die Rede. Hintergrund ist, dass die beschriebenen Verhältnisse und Möglichkeiten von den Grundregeln her nicht nur dann gelten, wenn die unveränderten Wortformen der Texte zugrunde gelegt werden, sondern auch, wenn durch bestimmte Verarbeitungsschritte eine Zusammenfassung von einander sprachlich oder in der Bedeutung nahestehenden Formen erfolgt.[303] Es liegt der Sache

[300] Laut Stock 2007, S. 223 wurde dieser Ansatz 1989 von Christopher Fox vorgeschlagen. Vgl. ebd. S. 222–225 mit weiteren Erläuterungen zu allgemeinen Stoppwortlisten.

[301] Vgl. ebd. S. 225 f.

[302] Diese Beschreibung des *idf*-Werts entspricht der von Stephen Robertson vorgestellten Variante, die den Effekt hat, dass Terme, die in allen Texten vorkommen, als *tf-idf*-Wert 0 erhalten. Vgl. ebd. S. 321–326, wo auch weitere Berechnungsmethoden vorgestellt werden.

[303] So Stock 2007, S. 321 über die Anwendung der Verfahren zur Termgewichtung. Dass das Zipfsche Gesetz nicht von der Zusammenfassung beziehungsweise Unterscheidung verschiedener Flexionsformen abhängt, zeigt sich in seiner Anwendbarkeit auf Sprachen mit unterschiedlich großem Formenreichtum, zum Beispiel auf das Englische und das Deutsche. Allerdings führt eine Zusammenfassung von Formen natürlich zu einem anderen Faktor, der für die Abschätzung der Häufigkeit einer Form anhand ihres Ranges zu verwenden ist, beziehungsweise zu einem entsprechend geänderten Kurvenverlauf.

nach insbesondere nahe, die Wortformen auf die Stichwörter zurückzuführen, wie sie in Wörterbüchern verzeichnet werden, sie also zu lemmatisieren (oder jedenfalls – unter Verzicht auf die Unterscheidung von Homonymen, die zum Teil vom jeweiligen lexikographischen Konzept abhängt und sprachwissenschaftlich fundierte Interpretation erfordert – die Grundformen zu ermitteln). Für computerlinguistisch hinreichend erschlossene Sprachen wie das heutige Deutsch kann man die sich bei einer automatischen Lemmatisierung stellenden Probleme zwar als weitgehend gelöst betrachten – jedenfalls wenn man bereit ist, eine gewisse Fehlerquote in Kauf zu nehmen –, ein sachgerechtes Verfahren ist jedoch nicht trivial und erfordert nicht nur ein umfassendes maschinenlesbares Wörterbuch mit Informationen zur Flexionsmorphologie, sondern auch eine syntaktische Analyse, da sich nicht selten die Zuordnung erst daraus ergibt.[304]

Anstelle einer Lemmatisierung kann auch – und leichter – ein *Stemming* durchgeführt werden, das heißt (nach einem verbreiteten Verständnis) eine Entfernung von Suffixen, so dass nicht nur Homonyme nicht weiter differenziert werden, sondern auch verschiedenen Wortarten zuzuordnende Formen zusammenfallen. Beim *Stemming* werden in der Regel keine Präfixe entfernt, da diese für die Bedeutung entscheidend sein können.[305]

Auf die Erkennung inhaltlicher Übereinstimmung auch bei unterschiedlicher Wortwahl zielen Verfahren, die Synonyme zusammenfassen.[306] Eine entsprechende automatische Umformung setzt voraus, dass diese Bedeutungsrelationen verzeichnet sind – für das Englische kann dafür auf *WordNet* zugegriffen werden, für das Deutsche auf *GermaNet*.[307] Bei vielen Wörtern ist allerdings zu berücksichtigen,

[304] Als ein Beispiel sei auf Partikelverben verwiesen, also zusammengesetzte Verben, bei denen das Anfangsglied der Grundform zum Beispiel in Hauptsatzkonstruktionen abgetrennt am Ende (gegebenenfalls nach syntaktisch zugehörigen Wörtern als schließendes Glied einer Verbalklammer) steht (vgl. Bussmann 1990 S. 562 s. v. „Partikelverb“ und S. 662 f. s. v. „Satzklammer“). Eine – im Sinne einer bestimmten Orthographie – korrekte Lemmatisierung erfordert zudem nicht nur, dass zum Beispiel ein *zu* oder *ab* am Ende eines Hauptsatzes entsprechend einer solchen Analyse als Verbzusatz erkannt wird, sondern auch die Entscheidung, ob eine entsprechende Konstruktion zum Beispiel mit „da“ oder „zusammen“ als Bestandteil des Verbs zu betrachten ist, was zumindest in manchen Fällen von der Bedeutung abhängt und deshalb nicht anhand der in einem Wörterbuch verzeichneten Lemmata entschieden werden kann. Vgl. Duden 9[7], S. 213 s. v. „da“ und S. 1057 s. v. „zusammen“.

[305] Vgl. Stock 2007, S. 232 f. Ebd. S. 233–239 werden neben verschiedenen allgemeinen *Stemming*-Verfahren auch solche beschrieben, die Kookkurrenzen im zugrunde gelegten Korpus berücksichtigen oder die auf einer N-Gramm-Zerlegung (vgl. dazu unten Unterkapitel 2.3.3) basieren.

[306] Im *Information Retrieval* können auch weitere Bedeutungsbeziehungen wie zum Beispiel Hyper- und Hyponymie berücksichtigt werden, vgl. Stock 2007, vor allem S. 285–287. Für die Ermittlung von nicht ganz wörtlichen Textübernahmen dürfte das eine geringere Rolle spielen, da in der Regel – jedenfalls in Sachtexten – die Wortwahl ja durch den jeweiligen Sachzusammenhang und die intendierte Satzbedeutung begründet ist, die sich bei der Ersetzung durch nicht synonyme Formulierungen ändert.

[307] Die Homepage von *WordNet* ist *http://wordnet.princeton.edu/*, die von *GermaNet http://www.sfs.uni-tuebingen.de/lsd/*. Vgl. Stock 2007, S. 271–283 sowie zu *GermaNet* Kunze/Lemnitzer 2007, S. 135–139.

dass sie polysem sind, also mehrere Bedeutungen haben. Wenn keine Anhaltspunkte vorhanden sind, anhand derer die intendierte Bedeutung erschlossen werden kann,[308] kann es sinnvoll sein, alle möglichen oder wahrscheinlichen Bedeutungen zu berücksichtigen.[309] Für den Vergleich von Texten bedeutet das allerdings eine erhebliche Erhöhung der Komplexität, nämlich entsprechend dem Produkt aus der Zahl der für jedes berücksichtigte Wort einbezogenen Bedeutungen.

Für die hier untersuchten Texte können die in diesem Unterkapitel vorgestellten Verfahren nicht oder allenfalls mit erheblichen Einschränkungen genutzt werden. Insbesondere eine Synonymenersetzung scheidet schon deshalb aus, weil es kein abgeschlossenes einigermaßen umfassendes Wörterbuch des Frühneuhochdeutschen gibt, das als Grundlage verwendet werden könnte.[310] Eher vorstellbar erscheint eine Lemmatisierung oder ein *Stemming*. Bryan Jurish hat für die Abbildung von aus verschiedenen Epochen stammenden historischen Schreibungen auf neuhochdeutsche Entsprechungen einen nichtdeterministischen „rewrite transducer“ mit 306 gewichteten Ersetzungsregeln entwickelt.[311] Die dafür angegebenen hohen Werte für *Precision* und *Recall*[312] lassen sich so aber wohl für ein frühneuhochdeutsches Korpus nicht reproduzieren, da frühneuhochdeutsche Texte – wie oben in Kapitel 1.3 beschrieben – in einem im Laufe der Zeit zwar abnehmenden, aber insgesamt doch recht erheblichen Maße von dialektalen Einflüssen und je nach Schreibsprache variierenden orthographischen Konventionen bestimmt sind.[313]

[308] Vgl. Stock 2007, S. 288 f.

[309] So (in etwa) das Verfahren in Nawab/Stevenson/Clough 2012.

[310] Das teilweise vorliegende Frühneuhochdeutsche Wörterbuch (FWB) bietet mit seiner ausführlichen Verzeichnung von bedeutungsverwandten Wörtern (insbesondere von solchen, die aus dem jeweiligen Textumfeld ermittelt werden können) zwar einerseits einen Datenbestand, der bei entsprechender elektronischer Aufbereitung in dieser Form genutzt werden könnte, zeigt aber auch die Schwierigkeiten, die sich bei der Anwendung des von klar abgrenzbaren Begriffen her gedachten Konzepts der Synonymenersetzung auf Wörter beziehungsweise Wortbedeutungen ergeben.

[311] Vgl. Jurish 2011, vor allem S. 42–47. Die Regeln werden in etwas modifizierter Form (und geringfügig verminderter Zahl) ebd. S. 90–101 vorgestellt. Ähnliche Ansätze, die ebenfalls Ersetzungsregeln anwenden, aber umgekehrt eine Transformation von Wörtern in heutiger Orthographie in mögliche historische Schreibungen leisten sollen (vgl. Ernst-Gerlach 2013), sollen hier nicht weiter betrachtet werden, da es dabei gerade nicht um eine 1:1-Abbildung geht, die als Basis für einen schnellen Textvergleich dienen könnte, sondern vielmehr um die Generierung verschiedener Varianten als Erweiterung einer Suchanfrage.

[312] Jurish 2011, S. 46.

[313] Ein Indiz dafür bieten schon die in Jurish 2011, S. 46 angegebenen Prozentwerte für eine Zuordnung von historischen Wortformen auf gleich geschriebene neuhochdeutsche Formen (99,9 % *Precision* und 70,8 % *Recall* für die *Types*, also die unterschiedlichen Wortformen, sowie 99,1 % *Precision* und 83,7 % *Recall* für die *Tokens*, die laufenden Wortformen). In den oben in Kapitel 1.3 wiedergegebenen – natürlich statistisch in keiner Weise repräsentativen – Zitaten wäre für eine solche Zuordnung, wenn man sie auch bei einer Abweichung hinsichtlich der Groß- und Kleinschreibung vornimmt, nach meiner Zählung der *Recall* in Bezug auf die laufenden Wortformen deutlich unter 50 % (wobei ca. 2/3 der Entsprechungen Stoppwörter betreffen) und die *Precision* – je nachdem, wie die Zuordnung genau erfolgt und was man als falsch bewertet –

2.3.3 Textsegmentierung und N-Gramme

Für die Ermittlung sprachlich oder inhaltlich ähnlicher Textpassagen reicht es in aller Regel nicht aus, den Wortbestand (oder auch den Wörtern zugeordnete Synonyme) kompletter Texte miteinander zu vergleichen, da der Übereinstimmungsgrad nur bei einer recht umfassenden Übernahme signifikant sein dürfte. Ähnlich wie im *Information Retrieval* eine Unterteilung der ausgewerteten Texte zum Beispiel in Sätze oder Absätze erfolgen kann, um die zu einer Anfrage am besten passenden Textstellen zu ermitteln,[314] legt sich eine Segmentierung auch für den Textvergleich nahe, um dann zwischen den gebildeten Abschnitten nach auffälligen Entsprechungen zu suchen. Durch eine Abgrenzung relativ enger Kontexte sinkt die Wahrscheinlichkeit zufälliger Wortübereinstimmungen (abgesehen von Stoppwörtern), und die Bedeutung einzelner Wörter lässt sich wesentlich besser erschließen, selbst wenn nicht die genaue Formulierung verzeichnet wird, sondern nur eine ungeordnete Liste aller enthaltenen Wortformen oder Wörter. Auf dieser Basis lassen sich zum Beispiel Schnittmengen der enthaltenen Wortformen bilden, wodurch sich Ähnlichkeiten auch bei Umstellungen innerhalb der jeweils betrachteten Passagen finden lassen. Voraussetzung für einen hohen Übereinstimmungsgrad ist freilich, dass dabei die Segmentgrenzen in einander entsprechenden Textstücken nicht an unterschiedlichen Stellen gesetzt werden.

Eine Alternative zu einer Gliederung in zumindest im Idealfall inhaltlich zusammengehörige Passagen besteht in einer rein formalen Bildung von Textausschnitten. Aus dem Bereich der Sprachstatistik und der Informationstheorie stammt die Zerlegung in sogenannte *N-Gramme*.[315] Dabei werden im zugrunde liegenden Textmaterial alle enthaltenen nicht unterbrochenen Folgen von jeweils genau n Elementen ermittelt, wobei n eine bestimmte Zahl ist und ein Element die jeweils zugrunde gelegte Analyseeinheit – insbesondere kann es sich bei den Elementen um Zeichen (vor allem Buchstaben) oder Wörter handeln.[316] Die Zahl der in einem zugrunde gelegten Text beziehungsweise Textstück enthaltenen N-Gramme

zwischen 94,5 und 96,8 %. Die in Jurish 2011, S. 90–101 beschriebenen Umwandlungsregeln berücksichtigen zwar eine ganze Reihe der Abweichungen gegenüber der neuhochdeutschen Standardsprache, aber durchaus nicht alle. So ist zwar als Möglichkeit vorgesehen, ein *v* durch ein *u* oder ein *u* durch ein *au* zu ersetzen, nicht aber ein *v* durch ein *au* oder ein *u* durch ein *f*. Und auch bei einer entsprechenden Erweiterung des Regelsatzes bliebe wohl das Grundproblem, dass die Regeln in diesem Ansatz nicht textspezifisch, sondern für das gesamte Korpus formuliert sind, die Wahrscheinlichkeit einer sachlich korrekten Anwendung in vielen Fällen aber vom jeweiligen Text beziehungsweise der zugrunde liegenden Schreibsprache abhängt.

[314] Vgl. Stock 2007, S. 498.

[315] Die Bezeichnung beziehungsweise ihr englisches Äquivalent *n-gram* wurde wohl von Claude E. Shannon geprägt, vgl. Stock 2007, S. 201 sowie Shannon 1948, S. 387. Dass die statistische Untersuchung von Einzelzeichen sowie Zeichen-Bi- und -Trigrammen älter ist, lässt sich daraus entnehmen, dass Shannon ebd. auf ein Werk von 1939 hinweist.

[316] Weitere Möglichkeiten sind zum Beispiel Phoneme oder Silben, vgl. *http://en.wikipedia.org/wiki/N-gram*.

(als *Tokens*) entspricht ohne Modifikation fast der Zahl der enthaltenen Elemente – beim ersten Element beginnt das erste N-Gramm, beim zweiten das zweite usw. bis auf die letzten Elemente, die weniger als $n - 1$ Elemente vom Ende entfernt sind. Damit die Elemente am Anfang und am Ende eines zugrunde gelegten Textes (beziehungsweise Textstückes) wie die übrigen in jeweils n N-Grammen vertreten sind, kann man aber auch an den Rändern jeweils $n - 1$ Platzhalter-Elemente (zum Beispiel Leerzeichen) ergänzen, so dass sich die Zahl der N-Gramme entsprechend erhöht.[317]

Kurze Zeichen-N-Gramme sind insbesondere sprachstatistisch aussagekräftig, da sich die häufigsten Zeichenkombinationen zwischen den Sprachen stark unterscheiden. Je höher das gewählte n ist, umso mehr unterscheiden sich die gebildeten N-Gramme; sie können deshalb bei einem gut gewählten n ähnlich wie Wort-N-Gramme für die Ermittlung von übereinstimmenden Formulierungen genutzt werden, wobei eine N-Gramm-Bildung entsprechend den Wortgrenzen wohl in aller Regel sachlich sinnvoller und auch üblicher ist.[318]

Wenn zwei Texte auffällig starke Übereinstimmungen in Einzelwörtern (insbesondere nach der Eliminierung von Stoppwörtern) haben, kann dies ein Anzeichen für eine ähnliche Thematik sein. Zusammenhängende Wortfolgen sind hingegen nur zu einem kleinen Teil (etwa in Redewendungen) sprachlich vorgegeben oder naheliegend. Da die Zahl der (ohne Berücksichtigung grammatischer und natürlich erst recht inhaltlicher Kriterien) theoretisch möglichen Wortkombinationen entsprechend dem gewählten n exponentiell wächst, ist die Wahrscheinlichkeit rein zufälliger Übereinstimmungen schon bei einem kleinen n recht gering, soweit die betreffenden Wortfolgen nicht sprachlich und gegebenenfalls auch sachlich naheliegen.[319] Schon Wort-Trigramme sind so spezifisch, dass der Anteil der Entsprechungen bei eigenständigen Texten zu verwandten Themen klein ist, selbst wenn sie vom selben Autor stammen.[320] Wenn ein Textvergleich auf N-Gramm-

[317] Vgl. Stock 2007, S. 201 f.

[318] Ein Vorteil der Verwendung von Zeichen-N-Grammen mag darin bestehen, dass längeren Wörtern automatisch ein größeres Gewicht beigemessen wird. Wenn eine Übereinstimmung nicht mit Wortgrenzen zusammenfällt, ist das wohl in den meisten Fällen ein Grund, sie als nicht oder weniger relevant zu betrachten, was für die Verwendung von Wort-N-Grammen spricht. Allerdings ist dieses Argument fragwürdig, wenn – wie in den hier untersuchten Texten – die Wortabgrenzung Schwankungen unterliegt. Vgl. zum praktischen Einsatz von Vergleichen auf N-Gramm-Basis die Kapitel 2.4 und 2.5.

[319] Aufgrund syntaktischer Muster ist zum Beispiel eine Drei-Wort-Folge, die aus einem Artikel, danach einem entsprechend flektierten Adjektiv und schließlich einem Substantiv besteht, wesentlich wahrscheinlicher als ein Wort-Trigramm mit der Reihenfolge Substantiv – Adjektiv – Artikel. Bei Nicht-Funktionswörtern ist die Häufigkeit einer Kombination stark von der jeweiligen Bedeutung und zum Teil auch vom typischen Wortgebrauch abhängig. Und wenn ein N-Gramm ausschließlich oder überwiegend aus hochfrequenten Stoppwörtern besteht, die sich sinnvoll miteinander kombinieren lassen, ist auch die Wahrscheinlichkeit für das N-Gramm recht hoch.

[320] In Lyon/Barrett/Malcolm 2004, S. 2 wird festgestellt, unabhängig voneinander geschriebene Texte mit einer Länge von 1000–5000 Wörtern hätten maximal 8 % übereinstimmende Trigramme.

Basis erfolgt, bedeutet das allerdings, dass Umstellungen oder Einschübe die Erkennung einer Entsprechung verhindern können, insbesondere, wenn für *n* ein relativ großer Wert gewählt wird.

2.3.4 Codierung nach phonetischen Kriterien

Für die Arbeit mit frühneuhochdeutschen Texten bieten Verfahren, die eine normierte Orthographie voraussetzen, nur begrenzt Unterstützung. Ein Ansatz, der es erleichtert, Schreibungsvarianten einander zuzuordnen, stammt aus der Erschließung von Personennamen. Bekanntlich gibt es insbesondere bei Familiennamen oft eine Vielzahl von ähnlichen Formen (als Beispiel sei nur *Meier/Meyer/Maier/Mayer* genannt), und lange Zeit war die Schreibweise auch für eine bestimmte Person oder Familie nicht so festgelegt, wie sie es heute ist.[321] Das in dieser Arbeit in Teil 3 vorgestellte Verfahren knüpft an verschiedene phonetische Algorithmen[322] an, deren typische Verwendung insbesondere in der Zusammenfassung gleich oder ähnlich lautender, aber unterschiedlich geschriebener Familiennamen unter gemeinsame Suchschlüssel liegt.[323]

Als wohl ältester und zugleich auch bekanntester dieser Algorithmen ist *Soundex* zu nennen, als dessen Urform eine erstmals 1918 patentierte Indizierungsmethode von Robert C. Russell gelten kann.[324] Nach dem *Soundex*-Algorithmus wird ein Wort auf vier Zeichen abgebildet. Der Anfangsbuchstabe wird unverändert (als Großbuchstabe) übernommen, dann folgen drei Ziffern, denen nach dem in Tabelle 2.3[325] angegebenen Schema Buchstaben zugeordnet sind. Mehrere aufeinander folgende Buchstaben mit dem gleichen Code werden durch eine einzige Ziffer repräsentiert. Vokale spielen (abgesehen vom Anfangsbuchstaben) nur insofern eine Rolle, als sie diese Zusammenfassung gleicher Codes verhindern, wenn sie zwischen den betreffenden Buchstaben stehen. Auch *h*, *w* und *y* werden nicht weiter

[321] Ein bekanntes Beispiel ist, dass Goethes Name lange Zeit auch „Göthe“ geschrieben wurde (so auch als Variante in der Gemeinsamen Normdatendatei unter *http://d-nb.info/gnd/118540238/about/html* zu finden).

[322] Vgl. *http://en.wikipedia.org/wiki/Phonetic_algorithm*.

[323] Vgl. zu diesem Thema Wilz 2005. Ebd. S. 5 werden Anwendungsbeispiele aufgeführt.

[324] Die Angaben in der Literatur sind häufig im Hinblick auf die historische Entwicklung ungenau. Der *Soundex*-Algorithmus wird zwar wohl fast durchgängig so wie hier auch beschrieben (so zum Beispiel im von Wilz 2005 angeführten Standardwerk von Donald E. Knuth von 1973 – in der 2. Auflage Knuth 1998, S. 394 f. – und im dort angeführten Aufsatz Bourne/Ford 1961, S. 544 f.), die angegebenen Transformationsregeln weisen aber einige deutliche Unterschiede zu dem Verfahren auf, das in Russells Patentanmeldung dokumentiert ist (Russell 1918 und weitgehend gleich Russell 1922; eine Beschreibung von Russells Verfahren findet sich in Stock 2007, S. 307–309). Die Herkunft der im Patent nicht verwendeten Bezeichnung *Soundex* und die Entwicklung der Modifikationen soll hier nicht weiter untersucht werden. Einen etwas ausführlicheren Überblick zur Geschichte von *Soundex* einschließlich der Auflistung einiger weiterer Patentanmeldungen von Russell bietet *http://creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm#History*.

[325] Ähnlich zum Beispiel in Wilz 2005, S. 13. Eine entsprechende Tabelle (allerdings mit anderen Zuordnungen) gibt es schon in Russell 1918 und Russell 1922.

Code	Buchstaben
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

Tab. 2.3: Buchstabencodierung nach *Soundex*

berücksichtigt, ebenso wenig alle Buchstaben am Wortende, sobald die codierte Version drei Ziffern umfasst. Führt das Codierungsschema zu weniger als drei Ziffern, wird am Ende mit Nullen aufgefüllt.[326]

Der *Soundex*-Algorithmus ist recht verbreitet. Er weist allerdings verschiedene Schwächen auf, die Anlass gegeben haben zur Entwicklung ähnlicher Verfahren.[327] Die 1969 von Hans Joachim Postel publizierte *Kölner Phonetik* berücksichtigt deutsche Schreibregeln und differenziert bei einigen Buchstaben die Codierung je nach Zeichenkontext. Eine ungefähre Übersicht über die dabei angewandten Codierungsregeln bietet Tabelle 2.4 auf S. 82.[328] Auch hier werden gleiche Codes zusammengefasst, wenn die entsprechenden Buchstaben aufeinander folgen, und auch hier werden Vokale in der codierten Form nur am Wortanfang sowie als Trenner, die die Zusammenfassung gleicher Codes verhindern, berücksichtigt. Anders als bei *Soundex* werden auch längere Wörter komplett codiert.[329]

Als drittes Beispiel soll noch *Metaphone* angeführt werden, das wie *Soundex* ursprünglich für die englische Sprache entwickelt wurde. Lawrence Philips publizierte die erste Fassung 1990. In einer verbesserten Version, die er 2000 unter dem Namen *Double Metaphone* veröffentlichte, sollen auch die Schreibregeln verschiedener anderer Sprachen berücksichtigt sein. *Metaphone 3* aus dem Jahr 2009 ist nur als kommerzielles Produkt erhältlich. *Metaphones* Codierungsregeln sind in vielen Fällen kontextabhängig und insbesondere ab *Double Metaphone* erheblich komplexer als die bisher vorgestellten. Anders als bei *Soundex* und der *Kölner Phonetik*

[326] Vgl. Wilz 2005, S. 12 f.

[327] Ebd. S. 13–21 werden *Extended Soundex*, *Metaphone*, *Phonix*, der Daitch-Mokotoff-Algorithmus, die *Kölner Phonetik*, das von Wilz als *Phonem* bezeichnete Codierungsverfahren sowie *Phonet* kurz vorgestellt

[328] Die Tabelle basiert zum einen auf Wilz 2005, S. 18, zum anderen auf *https://de.wikipedia.org/wiki/K%C3%B6lner_Phonetik* – die in den dort gezeigten Tabellen zu findenden Informationen stimmen nicht in allen Details völlig überein und wurden hier nach Plausibilität und unter Berücksichtigung der dort zusätzlich zu findenden Erläuterungen zusammengefasst. Laut Wilz 2005, S. 61 ist die Originalveröffentlichung zur *Kölner Phonetik* hinsichtlich der Regeln teilweise mehrdeutig. Dass die Umlaute und das *ß* nicht berücksichtigt sind, erklärt sich durch die damals übliche Form der digitalen Textrepräsentation, vgl. ebd. S. 17.

[329] Vgl. die Erläuterungen in *https://de.wikipedia.org/wiki/K%C3%B6lner_Phonetik* sowie die Implementierung in *Perl* in Wilz 2005, S. 61–64.

Code	Buchstaben
0	A, E, I, J, Y, O, U (nur am Wortanfang)
1	B; P außer vor H
2	D, T außer vor C/S/Z
3	F, V, W, PH
4	G, K, Q; C vor A/O/U/H/K/Q/X – aber nicht nach S/Z –, am Wortanfang auch vor L/R
48	X außer nach C/K/Q
5	L
6	M, N
7	R
8	S, Z; C nach S/Z, ansonsten außer vor A/O/U/H/K/Q/X, am Wortanfang auch nicht vor L/R; D, T vor C/S/Z; X nach C/K/Q

Tab. 2.4: Buchstabencodierung nach der *Kölner Phonetik*

werden als Codierungszeichen Buchstaben verwendet, deren typische Lautung zu der Lautklasse gehört, die dadurch repräsentiert wird. So steht zum Beispiel ein *T* für die Laute */d/* und */t/* und ein *F* für die Laute */f/* und */w/*. Wie in *Soundex* und in der *Kölner Phonetik* werden Vokale außer am Anfang nur als Kontext für die Nachbarbuchstaben berücksichtigt, haben aber in der codierten Fassung selbst keine Entsprechung.[330]

Auf die Beschreibung weiterer phonetischer Algorithmen soll hier verzichtet werden.[331] Die drei vorgestellten Verfahren lassen einige Gemeinsamkeiten erkennen, die auch der schon beschriebenen Varianz in der Schreibung frühneuhochdeutscher Texte weitgehend entsprechen:

- Konsonanten, denen eine ähnliche Lautung zugeordnet ist, werden durch einen gemeinsamen Code repräsentiert. Das betrifft insbesondere auch Buchstabenpaare, die für Fortis und Lenis (beziehungsweise stimmlose und stimmhafte Lautung) stehen. Nach dieser Regel können im Frühneuhochdeutschen Schreibvarianten, die durch die binnendeutsche Konsonantenschwächung[332] zu erklären sind, durch denselben Code wie die hochsprachliche Form repräsentiert werden.
- Wenn mehreren aufeinander folgenden Buchstaben (beziehungsweise Buchstabengruppen) der gleiche Code zugeordnet ist, werden sie zu einem einzigen Codezeichen zusammengefasst. Da Konsonanten im Frühneuhochdeutschen häu-

[330] Vgl. zu den verschiedenen Versionen von Metaphone und zu den Ersetzungsregeln von 1990 *http://en.wikipedia.org/wiki/Metaphone*.
[331] Einen Überblick bietet Wilz 2005, S. 13–22.
[332] Vgl. oben S. 44.

fig durch Doppelschreibungen oder eine Folge einander im Hinblick auf den Lautwert nahestehender Buchstaben wiedergegeben werden, können dadurch in vielen Fällen unterschiedliche Schreibmöglichkeiten zu einer einzigen Codeform vereinheitlicht werden.
- Vokale verhindern, dass vor und nach ihnen stehende Buchstaben mit gleicher oder ähnlicher Lautung zusammengefasst werden, werden aber ansonsten in der codierten Fassung nur am Wortanfang berücksichtigt. Das kann auch für bestimmte Konsonanten gelten. Auch dies lässt sich zum Teil recht gut auf frühneuhochdeutsche Texte anwenden, da bei ihnen nicht nur die Schreibung der Vokale variiert, sondern auch der jeweils intendierte Vokal in zahlreichen Fällen in einem gewissen Bereich wechseln und ein *<e>* in Nebentonsilben (der Schwa-Laut /ə/) gegenüber dem heutigen Deutsch fehlen oder hinzukommen kann.

Außerdem sind noch die folgenden Anregungen erwähnenswert:
- *Soundex* fasst nicht nur *<b>* und *<p>* zusammen, sondern stellt auch *<f>* und *<v>* in die gleiche Gruppe. Das könnte ein Verfahren sein, um Schreibvarianten aufgrund von Spirantisierung[333] zu berücksichtigen.
- In der *Kölner Phonetik* und in *Metaphone* werden bestimmte Buchstabenfolgen, deren Bestandteile verschiedenen Codes zugeordnet sind, als Einheit anders codiert, und die Codierung bestimmter Buchstaben hängt vom jeweiligen Buchstabenumfeld ab.[334] Auch im Frühneuhochdeutschen gibt es Fälle, bei denen das sinnvoll sein könnte, so die Zusammenfassung von *<sch>* und die differenzierte Behandlung von *<c>* je nach Folgebuchstabe.
- Die Verwendung von Buchstaben als Codes wie in *Metaphone*[335] ermöglicht zum einen bei Bedarf eine größere Differenzierung als die Verwendung von Ziffern und ist zum anderen für den menschlichen Leser leichter zu interpretieren.

Es gibt allerdings auch einige Punkte, die eine unveränderte Zugrundelegung eines der Codierungsschemata für die Untersuchung frühneuhochdeutscher Texte nicht angeraten sein lassen:
- Die Wortabgrenzung ist insbesondere bei Kompositen und Präfixbildungen nicht einheitlich und oft auch aus dem Schriftbild gar nicht sicher zu ermitteln. Wenn analog zu den vorgestellten Regeln Vokalbuchstaben je nach Stellung im Wort unterschiedlich codiert werden, ist deshalb davon auszugehen, dass manche Übereinstimmungen nicht gefunden werden.
- Für einige Buchstaben lässt sich keine allgemein gültige Regel über die Zuordnung zu einer Lautklasse aufstellen. Vor allem ist die Interpretation von *<i/j/y>* und *<u/v/w>* als Konsonant oder Vokal beziehungsweise Bestandteil eines Diphthongs

[333] Vgl. oben S. 44.
[334] Letzteres trifft auch auf die von Russell patentierte Vorstufe von *Soundex* zu (vgl. Russell 1918, S. 2, Z. 5–11), wurde bei der Weiterentwicklung des Verfahrens aber anscheinend fallen gelassen.
[335] Dies wird schon in Russells Patent von 1918 (ebd. S. 1, Z. 107 ff.) als Möglichkeit erwähnt.

häufig nicht sicher, wenn keine weiteren Kriterien einbezogen werden können. Wenn man das Buchstabenumfeld und Regeln über im Deutschen mögliche beziehungsweise wahrscheinliche Lautfolgen berücksichtigt, ist allerdings in zahlreichen dieser Fälle eine klare Kategorisierung möglich. Dennoch verbleiben recht viele, bei denen die Deutung die morphologische Analyse und/oder die Kenntnis der Schreibgewohnheiten im jeweiligen Text erfordern würde. Soweit die Codierungsregeln nur die jeweilige Zeichenfolge zugrunde legen, ist deshalb entweder in einem gewissen Maße mit der Zuordnung falscher Lautklassen zu rechnen oder die Lautklassen müssen so allgemein gewählt werden, dass sie die bestehenden Deutungsmöglichkeiten zusammenfassen.

- Das Zeicheninventar ist größer. Es umfasst nicht nur auch Satzzeichen, sondern insbesondere auch Diakritika und Kürzungszeichen, bei denen zu prüfen ist, inwieweit ihnen ein für die Codierung relevanter Lautwert zuzuordnen ist.

2.4 Plagiatserkennung

Textübernahmen im Sinne dieser Arbeit sind zwar allenfalls zum Teil als Plagiate zu betrachten, da dieser Begriff zumindest den Anspruch einer persönlichen Urheberschaft impliziert und damit etwa auf Gesetzestexte nicht anwendbar ist,[336] gleichwohl steht die Erkennung solcher Übernahmen natürlich der Plagiatserkennung nahe und damit einem Problem mit praktischer Bedeutung – insbesondere auch im Hinblick auf akademische Qualifikationsarbeiten –, das zu entsprechenden Forschungsaktivitäten geführt hat.

Wohl den besten Überblick über den dabei erreichten Stand[337] bieten die Dokumentationen zu den seit 2009 jährlich[338] ausgetragenen PAN-Wettbewerben und -Evaluierungen zu Plagiatserkennung, Autoridentifikation und ähnlichen Aufgaben.[339] Dem in der vorliegenden Arbeit untersuchten Problem entspricht dabei der

[336] Vgl. oben S. 24.

[337] Kommerzielle Softwareprodukte sollen hier nicht vorgestellt werden, da deren Erkennungsheuristiken zumindest im Detail nicht bekannt sind. Von der Forschungsgruppe um Debora Weber-Wulff wurden verschiedene Vergleichstests durchgeführt, die einen Überblick über den dabei bis 2013 erreichten Stand verschaffen können, vgl. *http://plagiat.htw-berlin.de/software/*.

[338] Oder, wenn man die ebenfalls mit der Bezeichnung *PAN* verknüpften, aber primär wohl nach der jeweils für die Aufgabenstellung gewählten Abkürzung bezeichneten Evaluierungen im Rahmen des *Forum for Information Retrieval Evaluation (FIRE)* mitzählt, zweimal jährlich. Sie hatten bisher die Ermittlung von *Text Reuse* über Sprachgrenzen hinweg sowie von *Source Code Reuse* zum Gegenstand und bleiben hier unberücksichtigt. Vgl. die Übersicht über die PAN-Reihe mit Links zu den Websites der PAN-Kampagne unter *https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/events/#webis-panEvaluationCampaign*.

[339] Vgl. Potthast u. a. 2009, 2010, 2011, 2012, 2013 und 2014 sowie Stamatatos u. a. 2015 und Rosso u. a. 2016. Im „evaluation lab“ von 2016 ging es nur in einer Teilaufgabe der „author diarization“ um Plagiatserkennung, und zwar nach textinternen Erkennungsmerkmalen (vgl. Rosso u. a. 2016, insbesondere S. [5]-[7]). *PAN* stand zumindest ursprünglich anscheinend als Abkürzung

Teilbereich der Plagiatserkennung über den Vergleich mit einem lokal vorliegenden großen Textkorpus.[340]

Allgemein wird dafür ein mehrschrittiges Verfahren angewendet: Am Beginn steht (nach einer Vorverarbeitung, die zum Beispiel Stoppwörter entfernt, die Wörter auf Wortstämme reduziert oder Synonyme ersetzt) die Prüfung, welche Texte des Korpus am ehesten als mögliche Vorlagen in Frage kommen. In einem zweiten Schritt wird ein Feinvergleich der Texte durchgeführt, und in einem dritten werden die gefundenen Übereinstimmungen nach bestimmten Kriterien gefiltert, so dass sich Bereiche ergeben, die als mögliches Plagiat beziehungsweise dessen Vorlage von einem Menschen zu prüfen sind.[341]

Für die Bewertung der Ergebnisse in den Wettbewerben wurde eine Formel entwickelt, die *Precision* und *Recall* sowie zusätzlich die sogenannte Granularität (*granularity*) berücksichtigt, die misst, inwieweit die plagiierten Stücke als Einheit oder in kleineren Teilstücken erkannt werden[342]. Die Ermittlung dieser Werte setzt freilich voraus, dass eindeutige Kriterien vorliegen, was als Plagiat zu bewerten ist und was nicht – das ist im Rahmen eines Wettbewerbs möglich, da hier eine Liste der zu ermittelnden Stellen existiert, nicht aber bei der Anwendung auf noch nicht untersuchte Texte.[343]

Unveränderte Plagiate können beim Vergleich mit einem größeren Korpus inzwischen offenbar problemlos erkannt werden,[344] und auch für die Ermittlung von mit gewissen Verfahren veränderten Plagiaten erreichen die besten Vergleichssysteme inzwischen hohe Werte zum Beispiel für *Precision* und *Recall*.[345]

---

für „Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection" – so auf der Website zum ersten PAN-Workshop 2007 (*http://www.uni-weimar.de/medien/webis/events/pan-07/pan07-web/*). Stamatatos u. a. 2015, S. [1] sowie Rosso u. a. 2016 verwenden allerdings die Langbeschreibung „Uncovering Plagiarism, Authorship, and Social Software Misuse" vor der Nennung des (dazu nicht ganz passenden) Kürzels. Diese Formulierung findet sich auch schon im Titel von Stein u. a. (Hg.) 2009.

[340] Andere Teilbereiche sind zum Beispiel die Erkennung anhand von textinternen Merkmalen und die Ermittlung von Vergleichskandidaten über Anfragen an eine Suchmaschine. Letzteres trat in späteren PAN-Wettbewerben an die Stelle des Vergleichs mit einem vollständig verfügbaren Korpus und stand in den Jahren 2012–2015 als Aufgabe neben dem Feinvergleich von möglichen Quell- und Zieldokumenten (vgl. Potthast u. a. 2012, 2013 und 2014 sowie Stamatatos u. a. 2015).

[341] Vgl. Potthast u. a. 2011, S. [6] und ähnlich schon Potthast u. a. 2009, S. 2. Vereinheitlichungen im Hinblick auf Groß- und Kleinschreibung, Zeichensetzung und Ähnliches werden in dieser Beschreibung nicht erwähnt, aber wohl als selbstverständlich vorausgesetzt. Explizit genannt werden sie zum Beispiel von Stamatatos 2009, S. 42 und Micol u. a. 2010, S. [5].

[342] Vgl. Potthast u. a. 2009, S. 5. Eine Weiterentwicklung, die zwischen verschiedenen Erkennungszielen differenziert, wird in Potthast u. a. 2014, S. [13]-[16] vorgestellt.

[343] Tatsächlich zeigte sich beim Wettbewerb 2009, dass das Korpus einige unbeabsichtigte Textübereinstimmungen enthielt (Potthast u. a. 2009, S. 4).

[344] Vgl. Potthast u. a. 2011, S. [8].

[345] Vgl. Potthast u. a. 2014, S. [21]-[28] (mit tabellarischer Zusammenfassung der in den letzten Jahren erreichten Werte).

Die Prüfung großer Korpora kann sehr zeitaufwendig sein. In den Beschreibungen einzelner Systeme für die Wettbewerbe 2009 und 2010 ist teilweise von mehr als eintägigen Laufzeiten die Rede.[346] 2011 wurde zwar teilweise ein sehr niedriger Zeitbedarf erreicht, vielfach waren aber mehrere Stunden erforderlich.[347]

Da die Zahl der erforderlichen Vergleichsoperationen zu stark steigen würde, wenn jedes zu prüfende Textstück (zum Beispiel jeder Absatz oder jeder Satz) mit jedem Stück der als Quellen in Frage kommenden Texte verglichen würde, ist es für die Verwendbarkeit eines Verfahrens auch bei sehr großen Textmengen wichtig, zunächst mit wenig Aufwand eine Vorauswahl der Vergleichskandidaten zu treffen, die einen möglichst hohen Anteil der tatsächlich relevanten Übereinstimmungen abdeckt.

Häufig werden zu diesem Zweck Textähnlichkeiten auf der Basis von N-Grammen[348] festgestellt. Im PAN-Wettbewerb 2009 machten dies von den sechs erfolgreichsten Teilnehmern vier, zugrunde gelegt wurden dabei 16-Gramme auf Zeichen- beziehungsweise 5-, 8- oder 1-Gramme auf Wortebene; für die Ähnlichkeitsbewertung waren verschiedene Maße im Einsatz.[349] Wie oben auf S. 79 dargestellt wurde, weisen schon die Wort-Trigramme zweier Texte nur eine geringe Schnittmenge auf, wenn die Texte unabhängig voneinander formuliert wurden, selbst wenn diese das gleiche Thema behandeln. Ein höherer Prozentsatz an Übereinstimmungen kann also ein Anhaltspunkt für eine nähere Überprüfung sein, aber auch bei nicht miteinander verwandten Texten ist (insbesondere bei einem ähnlichen Thema) mit einem gewissen Anteil an gleichen Drei-Wort-Folgen zu rechnen, die jeweils für sich genommen also wenig aussagekräftig sind.[350]

Wenn N-Gramme dazu eingesetzt werden sollen, verwandte Texte oder Textstellen zu ermitteln, müssen sie so lang sein, dass die Texte von den möglichen N-Grammen jeweils nur einen sehr kleinen Anteil enthalten und die Übereinstimmungen zwischen Texten ohne Abhängigkeitsverhältnisse gering sind. Das bedeutet aber, dass sich die Textmenge fast um den Faktor n (entsprechend der Länge der N-Gramme) erhöht, da die N-Gramme einander überlappen und nur wenige Kombinationen in einem zu vergleichenden Text (oder – wie auch im Folgenden – Textstück) mehr-

[346] Basile u. a. 2009, S. 21 u. 23 nennt für zwei Schritte ca. 40 beziehungsweise ca. 20 Stunden, Gottron 2010, S. [8] insgesamt über 38 Stunden. In beiden Wettbewerben waren Textdateien im Gesamtumfang von mehreren Gigabyte zu überprüfen.

[347] Potthast u. a. 2011, S. [6].

[348] Vgl. oben Unterkapitel 2.3.3.

[349] Potthast u. a. 2009, S. 6 bietet einen Überblick über die verwendeten Verfahren in Tabellenform.

[350] Das in Lyon/Barrett/Malcolm 2004 beschriebene System *Ferret* (s. u. S. 96) ist darauf ausgelegt, dass nach der Ermittlung der Häufigkeit von Trigramm-Übereinstimmungen bei Werten oberhalb eines gewissen Grenzwerts eine Überprüfung der Textparallelen vorgenommen wird. Während zufällig gleiche Wortfolgen in ihrem jeweiligen Textumfeld eher vereinzelt sind, häufen sich die Entsprechungen dort, wo sie auf Textübernahmen beruhen, und sind dadurch bei entsprechender Visualisierung schon optisch leicht zu erkennen (ebd. S. 3 f.).

fach vorkommen (und dann nur einmal verzeichnet werden, sofern es nicht um die Position, sondern nur um die Existenz eines N-Gramms geht).

Trotzdem erleichtert eine entsprechende Aufbereitung den Vergleich: Die ermittelten N-Gramme der einzelnen Texte lassen sich jeweils so verzeichnen, dass sich vergleichsweise schnell ermitteln lässt, welche beziehungsweise wie viele N-Gramme in zwei zu vergleichenden Texten übereinstimmen. Ein einfaches (und zudem leicht serialisierbares) Verfahren wäre etwa die alphabetische Sortierung der N-Gramm-Listen für jeden Text. Dann könnten anschließend für jedes zu untersuchende Textpaar die Listen parallel durchgeschaut werden (in der zweiten Liste würde also jeweils so weit weitergelesen, bis der zur Alphabetposition in der ersten Liste passende Eintrag erreicht oder bei Nichtexistenz überschritten wäre).

Wenn jeder Text mit jedem anderen verglichen wird, steht die Zahl der Textpaare allerdings in einem fast quadratischen Abhängigkeitsverhältnis zur Zahl der Texte.[351] Der Aufwand kann reduziert werden, wenn die N-Gramme über einen invertierten Index verzeichnet werden, das heißt über eine Datenstruktur, die zu jedem Eintrag (hier: zu jedem N-Gramm) die Texte verzeichnet, in denen er vorkommt.[352]

Als Alternative zum Vergleich aller N-Gramme wird in der Literatur auf Verfahren hingewiesen, die nur einen nach einem bestimmten Filterkriterium ausgewählten Teil der N-Gramme berücksichtigen.[353] Diese Verfahren stammen aus dem Bereich der Erkennung von Fast-Dubletten. In der Auswertung des PAN-Wettbewerbs 2010 wird angemerkt, die meisten Teilnehmer hätten einen *Brute-Force*-Vergleich durchgeführt, statt eine entsprechende Auswahl zu treffen.[354] Inwieweit sich die für eine weitgehende Textgleichheit entwickelten Ansätze auch für die Plagiatserkennung eignen, ist allerdings zu hinterfragen – da die Übereinstimmung eines Plagiats mit einer Quelle ja oft nur einen kleinen Textteil betrifft, kann es durchaus sein, dass sie hier die Erkennungsrate deutlich verschlechtern würde.

Bei Textübernahmen ist generell mit Veränderungen zu rechnen, sei es aus inhaltlichen oder aus stilistischen Gründen. In besonderem Maße gilt das für Plagiate, jedenfalls wenn bei ihrer Erstellung versucht wird, die Überprüfung auf Übernahmen durch Variation in den Formulierungen zu erschweren. Bei den PAN-Wettbewerben gehört deshalb auch die Erkennung solcher verschleierten Plagiate zur Aufgabenstellung. Die Textänderungen in den Korpora von 2009 bis 2011 wurden maschinell erstellt, im Korpus von 2012 teils maschinell, teils von Menschen – für

[351] Wenn t die Zahl der verzeichneten Texte ist, gibt es $t\,(t-1)\,/\,2$ entsprechende Paare. Der Divisor 2 beruht dabei auf der Annahme, dass es keinen Unterschied macht, ob Text *a* mit Text *b* vergleichen wird oder Text *b* mit Text *a*.

[352] Laut Potthast u. a. 2010, S. [5] wurde im PAN-10-Wettbewerb überwiegend mit invertierten Indices gearbeitet, in denen *Fingerprints* der N-Gramme verzeichnet wurden.

[353] Potthast 2011, S. 32 listet entsprechende Methoden auf.

[354] Potthast u. a. 2010, S. [5].

die automatisch generierten Texte wird ausdrücklich darauf hingewiesen, dass sie nicht unbedingt verständlich seien.[355]

Inwieweit automatisch generierte Fälle dieselben Anforderungen für die Erkennung stellen wie tatsächliche Plagiatsfälle, soll in dieser Arbeit nicht weiter untersucht werden.[356] Für den hier betrachteten Zusammenhang ist aber hervorzuheben, dass die für die Erkennung heutiger Plagiate zugrunde gelegten Annahmen für frühneuhochdeutsche Texte in verschiedener Hinsicht nicht zutreffen[357]:

- Schon bei der Zerlegung in laufende Wortformen, die für alle wortbasierten Verfahren erforderlich ist, ist mit Abweichungen zwischen eigentlich übereinstimmenden Textstücken zu rechnen, die teils darauf beruhen, dass die Abgrenzung durch Leerraum nicht immer den heutigen Kriterien entspricht, teils darauf, dass sie bei der Transkription der Texte vielfach rein optisch nicht klar zu entnehmen ist.
- Eine Untergliederung in Sätze ist insbesondere in den älteren Texten, aber auch später oft nicht sicher möglich, da sich die heutige Zeichensetzung erst allmählich entwickelte und die Interpunktion in frühneuhochdeutschen Texten oft anderen Prinzipien folgt, ohne dass diese immer einheitlich wären.[358]
- Die Schreibungen der Wörter weisen eine ganz erhebliche Varianz auf, die nicht nur auf orthographischer Freiheit (beziehungsweise dem Nebeneinanderbestehen verschiedener Schreibregeln) beruht, sondern auch auf dialektalen Einflüssen.
- Eine Abbildung auf Wortstämme wird nicht nur durch die Schreibungsvielfalt erschwert, sondern auch dadurch, dass keine auch nur einigermaßen umfassende Wortliste für das gesamte Alphabet vorliegt, die als Korrektiv dienen könnte. Erst recht ist es mangels einer entsprechenden Datenbasis nicht möglich, eine automatische Vereinheitlichung über die Ersetzung durch als synonym betrachtete Wörter – nach welchen Regeln auch immer sie erfolgen könnte – vorzunehmen.[359]

Angesichts dieser Punkte ist davon auszugehen, dass sich die Verfahren, die für die Plagiatserkennung entwickelt wurden, nur begrenzt auf die Erkennung von Übernahmen in frühneuhochdeutschen Texten übertragen lassen, jedenfalls sofern

[355] Potthast u. a. 2009, S. 4.

[356] Für den PAN-Wettbewerb 2012 wurden erstmals auch von Menschen umgeschriebene Texte in die Aufgabenstellung einbezogen, da die früher verwendeten Korpora als unzureichend eingestuft wurden, vgl. Potthast u. a. 2012, S. [3].

[357] Vgl. oben Kapitel 1.3.

[358] Eine Satzabgrenzung über den Vergleich mit syntaktischen Mustern würde nicht nur voraussetzen, dass diese – vom heutigen Deutsch nicht selten abweichenden – Muster hinreichend bekannt sind, sondern auch, dass die Wörter jeweils ihrer syntaktischen Rolle zugeordnet werden können, was ohne entsprechende aufwendige Korpusaufbereitung wohl kaum möglich ist.

[359] Das sowohl für eine Lexemliste als auch für die Ermittlung von bedeutungsverwandten Wörtern heranzuziehende *Frühneuhochdeutsche Wörterbuch* (FWB) deckt bisher die Anfangsbuchstaben *a*, *b/p*, *c/k*, *d/t*, *g*, *i/j*, *l* und *m* komplett und außerdem die Wortstrecken *e – er-*, *h – hexerei*, *n – neigen* und *st – stoszug* ab.

diese Texte nicht weiter aufbereitet sind – zum Beispiel durch eine Lemmatisierung – und auch keine Hilfsmittel zur Verfügung stehen, die eine automatische Normalisierung ermöglichen würden. Während die Schreibvarianz frühneuhochdeutscher Texte in der vorliegenden Untersuchung weitgehend berücksichtigt ist, wird hier nicht versucht, auch Passagen zu finden, die zwar auf Vorlagen basieren, aber umformuliert sind und keine etwas längere unveränderte Wortfolge enthalten.

Dies stellt natürlich eine Einschränkung dar. Allerdings ist zu vermuten, dass von den Verfassern der hier ausgewerteten Texte kein Anlass gesehen wurde, Formulierungen zu ändern, nur um Passagen nicht einfach wörtlich zu übernehmen, und dass dementsprechend der Erkennung auch von nicht völlig wortgleichen Textstücken wesentlich weniger Gewicht zukommt als bei der Untersuchung von Texten auf möglicherweise verschleierte Plagiate.

2.5 Erforschung von *Text Reuse*

Das im Vergleich zur Plagiatserkennung allgemeinere Problem der automatischen Ermittlung intertextueller Beziehungen hat bis vor kurzem allem Anschein nach nur zu vereinzelten Forschungsaktivitäten geführt.[360] In den letzten Jahren sind allerdings einige größere Arbeiten zu diesem Bereich erschienen, so dass sich hier eine neue Forschungsrichtung etablieren könnte.[361]

2.5.1 Definition und Kategorisierung

Der Gegenstand dieser Forschungen wird dabei auch in deutschsprachigen Publikationen als „*Text Re-use*“ (oder „Text-Reuse“[362]) bezeichnet. Ob damit stets das Gleiche gemeint ist, darf allerdings wohl bezweifelt werden.

So heißt es in einem Aufsatz von Rao Muhammad Adeel Nawab, Mark Stevenson und Paul Clough: „Text reuse is the process of creating new document(s) using text from existing document(s).“[363] Die Formulierung „using text“ lässt sich dabei wohl nur als Übernahme zumindest von mehreren Wörtern (in einer vielleicht unterbrochenen Folge oder in einer anderen Anordnung, aber doch in einem syntaktischen Zusammenhang) deuten. Dazu passt, dass der Terminus in verschiedenen Publikationen verwendet wird, die sich mit der Erkennung von Textabhängigkeiten im Zeitungswesen oder im WWW befassen.[364]

[360] So auch der Befund von Potthast 2011, S. 5, der zwei Projekte zu *Text Reuse* im Zeitungswesen – eines davon im historischen Rahmen – sowie ein Projekt zur *Text-Reuse*-Erkennung im WWW verzeichnet.

[361] Vgl. Büchler 2013, S. 45 f. und 53 (fokussiert auf den „*Historical Text Re-use*“).

[362] So Kümmel 2011. Büchler 2013, S. 56 Anm. 2 begründet den Bindestrich in „*Re-use*“ mit der Missverständlichkeit der Schreibung *Reuse* im Deutschen.

[363] Nawab/Stevenson/Clough 2012, S. 54.

[364] Vgl. Clough u. a. 2002, Potthast 2011, Nawab/Stevenson/Clough 2012 und Snowsill 2012.

Wesentlich umfassender liest sich die Beschreibung von Marco Büchler: „*Text Re-use* beschreibt die mit unterschiedlichen Absichten mündliche und schriftliche Wiedergabe von Textinhalten. Diese können im Sinne einer Definition das Anerkennen einer Autorität aber auch das Wiedergeben einer besonders interessanten Information sein."[365] Hier wird also keine wörtliche Übereinstimmung vorausgesetzt, sondern auch zum Beispiel eine Paraphrase unter diesen Begriff gefasst.

Zudem geht es für Büchler nicht nur wie in der zuerst angeführten Definition um Schriftstücke, sondern ausdrücklich auch um mündliche Äußerungen. Als „45 der wichtigsten und dominantesten *Meme* des *Historical Text Re-use*"[366] führt er auf: „*Adage, Abstract, Anagram, Aphorism, Apophthegm, Battle Cry, Bonmot, Cliché, Definition*", „*Edition, Epigram, Epithet, Epitome, Fact, Flowery Phrase, Gnome, Idiom*", „*Joke, Koan, Law, Legend, Loanword, Mantra, Maxim, Meme, Metaphor*", „*Motto, Palindrom, Pangram, Parable, Paroimia, Phraseme, Platitude, Proverb, Punch Line*", „*Quip, Rant, Saw, Sententiae, Simile, Slogan, Template, Truism, Wit*".[367] Ohne die einzelnen Kategorien näher zu betrachten, lässt sich aus dieser Zusammenstellung wohl schon unmittelbar entnehmen, dass es hier um im Prinzip eigenständige Einheiten mit oft mündlicher Verbreitung und häufiger Wiederverwendung geht, die kaum geeignet sind, als Indizien für die unmittelbare intertextuelle Beziehung zwischen zwei Schriftstücken zu dienen. Auch die direkte Bezugnahme auf einen anderen Text gehört aber für Büchler natürlich zum *Text Reuse*, der in seinem Verständnis „absichtliche und unabsichtliche Zitationsspuren" umfasst, wobei er als Beispiele für Ersteres „*Zitate*, *Paraphrasen* und *Allusionen*" nennt.[368]

Büchler bietet neben dem Kategorisierungsschema für die Textstücke, in denen *Text Reuse* festgestellt wird, auch ein Schema für die Klassifikation der Beziehung, die bei einem *Text Reuse* vorliegt (im Sinne der Darstellung in einem Graphen als „*Edge Type System*" bezeichnet). Dabei nennt er als Kriterien die Art des *Text Reuse* (entweder mehr oder weniger wörtlich oder aber nur semantisch ähnlich sowie „*Incomplete Text Re-use*", wobei der zuletzt genannte Typ von den anderen beiden – anscheinend rein pragmatisch – dadurch unterschieden ist, dass eine Quelle zugrunde liegt, die nicht zum Korpus gehört), den Umfang und das Umfangsverhältnis der „*Text Re-use Unit*" in Quelle und übernehmendem Text sowie gegebenenfalls noch den Grad der Veränderung. Daraus ergeben sich fünfzehn

[365] Büchler 2013, S. VI. Büchler u. a. 2010, S. 2 erläutert, die Bezeichnungen „*reuse*" und „*textual reuse*" anstelle von „*citation*" seien deshalb gewählt, weil zum Zitieren nach gängigem Verständnis die Nennung der Quelle gehöre, dem das nicht gekennzeichnete Plagiat gegenübergestellt werde – diese Unterscheidung treffe aber auf derartige Textbezüge zum Beispiel in den klassischen griechischen Werken nicht zu, weil darin ein gewisser Textkanon als allgemein bekannt vorausgesetzt werde, so dass die Nennung des Referenztextes oft unterbleibe.

[366] Büchler 2013, S. 76.

[367] Ebd. S. 71–75.

[368] Ebd. S. 51.

Typen,[369] die teilweise (mehr oder weniger) den Arten inter- beziehungsweise hypertextueller Beziehungen nach dem von Gérard Genette entwickelten Schema[370] entsprechen, nämlich im Hinblick auf Intertextualität im Sinne Genettes zum Beispiel „*Verbatim*“ und „*Near Verbatim*“ als Untertypen von „*Quotation*“, im Hinblick auf Hypertextualität zum Beispiel „*Summarizing*“. Das Klassifikationsschema umfasst aber auch den „*Idiomatic Text-Reuse*“ mit den Typen „*Idiom*“ und „*Winged Word*“, also Übereinstimmungen, die auf fest gefügten Formulierungen beruhen.

Insgesamt lässt sich wohl feststellen, dass das, was in dieser Arbeit untersucht werden soll, allenfalls einen kleinen Bereich des von Büchler beschriebenen Forschungsgegenstands ausmacht. Zum einen soll es nur um die Erkennung von zumindest weitgehend wörtlich übereinstimmenden Übernahmen gehen, zum anderen jedenfalls primär nicht um die Untersuchung einzelner Meme (die für eine derartige Untersuchung auch erst einmal eruiert werden müssten), sondern um den vollständigen Vergleich ganzer Texte, das heißt um die Ermittlung von Passagen mit vorher unbekannter Ausdehnung, die sich aufeinander abbilden lassen. Eine derartiger vollständiger Vergleich aller Texte eines größeren Korpus miteinander lässt sich für die Arten von *Text Reuse*, die Büchler im Blick hat, wohl nur schwer leisten – darauf deutet jedenfalls sein Vorschlag hin, ihn durch parallele Bearbeitung auf einer Vielzahl von Rechnern zu bewältigen.[371]

2.5.2 Analyseschritte

Neben dem vorgestellten Entwurf eines inhaltlichen Klassifikationsschemas für die Beschreibung von *Text Reuse* systematisiert Marco Büchler auch die Verarbeitungsschritte, die sich in den von ihm bearbeiteten Projekten für die Erkennung als sinnvoll erwiesen haben und im *TRACER*-Tool[372] so implementiert wurden, dass jeder Schritt separat konfiguriert werden kann, und er stellt Verfahren zur Evaluation der gefundenen Ergebnisse vor. Als Schritte beschreibt er „Segmentierung“, „Preprocessing“, „Featuring“, „Selection“, „Linking“, „Scoring“ und „Postprocessing“.[373] Die im Folgenden gebotenen kurzen Erläuterungen orientieren sich an dieser Darstellung, enthalten allerdings eigene Ergänzungen, die die Motivation für die Verarbeitungsschritte erhellen sollen. Sie dürften – soweit sie nicht aufgrund der Formulierung als Überlegungen zu erkennen sind – inhaltlich wohl unstrittig sein.

- Die Segmentierung ist der erste Schritt, um die Textstücke zu erhalten, deren – wie auch immer definierte – Ähnlichkeit miteinander ermittelt werden soll. Sie kann

[369] Eine graphische Darstellung bietet Büchler 2013 S. 77 (vgl. ebd. S. 76 und 78).
[370] Vgl. oben Unterkapitel 1.1.2.
[371] Büchler 2013, S. 226. Vgl. zum Skalierungsproblem auch Büchler u. a. 2014.
[372] Vgl. dazu unten S. 95.
[373] Im Überblick Büchler 2013, S. 89 f., ausführlich ebd. S. 91–122.

so erfolgen, dass die dabei gebildeten Textausschnitte einander überlappen, oder so, dass sie aufeinander folgen.[374] Für Letzteres bietet sich insbesondere eine Aufteilung in Sätze an, sie ist allerdings vor allem bei historischen Texten oft problematisch, weil die Zeichensetzung, anhand deren sie noch am ehesten – mit einer gewissen Fehlerquote – automatisch vorgenommen werden kann, vielfach nicht original ist, sondern auf editorischen Entscheidungen beruht.[375]

- Das *Preprocessing* dient der Reduzierung von Varianz, soweit diese die Erkennung von *Text Reuse* behindert. Das betrifft zum Beispiel unterschiedliche Möglichkeiten der Textcodierung wie etwa die alternative Erfassung von Buchstaben mit Diakritika als ein oder als zwei Zeichen, Schreibungsvarianten aufgrund von unterschiedlichen orthographischen Regeln, auf Dialekten oder auf der historischen Entwicklung beruhende sprachliche Unterschiede, unterschiedliche Flexionsformen eines Worts oder sogar – wenn auch inhaltliche Übereinstimmungen bei abweichender Wortwahl gefunden werden sollen – Synonyme und Kohyponyme.[376]
- Im *Featuring* wird ermittelt, welche nach bestimmten Kriterien gebildeten kleineren Einheiten sich aus den im *Preprocessing* erstellten Repräsentationsformen der Segmente extrahieren lassen. Das können zum Beispiel Wort-N-Gramme, Kookkurrenzen oder Textstücke sein, die einem bestimmten formalen Schema entsprechen (etwa typische Formeln zur Kennzeichnung von Zitaten).[377]
- In der folgenden *Selection* werden aus den ermittelten *Features* diejenigen ausgewählt, die für den (ersten) Vergleich berücksichtigt werden sollen. Grund für diesen Schritt ist offenbar die enorme Komplexität, die sich ohnehin beim Vergleich größerer Textmengen und erst recht bei einer vollständigen Berücksichtigung aller *Features* (jedenfalls wenn diese zahlreich sind) ergeben kann.[378]
- Das *Linking* erfolgt durch die Erstellung einer Link-Matrix, die die Übereinstimmungen in den ausgewählten *Features* aller zu untersuchenden Segmentpaare verzeichnet. Daraus ergibt sich eine quadratische Komplexität. Bei größeren Textmengen reicht oft der Arbeitsspeicher nicht aus. Die Wahl eines zu den jeweiligen

[374] Ebd. S. 91.

[375] Vgl. ebd. S. 91. Das Problem zeigt sich auch bei der Arbeit mit frühneuhochdeutschen Texten unter Zugrundelegung der originalen Schreibung, da die Satzzeichen darin oft nicht entsprechend den heutigen Konventionen gesetzt sind (vgl. Reichmann/Wegera 1993, S. 29–31) und gerade auch bei Textübernahmen nicht unbedingt übereinstimmen. Büchler 2013, S. 92 führt auch weitere Möglichkeiten der Segmentierung auf, zum Beispiel auch solche nach Seite oder Zeile der zugrunde gelegten Textausgabe.

[376] Vgl. Büchler 2013, S. 94–96 (dort auch Weiteres).

[377] Vgl. ebd. S. 100–103 (dort auch Weiteres).

[378] Büchler weist zum Beispiel ebd. auf S. 106 darauf hin, dass sich durch die Ausklammerung aller *Features* mit einer hohen Frequenz die Geschwindigkeit erhöhen lässt, führt allerdings zugleich als Problem an, dass für manche zu untersuchende Textstücke dann gar kein *Feature* mehr übrig bleibt. Dass sich bei der Ermittlung von *Text Reuse* generell leicht Laufzeitprobleme ergeben können, ist auch Büchler 2013 verschiedentlich zu entnehmen, zum Beispiel auf S. 123 und 165.

Daten passenden Verfahrens zur Zwischenspeicherung kann die Laufzeit stark beeinflussen.[379]

- Im *Scoring* wird ein wie auch immer bestimmtes Ähnlichkeits- beziehungsweise Distanzmaß für die verlinkten Segmentpaare berechnet. Es muss sich nicht unbedingt an den ausgewählten *Features* orientieren, sondern kann zum Beispiel auch auf dem Anteil von übereinstimmenden Wörtern beruhen.[380]
- Das *Postprocessing* dient der Auswertung der gefundenen Zuordnungen. Dabei kann die Treffermenge nach weiteren Kriterien bereinigt werden, um für eine bestimmte Fragestellung möglichst aussagekräftig zu sein. Beispiele hierfür sind die Ermittlung von Textabhängigkeiten und Textrezeption, die Untersuchung auf Merkmale wie Versmaß und Reim, die die Memorierbarkeit fördern, und die Zuordnung zu den oben angeführten Mem- und Beziehungs-Kategorien.[381]

Diese Abfolge von Schritten ist in ihrer ganzen Komplexität zwar nicht zwangsläufig erforderlich, es gibt aber zum Teil deutliche Ähnlichkeiten zu anderen Beschreibungen, auch aus dem Bereich der Plagiatserkennung. So ist es wohl allgemein anerkannt, dass ein Vorverarbeitungsschritt durchgeführt werden sollte, in dem zumindest gewisse Vereinheitlichungen in den Schreibungen durchgeführt werden – auch für moderne Texte legt sich zum Beispiel die Zusammenfassung von Groß- und Kleinschreibung nahe, da hierbei mit für den Vergleich irrelevanten orthographischen Varianten und Schreibunsicherheiten zu rechnen ist – und je nach Ziel und Umsetzbarkeit gegebenenfalls auch eine Lemmatisierung oder Synonymenersetzung[382]. Wenn man davon ausgeht, dass die zu untersuchenden Texte nicht in einer auf den reinen Zeichenbestand beschränkten Form vorliegen, sondern zum Beispiel als XML-Dateien (oder auch in einem binären Format), ist es zudem erforderlich, zunächst einmal diese Form zu extrahieren.[383] Umgekehrt sollte dann auch nach Abschluss der Stellenzuordnung wieder eine Abbildung zumindest auf die originalen Schreibungen erfolgen, soweit die ermittelten Textstücke als solche präsentiert werden. Die Bildung von Segmenten – sei es in Form von N-Grammen, sei es durch Unterteilung an Satz- oder Absatzgrenzen – ist gängige Praxis in der Plagiatserkennung, die Auswahl bestimmter *Features* für die schnelle Überprüfung auf auffällige Ähnlichkeiten zum Teil.[384]

[379] Vgl. ebd. S. 114 f.

[380] Vgl. ebd. S. 116–119 (dort auch Weiteres).

[381] Vgl. ebd. S. 120–122 (dort auch Weiteres). Dem *Postprocessing* kann wohl auch die Aufbereitung der ermittelten Daten für die Darstellung zugeordnet werden. Dieser Punkt wird von Büchler anscheinend nicht genannt, ist aber erforderlich, wenn es zum Beispiel darum geht, die ermittelten Stellen im Textzusammenhang zu zeigen beziehungsweise einen Text unter Hervorhebung und vielleicht auch Verlinkung dieser Passagen zu präsentieren.

[382] Entsprechende Schritte werden für die Plagiatserkennung beschrieben, vgl. oben S. 85. Die Umwandlung in Kleinbuchstaben scheint gängige Praxis zu sein, vgl. Tabelle 2.5 auf S. 98.

[383] Dieser Schritt wird in Büchler 2013 anscheinend als schon erledigt vorausgesetzt, wenn er nicht ein unerwähnter Teilschritt der Segmentierung ist (was etwa bei der Zugrundelegung von XML-Texten mit entsprechend auswertbaren *Tags* plausibel wäre).

2.5.3 Projekte und Programme

Was die automatische Ermittlung von *Text Reuse* derzeit zu leisten vermag und welche Erkennungskriterien dabei eingesetzt werden, soll anhand einiger Projekte, die sich mit der Untersuchung von Werken des kulturellen Erbes beschäftigen, und einiger vor allem in diesem Zusammenhang entwickelter Programme ein wenig näher betrachtet werden.

Wohl an erster Stelle sind die Arbeiten aus der Forschungsgruppe um Gerhard Heyer zu nennen, insbesondere *eAQUA* („Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft")[385] und *eTraces*[386], sowie die 2015 eingerichtete Nachwuchsgruppe *eTRAP* („electronic Text Reuse Acquisition Project")[387] unter Leitung von Marco Büchler.

Mit der Untersuchung der indirekten Überlieferung (im Sinne von „syntaktischen Übereinstimmungen"[388], also mehr oder weniger wörtlichen Formulierungsübernahmen) von Platons *Timaios* in der Antike beschäftigte sich ein Teilprojekt von *eAQUA*. Zur Ermittlung von Übereinstimmungen wurden dabei verschiedene Kriterien wie identische Wortketten und Kookkurrenzen unter Berücksichtigung des Abstands, jeweils bezogen auf durch Satzzeichen abgegrenzte Analyseeinheiten, eingesetzt; daran schloss sich eine manuelle Auswahl und Klassifikation der tatsächlich als indirekte Überlieferung zu bewertenden Stellen an.[389] Ein besonderes Anliegen im Projekt war die Entwicklung geeigneter Darstellungsformen, die zum Beispiel den Feinvergleich von Varianten erleichtern oder auch die Häufung von Übernahmen aus bestimmten Textabschnitten oder die Entwicklung der Rezeption eines Werks durch graphische Aufbereitung veranschaulichen.[390] Die „Zitationssuche" wurde – neben weiteren Komponenten – in einer zweiten Projektphase weiterentwickelt und kann über das Portal des Projekts für die Suche nach Übereinstimmungen zwischen Texten bestimmter Korpora genutzt werden.[391]

[384] Vgl. Potthast u. a. 2010, S. [5], wo die Berücksichtigung aller N-Gramme als „brute force" beschrieben wird. Die PAN-Wettbewerbe von 2012 und 2013 stellten ausdrücklich als Aufgabe, Plagiate durch möglichst wenige Suchmaschinenanfragen zu ermitteln, vgl. *http://www.uni-weimar.de/medien/webis/events/pan-12/pan12-web/plagiarism-detection.html* und *http://www.uni-weimar.de/medien/webis/events/pan-13/pan13-web/plagiarism-detection.html*.

[385] Vgl. *http://www.eaqua.net/*.

[386] Vgl. *http://asv.informatik.uni-leipzig.de/de/projects/25*.

[387] Vgl. *http://etrap.gcdh.de/*.

[388] Gessner 2010, S. 27.

[389] Ebd. S. 28.

[390] Vgl. ebd. S. 31–36 und Büchler u. a. 2010, S. 11–13. Die Verwendung des *CitationGraph* für die umgekehrte Frage, welche Quellen Plutarch in der Perikles-Vita zitiert, wird in Schubert 2010 S. 42–48 dargestellt, die Verwendung für die Untersuchung von nur fragmentarisch überlieferten Texten ebd. S. 48–54. Ergebnisse eines Forschungsseminars, in dem das Tool für die Ermittlung von Zitationen in einigen Werken Plutarchs eingesetzt wurde, werden in Schubert/Klank (Hg.) 2012 beschrieben.

Das Projekt *eTraces* zielte auf eine „Recherche und Analyse von Zitationsspuren und Wissenstransfer in sozialwissenschaftlichen Texten und deutschsprachiger Literatur“[392] und hatte unter anderem eine Untersuchung von Bibelzitaten in deutschsprachigen Romanen aus der Zeit von 1500 bis 1900 zum Gegenstand. Zunächst im Rahmen dieses Projekts wurde das *TRACER*-Tool entwickelt, in dem für die oben beschriebenen Verarbeitungsschritte jeweils vielfältige Konfigurationsmöglichkeiten zur Verfügung stehen. Die damit gegebene Anpassbarkeit führt freilich zu einer je nach gewählten Einstellungen teilweise recht erheblichen Verarbeitungskomplexität, die vor allem für die Untersuchung größerer Datenbestände den Einsatz von sehr leistungsfähiger Hardware nahelegt.[393]

Im Rahmen des Heidelberger Exzellenzclusters „Asia and Europe in a Global Context“ wurde das Programm *QuotationFinder* entwickelt.[394] Es soll zum einen die Ermittlung von Zitaten aus einem Text in einer Gruppe von Texten ermöglichen, zum anderen den Vergleich aller möglichen Textpaare in einer solchen Gruppe.

Dabei stehen jeweils mehrere Erkennungsverfahren zur Verfügung. Bei der Suche nach Zitaten aus einem Text kann zwischen einer Untergliederung in Sätze oder Teilsätze (jeweils anhand von vorgegebenen Satzzeichen abgegrenzt) und einer Bildung von Wort-N-Grammen (beziehungsweise von Zeichen-N-Grammen für Chinesisch und Japanisch) mit einer Größe von n zwischen 3 und 20 gewählt werden; allerdings können maximal 10.000 Segmente verarbeitet werden, so dass eine Untersuchung auf N-Gramm-Basis nur bis zu dieser Wortanzahl möglich ist.

Für den Vergleich aller Textpaare steht eine Untergliederung in Einzelwörter, Wort-Trigramme (beziehungsweise Zeichen-Trigramme für Chinesisch und Japanisch), Teilsätze und Sätze zur Verfügung, die jeweils nach den Kriterien „*Similarity*“, „*Resemblance*“, „*Index Resemblance*“, „*Containment*“ und „*Index Containment*“[395] untersucht werden können, außerdem können Wort-, Satz- und Absatzzahl sowie

[391] Das Tool kann von *http://www.eaqua.net/* aus bei aktiviertem Javascript über den Menüpunkt *Tools – Demonstration Zitationen* aufgerufen werden. In der Demonstration stehen eine tabellarische Anzeige sowie eine Auswertung in Diagrammform zur Verfügung. Vgl. zur Projektgeschichte *http://www.eaqua.net/*, Menüpunkt *Über eAQUA* (dort ist auch die zitierte Bezeichnung des Tools zu finden).

[392] So der Untertitel für das Projekt, zum Beispiel in *http://www.gesis.org/forschung/drittmittelprojekte/archiv/etraces/*.

[393] Ich danke Dr. Marco Büchler für die Möglichkeit, das Programm zu verwenden, und insbesondere auch für die Einführung in die Benutzung im Rahmen eines *Hackathons*. Die Software steht auf Anfrage zur Verfügung (vgl. *http://www.etrap.eu/research/tracer/*).

[394] Vgl. *http://www.asia-europe.uni-heidelberg.de/de/forschung/heidelberg-research-architecture/projekte/hra4-quotationfinder/about.html*. Die Entwicklung wurde anscheinend 2010 mit der Version 0.5 (beziehungsweise der Beta-Version 0.6) eingestellt, vgl. die Übersicht über die Programmversionen unter *http://www.asia-europe.uni-heidelberg.de/de/forschung/heidelberg-research-architecture/projekte/hra4-quotationfinder/download.html*. Erläuterungen zur Benutzung des Programms bietet ein *Tutorial*, das in der gepackten Download-Datei enthalten ist.

[395] Damit wird die Möglichkeit geboten, wiederholte Vorkommen derselben Elemente unterschiedlich zu gewichten und Unterschiede in der Textgröße in die Bewertung von Ähnlichkeiten

Vokabulargröße der Texte verglichen werden. Die Verarbeitung von XML-Dateien ist zwar – anders als die von einigen anderen, auch binären Formaten – nicht vorgesehen, lässt sich aber erreichen, wenn man das Dateinamensende in „.htm“ ändert, so dass sie als HTML-Dateien eingelesen werden. Es gibt allerdings keine Möglichkeit, Regeln für eine weitergehende Datenaufbereitung im Sinne des von Marco Büchler beschriebenen *Preprocessing* festzulegen.

Bei Tests mit Texten aus dem hier zugrunde gelegten Korpus zeigte sich zwar kein Erfolg für den Vergleich von Sätzen oder Teilsätzen, wohl aber, dass die Trigramm-Analyse trotz einer nicht allzu geringen Schreibungsvarianz für die Erkennung von Textübernahmen zumindest dann recht gut geeignet ist, wenn diese den Großteil der jeweiligen Texte ausmachen. Da die Textpaare allerdings anscheinend tatsächlich jeweils nacheinander verglichen werden, steigt die Laufzeit wohl in etwa quadratisch mit der Größe des Korpus.[396]

Das – inzwischen anscheinend nicht mehr zum Download angebotene – Programm *Ferret*[397] stammt zwar nicht aus dem Bereich der *Text-Reuse*-Forschung, sondern wurde mit dem Ziel der Plagiatserkennung entwickelt, kann aber ähnlich wie der *CollationFinder* für den Vergleich größerer Textmengen auf der Basis von Wort-Trigrammen genutzt werden. Weitere Vergleichsmöglichkeiten stehen hier nicht zur Verfügung, die Trigramm-Analyse wird allerdings – soweit der Arbeitsspeicher ausreicht – auch bei größeren Textbeständen mit sehr hoher Geschwindigkeit durchgeführt, da der Vergleich zunächst einmal nicht zwischen den einzelnen Textpaaren erfolgt, sondern die Trigramme aller Texte verzeichnet werden, um daraus das Maß an Übereinstimmung zwischen den Paaren zu ermitteln. Um nicht nur diesen Wert zu erhalten, sondern die jeweiligen Entsprechungen im Textzusammenhang nebeneinander zu sehen, muss anschließend für das jeweils zu betrachtende Textpaar eine zusätzliche Analyse durchgeführt werden, die allerdings vergleichsweise zeitaufwendig ist.[398] *Ferret* bietet wie der *CollationFinder* die Möglichkeit zur Verarbeitung von einigen binären Formaten, allerdings kein spezielles Verfahren für das Einlesen von HTML- oder XML-Dokumenten.

einfließen zu lassen. Die Berechnung wird im *Tutorial* (und im Programm selbst bei Klick auf das jeweilige Fragezeichen neben einem Kriterium) anhand von Beispielen erläutert.

[396] Nach meinen – nicht allzu umfangreichen – Tests scheint dabei vor allem die Zahl der zu vergleichenden Texte, nicht oder nicht so sehr ihre jeweilige Größe, entscheidend zu sein.

[397] Die folgende Darstellung basiert auf der Programmversion 4.0 (frühere URL: *http://homepages.feis.herts.ac.uk/~comqpcl/Downloads/ferret4.0-MSWindows-installer.exe*, am 28. 12. 2017 nicht mehr aufrufbar). Das Programm wurde für *Linux* wohl noch bis zur Version 5.4 weiterentwickelt, die letzte Projektseite *https://github.com/petercrlane/ferret* wurde inzwischen aber offenbar ebenfalls gelöscht.

[398] Während sich die Trigramm-Analyse kleinerer Textmengen auf dem Testsystem in wenigen Sekunden oder sogar noch schneller durchführen ließ, erforderte die Anzeige eines Textpaars mit einer auffälligen, aber noch relativ kleinen Anzahl an Entsprechungen auch bei nicht allzu großen Dateien schon mehrere Minuten.

Als Ergänzung zu *PhiloLogic*, einem im Rahmen des Projekts *American and French Research on the Treasury of the French Language (ARTFL)* entwickelten Volltext-such- und -analysesystem,[399] wurden zwei für die hier betrachtete Fragestellung einschlägige Tools erstellt: *PAIR* ermöglicht den Vergleich eines Textes mit einem in einer Datenbank aufbereiteten Textkorpus, *PhiloLine* den vieler Texte miteinander.[400] Das Erkennungsverfahren basiert darauf, dass in einer Vorverarbeitung XML-*Tags* und – je nach Konfiguration – auch der Inhalt bestimmter XML-Elemente, Stoppwörter und kurze Wörter entfernt sowie kleinere Schreibungsvarianten vereinheitlicht werden, dann nach übereinstimmenden N-Grammen gesucht und schließlich von diesen aus das jeweilige Textumfeld so weit erweitert wird, wie sich Übereinstimmungen entsprechend bestimmten Kriterien ermitteln lassen. Dabei sind unterschiedliche Einstellungen zum Beispiel zur Art der Vorverarbeitung oder zum geforderten Übereinstimmungsgrad für die Kontexterweiterung möglich.[401]

Die drei zuletzt vorgestellten Programme weisen einige Ähnlichkeiten, aber auch Unterschiede auf. Tabelle 2.5 auf S. 98 bietet dazu einen Überblick, der auch die technischen Rahmenbedingungen wie die Verfügbarkeit für verschiedenen Betriebssysteme und die unterstützten Dateiformate zumindest oberflächlich beinhaltet und anhand jeweils zweier Testläufe einen Eindruck vom benötigten Zeitaufwand vermittelt.[402] Offenbar gibt es Unterschiede zwischen den Leistungen und Vorgaben der Programme, die für die Verwendbarkeit in einem konkreten Fall entscheidend sein mögen. Für die hier vorliegende Fragestellung ist wohl insbesondere bemerkenswert, dass alle drei Programme den Vergleich von mehreren Texten miteinander ermöglichen und dass die Untersuchung auf Textentsprechungen jeweils über die Analyse der enthaltenen Wort-Trigramme in Kleinschreibung und ohne Satzzeichen erfolgt oder jedenfalls erfolgen kann.

Daneben sind natürlich – je nach konkretem Erkennungsziel – auch andere Kriterien für die Zuordnung von Textstellen möglich. Das Projekt *Tesserae*[403], das

[399] Vgl. *https://sites.google.com/site/philologic3/* sowie inzwischen *https://github.com/ARTFL-Project/PhiloLogic4* und über ARTFL *http://artfl-project.uchicago.edu/*.

[400] Die Programme konnten früher von *http://code.google.com/p/text-pair/downloads/list* aus heruntergeladen werden. Vgl. Horton/Olsen/Roe 2010 und Roe u. a. 2012.

[401] Voraussetzung für die Verwendung entsprechend der beiliegenden *readme.txt*-Datei ist offenbar ein *Linux*-System (oder Ähnliches), außerdem setzt die Funktion *LoadBibliography* im Skript *compareall2allinmemory.pl*, das als letzter Schritt ausgeführt wird, bibliographische Daten entsprechend den *PhiloLogic*-Strukturen voraus.

[402] Die Zeitangaben sind außerdem natürlich abhängig vom verwendeten Rechner und dessen aktueller sonstiger Belastung, und sie sind speziell für den *QuotationFinder* ungenau, weil sich der Zeitaufwand für den eigentlichen Vergleich aufgrund der jeweils manuell durchzuführenden Auswahl der Dateien und der gewünschten Vergleichsmethode (jedenfalls ohne entsprechende Tools) nicht sicher protokollieren ließ. Für *PhiloLine* wurden im Test die eigentlich vorgesehenen Aufrufe von *Linux*-Tools durch eigene *Perl*-Skripte ersetzt.

[403] *http://tesserae.caset.buffalo.edu*.

	Ferret	***PhiloLine***	***QuotationFinder***
Betriebssystem	*Linux*, für Version 4.0 auch *Windows*	*Linux* (o.Ä.)	System mit *Java*
Programmiersprache	C++ (in Version 5.3)	*Perl*	*Java*
Benutzerinteraktion	GUI, in Version 5.3 auch Kommandozeile	Kommandozeile	GUI
einlesbare Formate	*.txt, *.doc, *.rtf, *.pdf; Programm-Quellcode	*.xml	*.txt, *.htm, *.doc, *.rtf
Anzeige/Ausgabe	GUI und PDF: Gesamttabelle und Textpaar mit Hervorhebung übereinstimmender Trigramme (im GUI synoptische Anzeige entsprechend ausgewähltem Trigramm); in Version 5.3 auch XML	HTML: Textausschnitte (paarweise gefundene Übereinstimmungen mit Textumfeld)	Gesamttabelle; im GUI auch Textpaar mit Hervorhebung übereinstimmender Trigramme (synoptische Anzeige entsprechend ausgewähltem Trigramm)
Basis des Vergleichs	auf Kleinbuchstaben reduzierte Textform	auf Kleinbuchstaben (entsprechend Konfiguration) reduzierte Textform	auf Kleinbuchstaben reduzierte Textform
Segmentierung	Wort-Trigramme	Wort-N-Gramme	Wort-Trigramme, Einzelwörter, Teilsätze, Sätze
Konfigurationsmöglichkeiten	Vergleich als Text oder als Programmcode	Wort-N-Gramm-Größe; zu ignorierende Wörter; zu Wörtern gehörende Zeichen; Umgang mit Diakritika und Apostroph; Ausschluss des Inhalts bestimmter XML-Elemente; Schreibungsnormalisierung; Ausschluss häufiger N-Gramme; Größe des zu prüfenden Umfelds (und Weiteres)	Art der Segmentierung; Bewertung des mehrfachen Vorkommens von Segmenten; außerdem Verwendungsmodus „Find Quotations“ mit weiteren Möglichkeiten
Zeitbedarf Wort-Trigramm-Vergleich 20 Texte, ca. 3 MB	ca. 3 Sek. (ohne Abbildung auf entsprechende Textstellen)	144 Sek.	ca. 160 Sek.
Zeitbedarf Wort-Trigramm-Vergleich 40 Texte, ca. 7 MB	ca. 6 Sek. (ohne Abbildung auf entsprechende Textstellen)	366 Sek.	ca. 610 Sek.

Tab. 2.5: Programme zum Vergleich mehrerer Texte miteinander

zunächst zur Ermittlung von Allusionen in lateinischer Dichtung dienen sollte[404] und inzwischen auch für die Untersuchung anderer Werke der lateinischen, griechischen und englischen Literatur eingesetzt werden kann[405], ermöglicht den Online-Vergleich von Textpaaren des jeweils ausgewählten Korpus. Als grundlegendes Auswahlkriterium dient dabei, dass zwei Textpassagen (Verse beziehungsweise Sätze oder Teilsätze entsprechend der Interpunktion[406]) mindestens zwei Wörter

[404] Coffee u. a. 2013, S. 221.

[405] Vgl. die auf *http://tesserae.caset.buffalo.edu/index.php* in der zweiten Menüzeile genannten Rubriken.

[406] Unterhalb der Ebene der Satzgliederung dienen dabei nach Forstall u. a. 2014, S. 2 Semikolon und Komma als Abgrenzungszeichen; *http://tesserae.caset.buffalo.edu/help_advanced.php* nennt nur das Semikolon.

gemeinsam haben.[407] Dabei werden in der Grundeinstellung verschiedene Flexionsformen eines Worts als gleich behandelt und die zehn häufigsten Wörter des Korpus als Stoppwörter ausgeschlossen. Außerdem werden die ermittelten Übereinstimmungen auf der Basis der Vorkommenshäufigkeit und des Abstands der Wörter bewertet und entsprechend sortiert.[408]

In einer im Rahmen des Projekts durchgeführten Untersuchung zur Ermittlung von Einflüssen der *Aeneis* auf Lukans *Bellum civile* wurden die durch das Programm ermittelten Entsprechungen von Hand einer von fünf Klassen zugewiesen, die für unterschiedliche Grade an Korrektheit und Aussagekraft stehen; das Spektrum reicht von der Einordnung als Erkennungsfehler bis hin zu starker formaler Ähnlichkeit mit analogem Kontext. Der automatische Erkennungsschritt findet zwar zum großen Teil Übereinstimmungen, die letztlich als bedeutungslos eingestuft werden, es werden jedoch auch ähnlich viele interpretierbare Allusionen gefunden wie in Kommentaren zu den untersuchten Texten, und dabei handelt es sich oft um abweichende Stellen.[409]

Das in *Tesserae* gewählte grundlegende Erkennungskriterium dürfte freilich – auch wenn man bereit ist, eine Vielzahl an letztlich nicht aussagekräftigen Funden in Kauf zu nehmen – im Hinblick auf Rezeptionsanalysen nur für Texte in Frage kommen, die zum einen eine überdurchschnittlich signifikante Wortwahl aufweisen und zum anderen zumindest passagenweise in ihrem Wortlaut so bekannt waren, dass eine Entsprechung von zwei noch nicht einmal unbedingt aufeinander folgenden Wörtern als Allusion verstanden werden konnte. Das mag für die klassische lateinische Dichtung plausibel sein, dürfte ansonsten aber nur auf recht wenige Werke von zentraler kulturhistorischer Bedeutung sowie für geflügelte Worte, Sprichwörter und Ähnliches zutreffen. Das gemeinsame Auftreten nur zweier Wörter mit geringem Abstand ist wohl in aller Regel eher sprachlich beziehungsweise sachlich bedingt und nicht als Bezugnahme auf eine konkrete literarische Vorlage zu deuten. Zumindest ohne automatisierte Filtermaßnahmen, die zum Beispiel auf Kookkurrenzanalysen beruhen könnten, dürfte sich das beschriebene Verfahren deshalb kaum für den Vergleich von Textpaaren eignen, bei denen nicht ohnehin schon mit einer literarischen Beeinflussung des einen Textes durch den anderen zu rechnen ist.

Das Projekt *Phœbus* zielt auf die Erkennung von textuellen Übernahmen („reuses, citations, borrowings, etc.“) in recht großen Textmengen. Konkret wird insbesondere der Vergleich von Werken Balzacs mit anderen Schriften seiner Zeit sowie später

[407] Coffee u. a. 2013, S. 223.

[408] Vgl. *http://tesserae.caset.buffalo.edu/help_advanced.php*, wo die verschiedenen Einstellungsmöglichkeiten beschrieben werden, sowie ausführlicher zum Bewertungsverfahren Forstall u. a. 2014.

[409] Vgl. Coffee u. a. 2013, vor allem S. 223 f.

der Vergleich französischer Literatur des 19. Jahrhunderts mit der zeitgenössischen Presse angestrebt.[410] Dabei sollen nicht nur exakte Übereinstimmungen erkannt werden, sondern auch solche mit kleineren Änderungen im Wortlaut. Die „reuses and citations" im Sinne dieses Projekts schließen auch unbewusste Übernahmen ein.[411] Dabei geht es nicht um eine quantitative Untersuchung von Übernahmen, sondern vielmehr letztlich um eine Analyse, die eine Beschreibung der jeweiligen Art der Textbeziehung und eine entsprechende Aufbereitung als Hypertext mit annotierten Links ermöglicht.[412]

Für die Erkennung werden die Texte zunächst in eine Form ohne Stoppwörter und mit Abbildung der verbleibenden Wörter auf die Wortstämme transformiert. Anschließend werden ähnlich wie bei Wort-N-Grammen alle Folgen von n transformierten Wörtern, aber einschließlich solcher mit Lücken bis zu einer bestimmten Maximalgröße, miteinander verglichen.[413] Aus den gefundenen Übereinstimmungen werden entsprechend ihren Textpositionen größere Blöcke von einander ähnlichen Textpassagen gebildet. Schließlich erfolgt eine Filterung dieser Blöcke auf der Basis der Wortzahl und einer Bewertung der enthaltenen Wörter: Nur dann, wenn ein Block eine bestimmte Mindestzahl von als signifikant eingestuften Wörtern enthält, wird die Übereinstimmung als relevant betrachtet.[414] Die Präsentation der ermittelten Bereiche mit hoher Ähnlichkeit erfolgt auf zwei Ebenen: zum einen über eine Visualisierung in Form eines Dotplots, zum anderen in einer synoptischen Ansicht.[415]

Die Ziele von *Phœbus* stehen offenbar denen der vorliegenden Untersuchung recht nahe. Das darin letztlich angestrebte Korpus ist deutlich umfangreicher, allerdings geht es anscheinend nicht um einen Vergleich aller Texte miteinander, sondern vielmehr um ein Referenzkorpus, mit dem dann jeweils einzelne Texte verglichen werden sollen, was die Komplexität erheblich reduziert.[416] Die Erkennung

[410] Ganascia/Glaudes/Del Lungo 2014, S. 413 und 420. Die Homepage des Projekts – beziehungsweise eines umfassenderen Projekts unter dem Titel „eBalzac" – ist offenbar *https://phoebus.lip6.fr/*.

[411] Ganascia/Glaudes/Del Lungo 2014, S. 414.

[412] Ebd. S. 413.

[413] Für ein bestimmtes Textpaar, für das auch eine von Hand durchgeführte Ermittlung signifikanter Übereinstimmungen als Referenzgröße vorliegt, wurden verschiedene Parameter getestet. Danach erscheint $n = 3$ bei maximal zwei übersprungenen Wörtern als wohl optimal im Sinne einer Verbindung von hoher *Precision* mit hohem *Recall* (Ganascia/Glaudes/Del Lungo 2014, S. 417). Anscheinend als Weiterentwicklung ist auch der Vergleich ungeordneter Wortfolgen möglich (vgl. Boukhaled/Sellami/Ganascia 2015, S. [2]).

[414] Ganascia/Glaudes/Del Lungo 2014, S. 415 f.

[415] Ebd. S. 418 f. Eine Online-Version unter der URL *http://obvil-dev.paris-sorbonne.fr/phoebus/* kann für die Untersuchung zweier Texte auf Übereinstimmungen genutzt werden. Diese Version listet die ermittelten Übereinstimmungen auf; die Bildung größerer Blöcke, die Dotplot-Visualisierung und die synoptische Ansicht sind dort anscheinend nicht vorgesehen.

[416] Vgl. die in Ganascia/Glaudes/Del Lungo 2014, S. 417–420 genannten Beispiele. Für die Aufbereitung des Textes von Balzacs *Comédie humaine* wird ebd. S. 417 ein Zeitbedarf von weniger

auch von nicht zusammenhängenden und nicht ganz wörtlichen Übereinstimmungen geht ebenfalls weit über das hier Angestrebte hinaus; Voraussetzung für das beschriebene Verfahren ist freilich eine weitgehend einheitliche Sprachform, deren Wortformen sich ohne größere Probleme auf Wortstämme abbilden lassen und für die die Bewertung der Signifikanz von Wörtern beziehungsweise Wortstämmen einigermaßen zuverlässig geleistet werden kann. Für ein Textkorpus, wie es in der vorliegenden Arbeit untersucht werden soll, treffen diese Bedingungen aber bis auf Weiteres nicht zu; vielmehr besteht dafür das Problem, auch ohne sprachtechnologische Hilfsmittel die große Schreibungsvarianz so zu berücksichtigen, dass zumindest eine Erkennung von Übereinstimmungen im Wortlaut möglich ist.

Schließlich soll noch auf das Projekt *Viral Texts*[417] hingewiesen werden, das auf der Basis von nicht korrigierten OCR-Texten Textübernahmen in US-amerikanischen und in der zweiten Projektphase auch britischen und australischen Zeitungstexten des 19. Jahrhunderts untersucht. Das Interesse ist dabei insbesondere auch auf die Ermittlung von geographischen und zeitlichen Zusammenhängen sowie von sozialen Netzwerken gerichtet, die sich aus den Übernahmen erschließen lassen. Mehrfacher Abdruck von Texten in verschiedenen Nummern einer Zeitung soll ausdrücklich nicht verzeichnet werden, vielmehr geht es um Verbindungen zwischen den Zeitungen. Zudem sollten ursprünglich keine kurzen Übereinstimmungen wie etwa Zitate gefunden werden, sondern nur solche mit einer Mindestlänge von etwa 100 Wörtern.[418]

Die Erkennung der Übernahmen erfolgt in mehreren Schritten. Zunächst wird ermittelt, welche Dokumentpaare eine Häufung von gemeinsamen Wort-N-Grammen aufweisen und deshalb als Kandidaten für eine nähere Untersuchung in Frage kommen, dann wird ausgehend von den entsprechenden Textpositionen eine lokale Alinierung durchgeführt, und schließlich werden die alinierten Bereiche zu Regionen mit hoher Ähnlichkeit zusammengefasst. Für die Wort-N-Gramme hat sich für dieses Projekt eine Größe von n zwischen 5 und 7 als sinnvoll erwiesen, wobei N-Gramme mit einer sehr hohen Vorkommenshäufigkeit bei der Auswertung nicht berücksichtigt werden. Außerdem werden auch „skip n-grams“ untersucht, bei denen eine begrenzte Anzahl von Wörtern zwischen den Teilen des N-Gramms stehen kann.[419]

als zehn Minuten genannt, für den anschließenden Vergleich mit kompletten Romanen einige Minuten.

[417] *http://viraltexts.org/*.

[418] Vgl. Smith/Cordell/Dillon 2013, insbesondere Abschnitt 2 und 5, die Informationen auf der eben genannten Homepage des Projekts sowie Smith/Cordell/Mullen 2015, insbesondere Abschnitt IV über damals geplante Weiterentwicklungen.

[419] Diese Beschreibung basiert auf Smith/Cordell/Dillon 2013, Abschnitt 3. In Smith/Cordell/Mullen 2015, Abschnitt I ist nur noch von einer N-Gramm-Größe von 5 die Rede.

Die Aufgabenstellung dieses Projekts weist gerade in technischer Hinsicht deutliche Ähnlichkeiten zu dem in der vorliegenden Arbeit untersuchten Problem auf, allerdings auch einige Unterschiede. Gemeinsamkeiten bestehen insbesondere darin, dass ein Korpus vollständig auf Übereinstimmungen untersucht wird, also nicht nur einzelne Texte oder Textstücke mit einem Korpus verglichen werden, und dass es um die Erkennung von Übernahmen geht, deren Abgrenzung erst bei der Untersuchung ermittelt werden kann.[420] Unterschiedlich ist neben dem konkreten Untersuchungsgegenstand und der unterschiedlichen Fokussierung, die in der vorliegenden Untersuchung auf Texte und Textstücke, in *Viral Texts* aber wohl eher auf soziale Zusammenhänge gerichtet ist,[421] vor allem die Art der Textvarianz, die vom Erkennungsalgorithmus berücksichtigt werden muss. Während für *Viral Texts* insbesondere das Problem von OCR-Fehlern relevant ist,[422] geht es für den Vergleich frühneuhochdeutscher Texte vor allem darum, die große Variabilität der Schreibungen einschließlich der Wortabgrenzung zu berücksichtigen.

2.6 Programme zur automatischen Kollationierung

Der automatisierte Vergleich von verschiedenen Textfassungen beziehungsweise Textzeugen ist ein altes Anliegen beim Einsatz von Computern im Rahmen editorischer Arbeit – erste entsprechende Verfahren wurden wohl in den 1970er Jahren beschrieben.[423] Hier sollen einige bekanntere Lösungen insbesondere im Hinblick auf die für den Vergleich zugrunde gelegten Kriterien kurz vorgestellt werden.

Das vermutlich älteste zu nennende Kollationierungsprogramm, das auch heute noch Verwendung findet, ist das zu *TUSTEP*, dem seit den 1970er Jahren entwickelten „Tübinger System von Textverarbeitungs-Programmen“ gehörende *VERGLEICHE* (sowie einige damit zusammenhängende Kommandos).[424] Das Programm

[420] In Smith/Cordell/Dillon 2013 wird zu Beginn von Abschnitt 3 unter anderem auf diese Punkte als neu gegenüber früheren Untersuchungen hingewiesen.

[421] Auch wenn im Zusammenhang mit der Nichtberücksichtigung von Übereinstimmungen zwischen den verschiedenen Ausgaben einer Zeitung darauf hingewiesen wird, dass diese vielfach zum Beispiel auf Angaben im Impressum oder auf wiederkehrenden Anzeigen beruhen (Smith/Cordell/Dillon 2013, Abschnitt 2 und ähnlich Smith/Cordell/Mullen 2015, Abschnitt II 1 iii), ist doch sicherlich auch in diesem Rahmen der Wiederabdruck zum Beispiel von literarischen Texten gut vorstellbar, und auch die Nutzung journalistischer Beiträge als Vorlagen scheint durchaus plausibel. Wenn es um die Rezeption der betreffenden Texte ginge, wären auch solche Fälle durchaus von Interesse. Das ist aber offenbar nicht die Fragestellung des Projekts.

[422] Laut Smith/Cordell/Mullen 2015, Abschnitt I beträgt die in einer Pilotstudie ermittelte durchschnittliche Fehlerquote auf der Zeichenebene zwischen 5 und 15 %, auf der Wortebene sogar über 25 %.

[423] Vgl. Kochendörfer 1974 über eine Kollationierung von 19 Textzeugen zu 300 Versen des *Parzival* sowie Publikationen zu *TUSTEP*, zum Beispiel Reeg 1977. Vgl. auch die Literaturhinweise in Sahle 2013, Bd. 2, S. 5, Anm. 20; dort werden Titel ab 1970 genannt.

[424] Eine Dokumentation zur aktuellen Programmversion findet sich in Schälkle/Ott 2017, S. 1108–1132.

kann derzeit (soweit nicht zusätzliche Parameter Informationen über zu überspringende Textstücke enthalten) Löschungen und Einfügungen bis zum Abstand von etwa einer DIN-A4-Seite ermitteln, was allerdings als „zeitaufwändig“ beschrieben wird.[425]

Der Vergleich bezieht sich stets auf zwei Texte; gegebenenfalls müssen alle Texte nacheinander mit demselben Grundtext verglichen werden.[426] Die ermittelten Unterschiede werden in einem Vergleichsprotokoll festgehalten, das die einander zugeordneten, aber nicht gleichen Zeilen aliniert und mit einem Editierskript versieht; bei Löschung, Hinzufügung oder Ersetzung ganzer Zeilen erfolgt eine besondere Kennzeichnung.[427]

Für den Vergleich kann festgelegt werden, dass gewisse Unterschiede ignoriert werden sollen. So lassen sich zum Beispiel Schreibungsvarianten dadurch abfangen, dass Zeichenfolgen und zugehörige Ersetzungen angegeben werden, es können Zeichengruppen als gleichwertig deklariert werden, und es lassen sich Abkürzungszeichen festlegen.[428]

Nach welchem Verfahren die Ermittlung einer optimalen Zuordnung von Textstücken erfolgt, ist aus dem aktuellen Handbuch anscheinend nicht zu entnehmen. Eine kurze Beschreibung von 1977 weist darauf hin, dass Auslassungen oder eine geringe Anzahl von Wortgruppen mit mindestens drei Wörtern, die in beiden Texten übereinstimmten, zu einer Erhöhung des Zeitbedarfs führten.[429]

Die von Peter Robinson ursprünglich für den Vergleich von über 40 Manuskripten zu zwei altnordischen Gedichten entwickelte, inzwischen (zumindest in aller Regel) nicht mehr einsetzbare Programmsammlung *Collate* sollte es ermöglichen, zahlreiche Textfassungen miteinander zu vergleichen und einen kritischen Apparat zu erstellen. Nach der wohl ersten Beschreibung von 1989[430] basierte die Kollationierung auf einem Vergleich auf Wortebene, wobei den handschriftlichen Wortformen aufgrund der hohen Schreibungsvarianz jeweils eine normalisierte Wortform zugeordnet worden war. Die bei diesem Vergleich gefundenen Abweichungen wurden dann noch über eine Prüfung auf Ähnlichkeit in den Zeichenfolgen und auf Synonymie näher klassifiziert.

Das Verfahren, um zu erreichen, dass jeweils die einander entsprechenden Passagen verglichen wurden, basierte dabei anscheinend nicht auf einem der oben in

[425] Ebd. S. 1110.
[426] Ebd. S. 1111.
[427] Ebd. S. 1112.
[428] Ebd. S. 1117 f. und 1129 f.
[429] Reeg 1977.
[430] Robinson 1989. Das ist die konkreteste Darstellung des eingesetzten Verfahrens, die mir vorliegt – wie stark sich dies später noch veränderte, kann ich nicht einschätzen. Das Programm kann aufgrund von Betriebssystem-Änderungen auf aktuellen Systemen nicht mehr genutzt werden, vgl. Robinson 2009, S. 347. Dort werden auch die Leistungen des Programms etwas ausführlicher beschrieben.

Kapitel 2.1 beschriebenen Stringalgorithmen, sondern auf der Versgliederung.[431] Welche Mechanismen zur Zuordnung von Textpassagen beim Vergleich von Fassungen mit umfangreicheren Abweichungen zur Verfügung standen, ist nicht erkennbar (und spielte möglicherweise für die damals untersuchten Versionen keine Rolle). In einer Beschreibung von 2009 weist Robinson darauf hin, dass das Programm zwar hervorragend darin sei, Einzelwortzuordnungen vorzunehmen, und sehr gut in Satzzuordnungen, dass aber die Möglichkeiten zur Erkennung von Umstellungen begrenzt seien.[432]

Für den Vergleich von Textfassungen, die Umstellungen enthalten (konkret für die Vorbereitung einer textgenetischen Edition) wurde das Programm *MEDITE* entwickelt.[433] Wenn als Editieroperation neben Einfügung, Löschung und Ersetzung auch die Umstellung zulässig ist, ist die Ermittlung eines optimalen Editierskripts NP-vollständig,[434] also für nicht sehr kleine Textmengen (nach derzeitigem Erkenntnisstand) praktisch nicht lösbar. In *MEDITE* wird deshalb ein heuristisches Verfahren zugrunde gelegt, um eine Annäherung an eine optimale Lösung zu erreichen. Es basiert darauf, über einen Suffixbaum alle exakten Übereinstimmungen zu ermitteln, die in keinem der beiden verglichenen Texte gänzlich in einer anderen gefundenen Übereinstimmung enthalten sind, und aus diesen in einem rekursiven Verfahren diejenigen auszuwählen, die als nicht umgestellt betrachtet werden. Das Verfahren arbeitet – anders als *Collate* und wohl auch *TUSTEP* – auf Zeichenebene.[435]

Nach einer Beschreibung aus dem Jahr 2011 können zwei Fassungen eines Romans im Umfang von 500 Seiten auch bei zahlreichen Unterschieden in einigen Minuten miteinander verglichen werden.[436] Es ist allerdings zu vermuten, dass dies doch eine insgesamt recht große Nähe der beiden Texte voraussetzt, jedenfalls weist eine frühere Publikation darauf hin, dass der zeitliche Aufwand für die Auswahl der als nicht umgestellt bewerteten Blöcke quadratisch von der Zahl der Übereinstimmungen abhänge,[437] und bei eigenen Testläufen mit dem in die-

[431] Robinson 1989 erörtert das zwar nicht genau, führt aber auf S. 104 als letzte Möglichkeit beim Fehlschlagen der vorherigen Zuordnungsversuche an, dass dann die verbleibenden Wörter des jeweiligen Verses als Ersetzung betrachtet würden.

[432] Robinson 2009, S. 351–353.

[433] Das Programm steht unter anderem in der anscheinend letzten Version 3.8 aus dem Jahr 2009 über die URL *http://www-poleia.lip6.fr/~ganascia/Medite_Project* zum Download zur Verfügung. Kurze Hinweise zur Benutzungsoberfläche finden sich in Bourdaillet/Ganascia 2006, S. 467 f. (mit Abbildung auf S. 461) sowie in Ganascia 2007, S. 7 f. in der Preprint-Version.

[434] Bourdaillet/Ganascia 2007, S. 140 (mit Verweis auf eine Publikation von Dana Shapira und James A. Storer). Vgl. zur NP-Vollständigkeit zum Beispiel Harel/Feldman 2006, S. 201–217.

[435] Vgl. Bourdaillet/Ganascia 2007, S. [1]. Eine technische Beschreibung findet sich in Bourdaillet 2009.

[436] Ganascia 2011, S. [3] (ohne nähere Angaben zum eingesetzten System oder zum konkreten Umfang der Abweichungen zwischen den Versionen).

[437] Bourdaillet/Ganascia 2007, S. 145.

ser Arbeit untersuchten Textmaterial wurden deutlich größere Laufzeiten ermittelt.[438]

Das Programm *Juxta*[439] ermöglicht insbesondere verschiedene Visualisierungen der gefundenen Zuordnungen beziehungsweise Unterschiede. Es können zum Beispiel jeweils zwei Textfassungen nebeneinander angezeigt werden, je nach Wunsch mit oder ohne automatische Parallelisierung der als gleich erkannten Textstücke. Unterschiede werden dabei grün hervorgehoben, und wenn eine in dieser Weise gekennzeichnete Passage mit der Maus berührt wird, ändert sich ihre Farbe ebenso wie die des Gegenstücks in der anderen Fassung. Alternativ kann eine einzige Fassung angezeigt werden, wobei Abweichungen gegenüber anderen Versionen zu einer Blaufärbung führen, die um so dunkler wird, je mehr Fassungen es gibt, die sich an dieser Stelle unterscheiden.

Laut Kommentar im Quellcode[440] beruht der Vergleich auf einem von Paul Heckel 1978 vorgestellten Algorithmus, allerdings wird der Algorithmus mehrfach angewandt, um eine möglichst umfassende Zuordnung von Textstücken erreichen zu können.[441] Gibt es Umstellungen zwischen den verschiedenen Fassungen, lässt sich dies von Hand eingeben.

[438] So benötigte der Vergleich zweier voneinander unabhängiger Übersetzungen der *Institutiones* des *Corpus Iuris Civilis* sowohl bei Zugrundelegung einer nur geringfügig vereinfachten Textform (mit einem Umfang von etwa 770 beziehungsweise 720 kB) als auch auf der Basis einer im Folgenden noch vorzustellenden starken Reduktion der Formenvarianz über eine Codierung (mit einem Umfang von etwa 430 beziehungsweise 410 kB) etwa eine Stunde. (Eine genaue Messung ist aufgrund der graphischen Oberfläche schwierig.) Texte in größerem Umfang wurden nicht verglichen, da schon der Vergleich der Texte mit etwa 770 beziehungsweise 720 kB in einem frühen Verarbeitungsschritt mehr als 1 GB Arbeitsspeicher belegte.

[439] Für die folgende Beschreibung wurde die Desktop-Version 1.7.0 zugrunde gelegt, die in kompilierter Form von *http://www.juxtasoftware.org/download/* und im Quellcode von *https://github.com/performant-software/juxta-desktop* heruntergeladen werden kann. Unter dem Namen *Juxta Commons* (*http://www.juxtacommons.org/*) steht außerdem eine im Vergleich dazu neuere Online-Version zur Verfügung, deren Oberfläche teilweise anders gestaltet ist.

[440] *https://github.com/performant-software/juxta-desktop/archive/master.zip*, darin *juxta-desktop-master/src/main/java/edu/virginia/speclab/diff/DiffAlgorithm.java*.

[441] Vgl. im Quellcode *src/main/java/edu/virginia/speclab/legacy/diff/MultiPassDiff.java*. Heckels Algorithmus (Heckel 1978) basiert auf einer Zuordnung aller Elemente (zum Beispiel Zeilen), die in beiden verglichenen Dateien genau ein Mal vorkommen. Andere Übereinstimmungen werden nur zugeordnet, wenn sie in beiden Dateien unmittelbar vor oder nach einer nach diesem Schema gefundenen Entsprechung stehen. Das ermöglicht einerseits eine wenig aufwendige Zuordnung auch von umgestellten Passagen, führt aber auch leicht zu einer sehr geringen Erkennungsquote, wenn viele der untersuchten Elemente häufiger vorkommen oder aber die einmaligen Vorkommen einander nicht genau entsprechen. Bei einer Verwendung dieses Verfahrens auf einzelne Wörter, wie es sich für den Vergleich von Texten ohne verlässliche Zeileneinteilung nahelegt, wäre eine einfache Anwendung von Heckels Algorithmus wohl wenig erfolgreich, und er setzt jedenfalls eine große Ähnlichkeit der beiden Texte voraus, da sonst über das unmittelbare Umfeld der jeweiligen Hapax legomena keine auch nur annähernd umfassende Abdeckung der tatsächlichen Übereinstimmungen zu erreichen wäre. Das Verfahren basiert inhaltlich auf der Annahme, dass die gefundenen singulären Wortformen tatsächlich aufeinander beziehbar sind, was auf unabhängig entstandene Texte natürlich allenfalls bei thematischer Ähnlichkeit passt.

Juxta lässt das Einlesen von Dateien bis einer Größe von 1 MB zu; allerdings steigt beim Vergleich größerer Dateien (zumindest wenn diese stark voneinander abweichen) die Verarbeitungszeit erheblich. Beim Vergleich kann eingestellt werden, ob Leerraum, Zeichensetzung und/oder Groß-/Kleinschreibung berücksichtigt werden sollen oder nicht. Außerdem werden bei XML-Dateien *Tags* automatisch vom Vergleich ausgeschlossen und es kann festgelegt werden, dass der Inhalt bestimmter XML-Elemente beim Vergleich nicht berücksichtigt wird. Es ist aber nicht möglich, andere Schreibungsvarianten als die eben genannten differenziert zu behandeln, sie führen vielmehr generell zu einer Einstufung der betroffenen Wörter als Ersetzung.

Das Programm *CollateX* war ursprünglich als Nachfolger für *Collate* gedacht, zielt inzwischen aber nicht mehr darauf, eine eigenständige Gesamtlösung für die Erstellung eines kritischen Apparats zu bieten, sondern soll vielmehr die Teilaufgabe des eigentlichen Vergleichs der verschiedenen Fassungen flexibel und in einer Form lösen, die eine Einbettung in andere Systeme ermöglicht.[442] *CollateX* kann (neben einem Web-Service) auf der Kommandozeile aufgerufen werden und bietet verschiedene Möglichkeiten, die Vergleichsergebnisse auszugeben (unter anderem mit TEI-*Tags* für Varianten).

In *CollateX* stehen drei Vergleichsalgorithmen zur Verfügung, nämlich das oben in Unterkapitel 2.1.2 auf S. 61 vorgestellte Verfahren von Needleman und Wunsch zur Ermittlung möglichst großer Ähnlichkeiten, der Algorithmus von *MEDITE* sowie ein Algorithmus von Ronald H. Dekker, der auf die Optimierung lokaler Alinierung von Wortketten und die Erkennung von Umstellungen zielt.[443]

Schließlich ist das Projekt *Semi-automatische Differenzanalyse von komplexen Textvarianten* (*SaDA*)[444] zu nennen, in dem im Zusammenhang mit zwei Editionsprojekten Tools unter anderem für die automatische Kollationierung entwickelt wurden. Eines der beiden Editionsvorhaben hatte mit der *Wundarznei* Heinrichs von Pfalzpaint einen frühneuhochdeutschen Text zum Gegenstand; deshalb spielte das Problem der Formenvarianz eine erhebliche Rolle.

Der Vergleich der in ihrem Textbestand recht erheblich variierenden Textzeugen erfolgte auf der Basis lemmatisierter und in Teilsätze aufgeteilter Texte, indem

[442] Vgl. *http://collatex.net/about/*.

[443] So die Beschreibung in der Dokumentation des Projekts (*http://collatex.net/doc/*). Vgl. Dekker/Middell 2011, S. [4]-[6]. Leider ließ sich in einem Testlauf mit der Version 1.7.1 des Programms schon beim Vergleich zweier Fassungen eines Textes in einem Umfang von ca. 24 kB mit dem *MEDITE*-Algorithmus auch bei Verwendung von 1 GB Arbeitsspeicher keine Ausgabe erzielen; der Vergleich dieser beiden Dateien dauerte mit dem Needleman-Wunsch-Algorithmus wenige Sekunden, mit dem Dekker-Algorithmus ca. 15 Minuten. Ein Vergleich zweier ähnlicher Texte im Umfang von ca. 68 beziehungsweise 106 kB benötigte mit dem Needleman-Wunsch-Algorithmus mehr als 1 GB Arbeitsspeicher und dauerte ca. 5 Minuten.

[444] Eine kurze Projektbeschreibung findet sich unter *http://www.informatik.uni-halle.de/ti/forschung/ehumanities/sada/*.

zum einen einander entsprechende Teilsätze ermittelt wurden und zum anderen die Ähnlichkeit von Absätzen durch Vergleich der enthaltenen Wörter bestimmt wurde.[445]

Was lässt sich nun aus der Betrachtung der verschiedenen Tools zur automatischen Kollationierung für die hier untersuchte Fragestellung entnehmen? Vor allem wohl dies, dass ein Komplettvergleich aller Texte, die auf Gemeinsamkeiten untersucht werden sollen, mit Software, die auf die minutiöse Verzeichnung feinster Unterschiede ausgerichtet ist, kaum zu leisten ist. Die Grundannahme bei diesen Programmen ist stets, dass die verglichenen Texte in ganz erheblichem Maße Übereinstimmungen aufweisen und dass sie überhaupt in einer – direkten oder indirekten – Beziehung zueinander stehen, so dass eine Abweichung auch tatsächlich als Einfügung, Löschung, Ersetzung oder Umstellung interpretiert werden kann. Der Aufwand an Rechenzeit (und je nach Algorithmus auch Speicherplatz), um zu erkennen, welche Textpassagen aufeinander abgebildet werden können, ist bei nichttrivialen Fällen erheblich und kann insbesondere dann noch einmal stark ansteigen, wenn auch Umstellungen erkannt werden sollen.

Dass ein Verfahren, das dies mit einer einigermaßen hohen Erfolgsquote leistet, für den Textvergleich im Rahmen einer kritischen Edition sehr wünschenswert ist und auch eine Verlangsamung (in gewissen Grenzen) rechtfertigt, dürfte unmittelbar einleuchten. Für die hier betrachteten Fälle wird ein derartiger Feinvergleich aber allenfalls dann relevant, wenn in einem Textpaar schon Übereinstimmungen ermittelt wurden, die darauf hindeuten, dass nicht nur einzelne Passagen einander zuzuordnen sind, sondern die Texte insgesamt in ein Abhängigkeitsverhältnis zueinander oder gemeinsam zu einem früheren Text eingeordnet werden können.

Neben einer Zuordnung über die Ermittlung einer längsten gemeinsamen Teilsequenz kommen unter diesen Bedingungen wohl insbesondere zum einen der Algorithmus von Paul Heckel in der für *Juxta* vorgenommenen Modifikation in Betracht, zum anderen der des *MEDITE*-Projekts, da mit beiden allem Anschein nach auch ziemlich umfangreiche Textfassungen mit einem Aufwand verglichen werden können, der die Anwendung auch auf eine etwas größere Zahl von Fällen ermöglicht.

[445] Vgl. Medek u. a. 2015 sowie (vor allem zum Text und zu den Handschriften) Leipold/Ritter/Solms 2014. Im Rahmen des Projekts wurde das Tool *CATView* entwickelt, das zur Visualisierung der Ähnlichkeiten und Unterschiede zwischen zuvor über eine Alinierung einander zugeordneten Textstücken dient (vgl. *http://catview.uzi.uni-halle.de/*). Eine Arbeitsumgebung, über die solche Alinierungen erstellt werden können, ist anscheinend zur Zeit noch in Bearbeitung (vgl. *http://www.informatik.uni-halle.de/ti/forschung/ehumanities/sada/pub/*).

3 Technische Beschreibung

Teil 3 stellt die für diese Untersuchung entwickelten Verfahrensschritte und Auswertungsansätze vor, um wörtliche Übereinstimmungen in frühneuhochdeutschen Texten zu ermitteln, hinsichtlich ihrer Aussagekraft einzuschätzen und daraus Informationen über textuelle Beziehungen abzuleiten. Kapitel 3.1 beschreibt, wie sich die Texte so aufbereiten lassen, dass die Schreibungsunterschiede weitgehend nivelliert werden, und erläutert, wie sich eine Rückbindung der auf dieser Basis zu ermittelnden Übereinstimmungen an die entsprechenden Textstücke im Original und gegebenenfalls an ihnen zugeordnete Angaben zur kanonischen Zitierweise (oder sonstige Informationen zur Stellenidentifikation) erreichen lässt. Kapitel 3.2 behandelt die Ermittlung von *maximal exact matches (MEMs)* in Texten und bietet verschiedene quantitative Analysen der für das hier ausgewertete Korpus ermittelten MEMs. Kapitel 3.3 geht auf Möglichkeiten ein, die Qualität der Matchdaten durch nachträgliche Bearbeitung beziehungsweise Auswahl auf der Basis einfacher Kriterien zu verbessern. Kapitel 3.4 schließlich zeigt verschiedene Möglichkeiten, wie sich die Matchdaten nutzen lassen, um einen Überblick über Textbeziehungen und Gruppenbildungen innerhalb des Korpus zu gewinnen und um Textpaare im Detail miteinander zu vergleichen.

3.1 Vorbereitende Texttransformationen

Wie bereits in Unterkapitel 2.1.2 dargestellt wurde, steigt der Aufwand für einen Vergleich, der Ersetzungen, Einfügungen und Löschungen in bestimmter Art und/oder bis zu einem bestimmten Maße zulässt, umso stärker, je mehr Zeichen (oder Wörter oder sonstige als Einheiten behandelte, nicht näher untersuchte Textfragmente) in dieser Weise verglichen werden,[446] so dass sich schnell praktische Probleme ergeben können, wenn ein solches Vergleichsverfahren für ganze Texte oder gar Textkorpora eingesetzt wird.

Es legt sich deshalb nahe, zunächst einmal nach exakten Übereinstimmungen zu suchen, die bestimmten Auswahlkriterien genügen, und auf dieser Basis gegebenenfalls einen detaillierteren Vergleich der ermittelten Entsprechungen und/oder ihres Textumfelds anzuschließen. Um trotzdem auch möglichst viele Entsprechungen ermitteln zu können, die nicht ganz exakt sind, sondern kleinere Abweichungen einschließen, bietet es sich an, die Texte so zu transformieren, dass sie möglichst wenige als irrelevant oder wenig wichtig betrachtete Unterschiede enthalten. Solche Vorverarbeitungsschritte sind, wie oben in den Kapiteln 2.4 und 2.5 beschrieben wurde, sowohl in der Plagiatserkennung als auch in der *Text-Reuse*-Ermittlung allgemein üblich.

[446] Genauer gesagt steigt der Aufwand (soweit Einzelvergleiche jeweils gleich zu Buche schlagen) entsprechend dem Produkt der Zahlen der als Einheiten behandelten Textfragmente in den beiden verglichenen Texten.

Bei der Datentransformation ist insbesondere zwischen zwei Schritten zu unterscheiden, nämlich zum einen der Textextraktion und zum anderen der Reduzierung von Textvarianz.

3.1.1 Textextraktion

Zunächst einmal ist davon auszugehen, dass die Dateien nicht nur die reinen Zeichenfolgen der zu untersuchenden Texte enthalten, sondern mit zusätzlichen Informationen angereichert sind, die die Formatierung beschreiben oder auch Strukturangaben, inhaltliche Klassifikationen oder Ähnliches enthalten können. Grundannahme soll hier – entsprechend dem in langfristig orientierten Editionsprojekten etablierten Standard – sein, dass es sich um XML-Dateien handelt, also um ein Datenformat, in dem sowohl der Text als auch die Zusatzinformationen mit einem weitgehend beliebigen Texteditor[447] in einer lesbaren Form angezeigt werden. „Lesbar“ bedeutet dabei nicht unbedingt, dass eine flüssige Lektüre des hier relevanten eigentlichen Textes (nach der Terminologie der XML-Spezifikation[448] die *character data* des Dokuments, allerdings einschließlich der Zeichen- und Entitätsreferenzen beziehungsweise der ihnen zugeordneten Zeichenfolgen, soweit diese nicht selbst zum *Markup* gehören) möglich ist – das hängt von den Darstellungsmöglichkeiten im verwendeten Texteditor und bei einer vollständigen Anzeige insbesondere auch vom Umfang der im *Markup* verzeichneten Zusatzinformationen ab –, sondern dass es keine nicht anzeigbaren Steuerzeichen gibt und dass sich nach einfachen Regeln eine Abgrenzung des *Markups* vornehmen lässt.

Die Bezeichnung *eigentlicher Text* ist hier mit der Vorstellung gewählt, dass XML zur Codierung von Texten eingesetzt wird und dass genau die Zeichenfolge des jeweiligen Textes übrigbleibt, wenn man das *Markup* bis auf Entitäten entfernt und Entitäten nach dem vorgesehenen Verfahren umwandelt. Das ist von den XML-Regeln her nicht zwingend – es lassen sich ohne Mühe andere Anwendungssituationen finden und auch Argumente dafür anführen, bei der XML-Repräsentation von Texten davon abzuweichen[449] –, aber es ist eine naheliegende Verfahrensweise und trifft jedenfalls auf das hier zugrunde gelegte Textkorpus zu.[450]

[447] Voraussetzung ist allerdings, dass die in der Datei verwendete Codierung unterstützt wird. Da dafür aber nur wenige Standards etabliert sind – neben ASCII und den ISO-8859-Erweiterungen zu ASCII sind dies vor allem *Unicode*-Codierungen wie UTF-8 –, stellt das auf einigermaßen aktuellen Systemen kein Problem dar.

[448] Vgl. in der aktuellen Spezifikation Abschnitt 2.4 (*http://www.w3.org/TR/2006/REC-xml11-20060816/#syntax*).

[449] Für die Codierung stark strukturierter Informationen kann es sinnvoll sein, überhaupt nur *Markup* zu verwenden – in dieser Weise wird XML zum Beispiel in *XML Schema* eingesetzt, um die formalen Regeln für bestimmte XML-Dokumente festzuhalten. Bei der Codierung von Texten kann es unterschiedliche Sichtweisen geben, was überhaupt als eigentlicher Text zu betrachten ist, so zum Beispiel bei editorischen Verbesserungen. Die TEI sieht hierfür verschiedene Codierungsvarianten vor – neben der Kennzeichnung des betreffenden Textstücks in einer Version mit einem *Tag*, dem die andere Version als Attribut zugeordnet ist, kann auch das Element *choice*

In einem ersten Schritt ist also der eigentliche Text aus den XML-Dateien zu extrahieren. Das lässt sich mit einfachen Parsing- oder auch Stringverarbeitungsfunktionen leicht bewerkstelligen, dabei gehen aber ohne weitere Maßnahmen alle Informationen verloren, die im XML-*Markup* enthalten sind. Dann lässt sich später nicht ohne eine gegebenenfalls relativ aufwendige Suche[451] die Position, an der in der XML-Datei beziehungsweise in der ihr entsprechenden Datenstruktur eine bestimmte Textstelle zu finden ist, und damit auch die Zuordnung zum Beispiel zu einem bestimmten Absatz feststellen.

Dies ist aber für viele Anwendungssituationen eine wichtige Information. Insbesondere kann dabei auch von Interesse sein, Textstellen so zu identifizieren, dass die Angaben auch nach Veränderungen in einer Datei – etwa bei der Korrektur von Fehlern oder der Einfügung von Kommentaren – mit zumindest einigermaßen hoher Wahrscheinlichkeit ihre Gültigkeit behalten und vielleicht sogar einer kanonischen Referenz für den jeweiligen Abschnitt entsprechen, also für entsprechend fachkundige Leser auch ohne weitere Erläuterungen verständlich sind.

Für eine relativ stabile Adressierung von Teilen eines XML-Dokuments bietet sich die Verwendung von *XPath*-Ausdrücken an. Damit lässt sich unter anderem ein konkretes XML-Element innerhalb eines Dokuments dadurch bezeichnen, dass vom Wurzelelement des Dokuments (also dem alle übrigen Elemente umschließenden Element) aus die ineinander geschachtelten Elemente benannt werden (gegebenenfalls mit Angabe, um das wievielte entsprechende Element es sich handelt), deren letztes (beziehungsweise in der Schachtelung innerstes) das Element ist, um das es geht. Eine alternative Adressierungsmöglichkeit besteht, wenn das betreffende Element ein Attribut hat, das eine eindeutige Identifikation ermöglicht. Die XML-Spezifikation sieht einen speziellen Datentyp für solche Identifikatoren vor, dabei gibt es allerdings relativ enge Vorschriften, welche Zeichenfolgen zulässig

verwendet werden, dem die Alternativen gleichrangig als Elemente untergeordnet sind, vgl. *http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-choice.html*.

[450] Sahle 2013, Bd. 3, S. 236–239 stellt Äußerungen in der Literatur zusammen, die voraussetzen, dass sich in dieser Weise eine Trennung zwischen Text und *Markup* vornehmen lässt. Die ebd. S. 239–243 erörterten Probleme spielen für die vorliegende Untersuchung nur teilweise eine Rolle, da hier eben eine Textauszeichnung vorausgesetzt wird, die auf einer einzigen Textfassung aufsetzt und für alle untersuchten Texte auf dem gleichen Zeichensatz basiert. Allerdings ist bei der Auswertung nur der Inhalt des *text*-Elements (nach dem Auszeichnungsschema der TEI) zu berücksichtigen, und in den Haupttext eingeschobene Stücke wie Marginalien und Fußnoten sind für das hier angewandte Verfahren zur Erkennung von Textübereinstimmungen sinnvollerweise zu eliminieren beziehungsweise so umzustellen, dass sie wortgleiche Passagen möglichst nicht unterbrechen.

[451] Eine Suche mit einem Skript oder in einem Programm, das keine spezielle Repräsentation für die *character data* eines XML-Dokuments bietet, lässt sich nur über einen regulären Ausdruck bewerkstelligen, weil zwischen allen Zeichen eines gesuchten Textstücks noch XML-*Tags*, XML-Kommentare oder Ähnliches stehen können. Steht eine Repräsentation der *character data* zur Verfügung, kann – wenn das Programm eine entsprechende Funktion hat –, natürlich darin gesucht werden. Jedenfalls ist dafür aber ein eigener Suchschritt erforderlich.

sind. Man kann freilich auch ein Attribut mit einem anderen Datentyp in diesem Sinne verwenden, muss dann aber ohne Unterstützung durch einen XML-Parser dafür sorgen, dass die Attributwerte jeweils eindeutig sind. In dieser Weise wird im DRQEdit-Korpus das Attribut *n* verwendet, das nach der TEI für die Abbildung eines Referenzsystems genutzt werden kann.[452] Damit ist für alle Texte, die schon im Original eine systematische Zählung haben, eine Adressierung über diese Zählung möglich.

Wenn der eigentliche Text aus einem XML-Dokument extrahiert und dabei zugleich die Zuordnung von Textteilen zu Elementen protokolliert wird, stellt sich die Frage, bis zu welcher Auszeichnungstiefe diese Protokollierung erfolgen sollte. Für die vorliegende Untersuchung wird angenommen, dass es ausreicht, wenn die herausgeschriebenen Positionen jeweils Textstücken entsprechen, die nahtlos aneinander anschließen. Das heißt, dass die Positionen von übergeordneten Elementen nicht verzeichnet werden (sie ergeben sich allerdings aus den Positionen der enthaltenen Elemente) und dass insbesondere auch darauf verzichtet wird, die Zuordnung zu XML-Elementen zu dokumentieren, die eine besondere Kennzeichnung innerhalb von gemischtem Inhalt (*mixed content*[453]) darstellen.

Nach diesem Ansatz ist es möglich, zum Beispiel eine Untergliederung in Überschriften und Absätze zu protokollieren oder auch eine in Sätze, in Wörter oder sogar Buchstaben, soweit diese Strukturen im jeweiligen Textstück durchgängig ausgezeichnet sind. Hingegen werden zusätzliche Markierungen wie zum Beispiel von Personennamen oder von editorischen Hinweisen ignoriert.

Auf diese Weise ist gewährleistet, dass einer Textposition stets genau ein *XPath*-Ausdruck zugeordnet werden kann, und bei entsprechender Auszeichnung des gesamten Korpus, dass dabei jeweils eine gleiche oder ähnliche Elementart zugrunde gelegt wird. Dementsprechend kann bei der Ermittlung des zu einer Position passenden *XPath*-Ausdrucks problemlos mit einem binären Suchverfahren gearbeitet werden. Der zu durchsuchende Bereich lässt sich also in mehreren Schritten jeweils in zwei in etwa gleich große Hälften aufteilen, wobei anhand des Werts in der Mitte entschieden wird, ob anschließend in der Hälfte mit den größeren Werten oder in der mit den kleineren weitergesucht wird.[454]

---

[452] Vgl. *http://www.tei-c.org/release/doc/tei-p5-doc/de/html/CO.html#CORS*. Das Attribut kann nach dieser Beschreibung zur Bezeichnung struktureller Einheiten genutzt werden, wobei als Bezeichnung alternativ allerdings auch ein Wert gewählt werden kann, der nur die Identifikation innerhalb der jeweils übergeordneten Einheit ermöglicht. „n“ steht primär für „number“, kann aber ausdrücklich auch ein anderes „label“ für ein Element sein (*http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ST.html#STGA*).

[453] Vgl. die aktuelle XML-Spezifikation, Abschnitt 3.2.2 (*https://www.w3.org/TR/2006/REC-xml11-20060816/#sec-mixed-content*).

[454] Vgl. zur binären Suche zum Beispiel Ottmann/Widmayer 1996, S. 154–156 oder – mit der Implementierung einer binären Wortsuche in *Perl* – Orwant/Hietaniemi/Macdonald 2000, S. 2 f.

3.1.2 Reduzierung von Textvarianz

Auch in Texten aus der heutigen Zeit ist bei wörtlich übernommenen Textstücken – zumindest dann, wenn sie nicht per *Copy and paste* übernommen, sondern abgeschrieben werden – mit kleineren Abweichungen zu rechnen, da es zum Beispiel in der Zeichensetzung und in der Groß- und Kleinschreibung manchmal mehrere nach der Orthographie zulässige Varianten oder auch etwa durch die Einführung der sogenannten neuen Rechtschreibung eine orthographische Entwicklung gibt und natürlich auch Abweichungen von der Normschreibung vorkommen. Dementsprechend ist eine Vereinheitlichung in diesen Bereichen zum Beispiel in der Plagiatserkennung und in der Untersuchung von *Text Reuse* verbreitet.[455]

Eine möglichst weitgehende Erkennung von Übernahmen in frühneuhochdeutschen Texten erfordert aber wesentlich umfassendere Umformungen, um die Vielzahl möglicher Schreibungen (und bei dialektaler Varianz auch Lautungen) eines Worts zumindest zum größeren Teil auf eine gemeinsame Form abzubilden.[456] Wie bereits in Unterkapitel 2.3.4 dargestellt wurde, gibt es in der Erkennung von Namensvarianten sehr ähnlich gelagerte Probleme, so dass es sich anbietet, analog zu den dort beschriebenen Codierungen eine Transformation von frühneuhochdeutschen Schreibungen in eine Form durchzuführen, die möglichst adäquat Varianten mit ähnlichen, gegebenenfalls austauschbaren Lauten zusammenfasst.

Ein Umwandlungsregelwerk, das auf alle dem Frühneuhochdeutschen zuzuordnenden Texte anwendbar ist, lässt sich allerdings wohl nicht – oder allenfalls unter Einbeziehung hochkomplexer Fallunterscheidungen, die letztlich auf eine Differenzierung nach Schreibsprachen oder Ähnlichem hinauslaufen – so konstruieren, dass tatsächlich fast alle einander entsprechenden Varianten zusammengefasst werden, ohne zugleich signifikante Unterschiede zu nivellieren. Es lässt sich hier das häufige Problem bei Kategorisierungen feststellen, dass eine Steigerung der *Precision* leicht eine Verringerung des *Recalls* nach sich zieht und umgekehrt.

Um das denkbare Spektrum aufzuzeigen: Wenn man davon ausgeht, dass bei der Umwandlung keine aus dem jeweiligen Text gewonnenen Erkenntnisse über die darin festzustellenden Schreibgewohnheiten einbezogen werden können, wäre eine maximale *Precision* dadurch zu erreichen, dass überhaupt keine Veränderung am jeweiligen Buchstabenbestand (abgesehen von einer Zusammenfassung von Groß- und Kleinschreibung) vorgenommen wird. Auch damit würde allerdings eine falsche Zusammenfassung von Formen wie *grauen* vorgenommen, bei denen die korrekte Lesung von den jeweiligen Schreibgewohnheiten abhängt (im angeführten Beispiel wäre die Interpretation als *Grafen* oder als *grauen* beziehungsweise *Grauen* nach der heutigen Orthographie möglich). Der dabei erreichte *Recall* wäre allerdings

[455] Vgl. oben S. 85, Anm. 341 und S. 93. Das entspricht auch dem allgemein üblichen Verfahren in Retrievalsystemen, die eine Volltextsuche anbieten und dafür das Textmaterial wortweise indizieren.

[456] Vgl. oben Kapitel 1.3.

relativ gering.[457] Umgekehrt wäre der *Recall* theoretisch dadurch zu maximieren, dass alle Wortformen auf einen einzigen Code abgebildet würden – die *Precision* wäre in diesem Fall allerdings extrem niedrig.

Ziel der hier beschriebenen Textumformung braucht es allerdings nicht zu sein, eine korrekte Charakterisierung der den Buchstaben zuzuordnenden Lautfolgen vorzunehmen oder gar die Wortformen so zu repräsentieren, dass sich die zugrunde liegenden Lexeme jeweils zweifelsfrei rekonstruieren lassen. Es geht nicht um eine Kategorisierung, die tatsächlichen Verhältnissen beziehungsweise ihrer sprachwissenschaftlichen Darstellung entspricht, sondern darum, eine Abbildung vorzunehmen, bei der möglichst viele der durch die Varianz in Schreibung und Lautung bedingten Unterschiede entfallen, verschiedenen Lexemen (beziehungsweise flektierten Wortformen) zuzuordnende Schreibformen aber möglichst nicht zusammengefasst werden. Dass dabei auch recht erhebliche Unschärfen tolerabel sind, wird unten ab S. 125 dargelegt.

Während oben in Kapitel 1.3 die Entwicklungen in Lautung und Schreibung beziehungsweise die entsprechenden Abweichungen vom heutigen Standarddeutsch zusammengefasst wurden, soll im Folgenden zur Vorbereitung der Entwicklung von Transformationsregeln in alphabetischer Reihenfolge ein Überblick gegeben werden, mit welchen Austauschverhältnissen zwischen Buchstaben beziehungsweise Buchstabengruppen zu rechnen ist, wobei die Erläuterungen von den in den Texten zu findenden Buchstaben (also nicht von den ihnen zuzuordnenden Lauten und nicht von den neuhochdeutschen Schreibungen) ausgehen und jeweils Entsprechungen in der modernen Orthographie nennen, da sich darin allgemein Schreibungen durchgesetzt haben, die auch im Frühneuhochdeutschen schon verbreitet sind.[458] Die Vielzahl der im Frühneuhochdeutschen möglichen Buchstabenfolgen wird in dieser Zusammenstellung allerdings weitgehend ausgeklammert, und es wird keine Unterscheidung zwischen Lang- und Kurzvokalen vorgenommen.[459] Dass alle Grundbuchstaben (nicht allerdings die meisten Kombinationen mit Diakritika) in gleicher Verwendung auch in neuhochdeutschen Schreibungen zu finden sind, wird

[457] Hier einen konkreten Wert zu ermitteln, wäre nur begrenzt aussagekräftig, da er stark von den jeweils zugrunde gelegten Texten abhängen würde und sich schlecht allgemein sagen lässt, wie zum Beispiel die verschiedenen Schreibsprachen angemessen gewichtet werden sollten. Zudem wäre es für die Ermittlung erforderlich, schon zu wissen, was tatsächlich als bis auf Schreib- und Lautvarianten gleich zu betrachten ist. Vgl. zur Illustration aber oben S. 77, Anm. 313.

[458] Diese Zusammenstellung beansprucht nicht, die in der zugrunde gelegten Grammatik (Reichmann/Wegera 1993) beschriebenen Phänomene auch nur annähernd vollständig zu berücksichtigen, was angesichts der Komplexität und des Wechsels der Beschreibungsebene von den Lauten zu den Buchstaben nur schwer und wohl nicht ohne umfangreiche weiterführende Lektüre oder entsprechende eigene Untersuchungen zu leisten wäre. Insbesondere werden auch die ebd. genannten Abhängigkeiten von Nachbarlauten sowie von Zeit, Raum und Schreibtradition und die Hinweise zur Häufigkeit hier nur in Einzelfällen und verkürzt wiedergegeben.

[459] Vgl. ebd. S. 38–49 zu den Kurz- und 49–57 zu den Langvokalen.

in dieser Auflistung als selbstverständlich vorausgesetzt und nicht weiter erwähnt. Soweit das sinnvoll erscheint, dient /.../ zur Kennzeichnung, dass der betreffende Laut bezeichnet werden soll, und <...> zur Kennzeichnung, dass es sich um eine Schreibung handelt.

- *a* kann – abhängig von der jeweiligen Schreibsprache – auch einem neuhochdeutschen *o*, *e* oder *ä* (auch als Bestandteil eines Diphthongs) oder auch selten ohne weiteren Zusatz einem Diphthong entsprechen.[460]
- *ä*/*å*/*æ* kann auch einem neuhochdeutschen *e* und in einigen Fällen, zum Beispiel vor */sch/* oder bei Verwendung des übergeschriebenen *e* als Dehnungszeichen, einem *a* entsprechen.[461]
- *ai* und *ae* können vor allem in westmitteldeutschen Texten als langes *a* zu lesen sein.[462] *ai* und *ay* werden aber auch zum Teil recht häufig anstelle eines neuhochdeutschen *ei* gebraucht.[463]
- *au*, *aw* (und vielleicht auch *auw*) können in einigen Regionen auch einem neuhochdeutschen *a*, *o* oder *eu* zuzuordnen sein.[464]
- *b* kann in einigen Wörtern einem neuhochdeutschen *p* entsprechen,[465] außerdem infolge von Hyperkorrektur einem neuhochdeutschen *w* oder *f*. Die Zeichenfolge *mb* ist eine häufige Schreibvariante zu *m*.[466]
- Ein allein stehendes *c* vor einem dunklen Vokal (also insbesondere *a*, *o* und *u*) oder einem Liquid (*l* oder *r*) entspricht einem neuhochdeutschen *k*, ein allein stehendes *c* vor einem hellen Vokal (insbesondere *e* und *i*) einem neuhochdeutschen *z*. Wie in der heutigen Orthographie kann ein *c* insbesondere in die Buchstabenfolgen *ch*, *ck* und *sch* eingebunden sein, daneben auch in *cz* (wie ein neuhochdeutsches *z* oder *tz* zu lesen) und in verschiedene Erweiterungen und Varianten der genannten Buchstabenfolgen (zum Beispiel durch Konsonantenverdopplungen, aber auch *sc* sowie die Zeichenfolgen *chs*, *chß* und *chss*, die teilweise als *sch* zu lesen sind). Die Verwendung eines *c* ohne weiteren Zusatz als Entsprechung zu einem neuhochdeutschen *ch* ist selten und läuft im 14. Jahrhundert aus.[467]

[460] Vgl. ebd. S. 38 f. (zu *a* anstelle von *o* oder *e*), 49 f. (zu *a* anstelle eines langen *e* oder *ä*, eines *ei* oder *au*) und 60 (zu *au* und ähnlichen Schreibungen anstelle von *eu* beziehungsweise *äu*).

[461] Vgl. ebd. S. 40–42 (zu den verschiedenen Schreibungen für den kurzen *e*-Laut sowie zum Gebrauch anstelle von *a*), 49 (zum übergeschriebenen *e* als Dehnungszeichen) und 52 f. (zum Langvokal *ä*).

[462] Vgl. ebd. S. 33. Die nach der dortigen Beschreibung zu vermutende Schreibung *ay* wird ebd. S. 49 nicht aufgeführt.

[463] Vgl. ebd. S. 58.

[464] Vgl. ebd. S. 59 f. Die Schreibung *auw* für den Diphthong wird dort nicht explizit aufgeführt (vgl. unten Anm. 488 zu *ouw*), ist im DRQEdit-Korpus allerdings recht häufig belegt. Ob sie auch in Fällen zu finden ist, in denen die hier beschriebene Lautänderung vorliegt, ist damit natürlich noch nicht sichergestellt.

[465] Entsprechende Beispiele finden sich in Reichmann/Wegera 1993, S. 90. Da der Gebrauch von *p* in der gehobenen Schreibtradition und auch in der heutigen Standardsprache gewissen Beschränkungen unterliegt (vgl. ebd. S. 87 f.), betrifft das wohl nur Wörter, bei denen ein ursprüngliches *b* im Anlaut durch ein *p* ersetzt wurde.

[466] Vgl. ebd. S. 105 (zu *b* anstelle von *w*) und 134 f. (zum Gebrauch von *b* nach *m*).

- *ch* kann anstelle eines *g*, eines *h* und (vor allem in älteren Texten) anstelle eines *k* stehen; im Suffix *icheit* beziehungsweise *echeit* entspricht es der Buchstabenfolge *gk*. Vor einem *t* kann es einem *f* entsprechen. Vor allem nach *i* kann es seit dem 15. Jahrhundert anstelle von *sch* stehen.[468]
- *d* kann auch einem neuhochdeutschen *t* entsprechen und vor allem nach einem *n* im Neuhochdeutschen ohne Entsprechung sein. Selten steht es anstelle eines *g* oder ist zwischenvokalisch eingeschoben.[469]
- *e* kann auch einem neuhochdeutschen *i*, *ie* oder *ä* und in einigen Fällen, zum Beispiel vor */sch/*, einem *a* entsprechen. Vor allem im Westmitteldeutschen kann es zur Bezeichnung der Dehnung eines vorangehenden Vokals dienen. Als Sprossvokal oder in Epithese kann es ohne Entsprechung im Neuhochdeutschen sein.[470]
- *ei* und *ey* können aufgrund von Monophthongierung oder bei Verwendung von *i* als Dehnungszeichen einem neuhochdeutschen *e* entsprechen.[471] Außerdem wird der Diphthong im Neuhochdeutschen in einigen Wörtern als *ai* geschrieben.[472]
- *eu*, *eü*, *eů* und ähnliche Schreibungen mit einem *w* herrschen zunächst – bis auf den bairischen Sprachraum – auch bei Wörtern vor, bei denen sich später aufgrund einer morphologischen Verwandtschaft zu Wörtern mit *au* die Schreibung mit *äu* durchsetzt.[473]
- *f* kann an Stellen, an denen eine Auslautverhärtung eintritt, in Texten aus dem mittelfränkischen Raum einem standardsprachlichen *b* (im Auslaut */p/*) entsprechen. Selten entspricht ein initiales *f* (so wie in der ostmitteldeutschen Lautung) einem hochsprachlichen *pf*.[474]
- *g* kann zwischenvokalisch in der neuhochdeutschen Orthographie ohne Entsprechung oder durch *h* ersetzt sein. In einigen Regionen kann es anstelle eines *k*, in einigen anstelle eines *j* oder auch eines *ch* stehen.[475]
- *h* kann als *ch* zu lesen sein. Zwischenvokalisch, am Wortanfang und neben vielen

[467] Vgl. ebd. S. 101 (zum Gebrauch anstelle von *k* und zum Teil in Kombination damit), 116 f. (zu *sch*), 122 (zu *ch*) und 132 (zum Gebrauch anstelle von *z*, ebenfalls zum Teil in Kombination damit).

[468] Vgl. ebd. S. 101 (zu *ch* als Schreibung für *k* und zum Suffix *icheit*), 118 (zur Schreibung von *ch* anstelle von *sch*), 122 f. (zum Gebrauch anstelle von *g*, zum Suffix *echeit*, zur Ersetzung von *ft* durch *cht*, zur Ersetzung von nachliquidischem *ch* durch *k* und zur Schreibung von *ch* anstelle von *sch*) und 125 (zur Verwendung statt *h*).

[469] Vgl. ebd. S. 90–92 (zum Verhältnis von *d* und *t*, zum eingeschobenen *d* und zur Verwendung anstelle von *g*) sowie 93 (zum Verhältnis von *d* und *t*).

[470] Vgl. ebd. S. 33 (zu *e* als Längenzeichen), 42 (zu *e* anstelle von *a*), 69 f. (zu *e* anstelle von *ie* und *i*) und 82 f. (zu Sprossvokal und Epithese).

[471] Vgl. ebd. S. 50 f.

[472] Schreibungen mit *a* sind im Frühneuhochdeutschen typisch für oberdeutsche Texte, vgl. ebd. S. 58.

[473] Vgl. ebd. S. 61.

[474] Vgl. ebd. S. 110.

[475] Vgl. ebd. S. 99 f. (zur zwischenvokalischen Verwendung und zum Gebrauch anstelle von *k*), 120 f. (zu *g* anstelle von *j*), 123 (zu *g* anstelle von *ch*) und 125 f. (zur zwischenvokalischen Verwendung anstelle von *h*).

anderen Konsonanten kann es in der heutigen Schreibung ohne Entsprechung sein.[476]

- *i* kann auch so wie ein neuhochdeutsches *j* verwendet werden. Vor allem im Westmitteldeutschen kann es zur Kennzeichnung der Länge des vorangehenden Vokals dienen. Aufgrund von Entrundung beziehungsweise Rundung kann es (beziehungsweise *ie*) einem neuhochdeutschen *ü*, bei noch nicht erfolgter oder nicht in der Schreibung abgebildeter neuhochdeutscher Diphthongierung einem *ei* und unter anderem als Nebensilbenvokal einem *e* entsprechen. Bei einigen Wörtern steht es anstelle eines neuhochdeutschen *u*. In dialektal geprägten Texten verschiedener Regionen kann ihm ein standardsprachliches *g* zuzuordnen sein.[477] Vermutlich kann es zwischenvokalisch auch einem *h* entsprechen.[478]
- *j* kann auch zur Bezeichnung des Vokals *i* dienen und wie *i* unter dialektalem Einfluss einem *g* der heutigen Schriftsprache entsprechen.[479]
- *k* kann in einigen Wörtern anstelle eines *g* stehen.[480]
- *l* kann aufgrund von Dissimilation einem *n* oder *r* entsprechen. Vereinzelt kann es ohne Entsprechung sein.[481]
- *m* kann auch einem neuhochdeutschen *w*, *nt*, *nd* oder *n* entsprechen. Anstelle des Suffixes *ung* kann *um* (oder *umb*) stehen, anstelle des Wortendes *ben m* oder *me*.[482] Abweichungen zwischen *m* und *n* können sich auch daraus ergeben, dass beide Konsonanten am Silbenende häufig durch den Nasalstrich vertreten werden und sich bei einer Transkription nicht immer sicher entscheiden lässt, welcher Buchstabe intendiert ist.[483]
- *n* kann anstelle eines *m* oder vereinzelt auch eines *ng* stehen, und es kann im Neuhochdeutschen ohne Entsprechung sein. *nn* kann auch einem *nd* entsprechen.[484]
- *o* kann einem neuhochdeutschen *a*, bei fehlender Kennzeichnung des Umlauts einem *ö* und aufgrund von Senkung einem *u* entsprechen. In einigen Fällen kann ihm ein *ü* zuzuordnen sein. In einigen Regionen kann es anstelle eines heutigen

[476] Vgl. ebd. S. 122 (zum Gebrauch anstelle von *ch*) und 125 f. (zu Schreibungen mit *h* ohne Entsprechung im Neuhochdeutschen) sowie passim.

[477] Vgl. ebd. S. 33 (zu *i* als Dehnungszeichen), 44 (zu Wörtern mit *i* anstelle von *u*), 64 f. (zur neuhochdeutschen Diphthongierung), 75–77 (zu Entrundung und Rundung), 78 (zur Verwendung als Nebensilbenvokal), 119 (zum Verhältnis von *i* und *j*), sowie 120 f. (zur Verwendung anstelle von *g*).

[478] Ebd. S. 119 werden *säjen* und *mäjen* genannt, allerdings heißt es ebd. auf S. 43 in Bezug auf das Verhältnis zu *i*, die Schreibung mit *j* sei medial „sehr selten vertreten".

[479] Vgl. ebd. S. 43 und 53 (zum Verhältnis von *i* und *j*) sowie 120 f. (zur Verwendung anstelle von *g*).

[480] Vgl. ebd. S. 103 f.

[481] Vgl. ebd. S. 148.

[482] Vgl. ebd. S. 136.

[483] Schwierigkeiten treten insbesondere auf, wenn ein Flexionsmorphem betroffen ist und nicht nur morphologisch, sondern auch syntaktisch verschiedene Auflösungen denkbar scheinen.

[484] Vgl. Reichmann/Wegera 1993, S. 138–140.

au stehen. Als Nebensilbenvokal in einem Präfix kann es äquivalent zu einem heutigen *e* sein.[485]

- *ö/o̊* kann aufgrund von Senkung einem neuhochdeutschen *ü* entsprechen. Es kann in oberdeutschen Texten vor *ei* auf einem durch diese Position bedingten Umlautprozess beruhen, dann ist dieser Schreibung ein standardsprachliches *o* zuzuordnen. In älteren Handschriften des alemannischen Raums kann es einem neuhochdeutschen *eu* beziehungsweise *äu* entsprechen.[486]
- *oi*, *oy* und *oe* können vor allem in westmitteldeutschen Texten als langes *o* zu lesen sein.[487]
- *ou*, *ow* und *ouw* entsprechen, soweit es sich um einen Diphthong handelt, meist einem neuhochdeutschen *au*, in einigen Regionen können sie (wie *au*) aber auch einem neuhochdeutschen *a* oder *o* zuzuordnen sein.[488]
- *p* wird in Texten, die von dialektaler Lautung beeinflusst sind, häufig anstelle eines initialen *b* oder eines *pf* nach der Hochlautung verwendet, nach *l* oder *r* kann es in älteren mittelfränkischen Texten auch einem neuhochdeutschen *f* entsprechen. *mp* ist wie *mb* zum Teil als Schreibvariante zu *m* zu lesen.[489]
- *pf* kann aufgrund von Hyperkorrektur einem neuhochdeutschen *f* beziehungsweise *v* entsprechen.[490]
- *ph* kann auch für *pf* stehen.[491]
- *r* kann in einer Metathese die Position mit dem benachbarten Vokal tauschen, und es kann selten im Neuhochdeutschen ohne Entsprechung sein.[492]
- *s* kann auch einem neuhochdeutschen *z* entsprechen. Vor *l*, *m*, *n* und *w* steht es in älteren Texten zum Teil noch anstelle des späteren *sch*.[493]
- *sch* kann insbesondere nach einem *r* beziehungsweise vor einem *t* einem neuhochdeutschen *s* entsprechen.[494]
- *ss/ß* kann auch einem neuhochdeutschen *s* (mit stimmhaftem *s* in der Standardsprache) oder durch Assimilation auch <*chs*> (*/ks/*) entsprechen.[495]

[485] Vgl. ebd. S. 45 f. (zum Kurzvokal *o* anstelle von *a* oder *ö*), 48 (zu *o* anstelle von *ü*), 54 f. (zum Langvokal *o* anstelle eines *a* oder *au* und zu *ö*-Schreibungen), 70 f. (zur Senkung) und 78 (zur Verwendung als Nebensilbenvokal).

[486] Vgl. ebd. S. 46 (zu *ö* vor *ei*) und 56 (zu *ö* anstelle von *eu* beziehungsweise *äu*) und 70 f. (zur Senkung).

[487] Vgl. ebd. S. 33.

[488] Vgl. ebd. S. 59 f. Die Schreibung *ouw* wird ebd. in der Auflistung möglicher Schreibungen nicht genannt, vielmehr heißt es, sie sei „bereits in mhd. Zeit zunehmend zu *ow~aw* gekürzt“ worden. Im DRQEdit-Korpus finden sich dafür vor allem mittelniederdeutsche, aber auch in einigen Texten frühneuhochdeutsche Belege.

[489] Vgl. Reichmann/Wegera 1993, S. 88 f. (zu *p* anstelle von *b*, *pf* oder *f*) und 135 (zu *mp*).

[490] Vgl. ebd. S. 129.

[491] Vgl. ebd. S. 128 f.

[492] Vgl. ebd. S. 150 f.

[493] Vgl. ebd. S. 115 (zu *s* anstelle von *z*) und 116 (zum Graphiewechsel vor *l*, *m*, *n* und *w*).

[494] Vgl. ebd. S. 117 f.

[495] Vgl. ebd. S. 111 (zum Verhältnis von stimmhaftem *s* und *ss* beziehungsweise *ß*) und 115 (zur Verwendung anstelle von *chs*).

- *t* kann auch einem neuhochdeutschen *d* entsprechen. Im Mittelfränkischen entspricht es im Auslaut einiger einsilbiger Wörter einem neuhochdeutschen *s*. Ein *tw* im Anlaut kann einem heutigen *qu* oder *zw* entsprechen. Schließlich gibt es bestimmte Fälle, in denen ein *t* vor oder nach einem anderen Konsonanten eingeschoben beziehungsweise angehängt sein kann.[496]
- *u* kann zum einen vokalisch (auch als Bestandteil eines Diphthongs) zu lesen sein, zum anderen konsonantisch. In vokalischer Funktion steht es auch für ein *ü*, da sich die Umlautkennzeichnung erst im Laufe der frühneuhochdeutschen Periode – je nach Region zu unterschiedlichen Zeiten – durchsetzte.[497] Es kann einem neuhochdeutschen *o* oder *au* sowie in Präfix-Nebensilben einem *e* entsprechen.[498] In konsonantischer Funktion kann es anstelle eines *w* oder *f* stehen.[499]
- *ů* entspricht meist einem neuhochdeutschen *u*, kann aber auch für ein *ü* oder *f* stehen.[500]
- *ü/ů* kann bei noch nicht erfolgter oder nicht in der Schreibung abgebildeter neuhochdeutscher Diphthongierung einem *eu* entsprechen.[501]
- *ui*, *uy* und *ue* können vor allem in westmitteldeutschen Texten als langes *u* zu lesen sein.[502]
- *v* kann grundsätzlich zur Bezeichnung der gleichen Laute verwendet werden wie *u*, allerdings steht es typischerweise – zumindest in den hier ausgewerteten Drucken – in initialer Stellung.[503]
- *w* kann selten einem neuhochdeutschen *f* entsprechen.[504] Es kann zusätzlich zu einem *u* oder auch alleine den zweiten Teil eines Diphthongs darstellen.[505] Auch der Gebrauch zur Repräsentation des Monophthongs *u* kommt vor.[506] Es kann vor allem zwischen Vokalen und nach einem *l* oder *r* und zum Teil auch initial anstelle eines neuhochdeutschen *b* stehen und zwischen Vokalen auch einem

[496] Vgl. ebd. S. 91 (zur Verwendung anstelle von *d*), 97 (zu mittelfränkischen Wortformen mit *t* im Auslaut und zu *t* in Epi- und Epenthese) und 133 (zu *tw*).
[497] Vgl. ebd. S. 35 und 48.
[498] Vgl. ebd. S. 47 (zum Kurzvokal *u* anstelle von *o*), 56 (zum Langvokal *u* anstelle von *o*), 64 f. (zur neuhochdeutschen Diphthongierung) und 78 (zur Verwendung als Nebensilbenvokal).
[499] Vgl. ebd. S. 104 f. (zu *u* anstelle von *w*) und 108 (zu *u* anstelle von *f*).
[500] Vgl. ebd. S. 63.
[501] Vgl. ebd. S. 65.
[502] Vgl. ebd. S. 33. *ui* wird in der Auflistung möglicher Schreibungen für langes *u* ebd. S. 56 zwar nicht genannt, ist aber durch das ebd. S. 33 genannte Beispiel *huis* belegt.
[503] Vgl. ebd. S. 46. Die ebd. S. 105 und 108 angeführten medialen Schreibungen treten im hier zugrunde gelegten Korpus nicht auf (jedenfalls soweit man Wörter in Großbuchstaben ausschließt, da für *U* und *V* in den gebrochenen Schriften der Zeit nur eine einzige Glyphe zur Verfügung steht, die der Form nach einem *V* entspricht). Die Verwendung von *v* nach Präfixen oder zu Beginn des Grundworts eines Kompositums ist allerdings in einem Teil der Texte gängige Praxis.
[504] Vgl. Reichmann/Wegera 1993, S. 109.
[505] Vgl. ebd. S. 59 f.
[506] Vgl. ebd. S. 47. Die ebd. S. 56 angeführte Verwendung zur Bezeichnung eines langen *ü* wird im dort als Verweis genannten Paragraphen L 17 nicht beschrieben, wäre aber dadurch zu erklären, dass sich die Umlautskennzeichnung, wie ebd. in L 17 erläutert, erst allmählich ausbreitete.

neuhochdeutschen *h* entsprechen beziehungsweise in der heutigen Schreibung ohne Gegenstück sein.[507]
- *x* kann auch einer der Buchstabenfolgen *ks*, *chs* oder *gs* entsprechen.[508]
- *y* kann einem neuhochdeutschen *i* oder seltener auch *j* entsprechen.[509]
- *z* kann auch einem neuhochdeutschen *s*, *ss*, *ß* oder *ts*, in der Zeichenfolge *sz* einem *ss* beziehungsweise *ß* und in Verbindung mit *c* und/oder *t* einem *tsch* entsprechen.[510]

Diese Zusammenstellung kann die Komplexität der möglichen Zuordnungen nur teilweise erkennbar machen. Generell ist darauf hinzuweisen, dass einzelne Laute im Frühneuhochdeutschen recht häufig durch mehrere Buchstaben repräsentiert werden, wobei nicht nur einfache Zeichenwiederholungen auch dort stehen können, wo sie nach der heutigen Orthographie nicht vorgesehen sind (und umgekehrt nicht immer vorhanden sind, wenn sie danach zu erwarten wären), sondern daneben oft auch solche Zeichenfolgen als lautliche Einheit zu lesen sind, deren Einzelbuchstaben für hinsichtlich der Artikulationsstelle ähnliche Laute stehen. So treten bei den Verschlusslauten zum Teil Fortis und Lenis nebeneinander auf. Viele Konsonanten können auch mit einem *h* kombiniert werden. Der Schwund von Lauten, die in der Standardsprache erhalten geblieben sind, wurde hier gar nicht verzeichnet.

Um etwas besser erkennbar zu machen, welche Buchstaben und Buchstabenfolgen durch Austauschbarkeitsbeziehungen miteinander verbunden sind, bietet Abbildung 3.1 eine entsprechende Graphdarstellung. Dabei wird aufgrund der Komplexität nur ein nach subjektiver Einschätzung der Relevanz ausgewählter Teil der eben beschriebenen Zuordnungsmöglichkeiten abgebildet, und bei Buchstabenfolgen sind fast nur solche berücksichtigt, die auch in der heutigen Orthographie als Repräsentation eines einzigen Phonems verstanden werden können.[511] Buchstabenwiederholungen sind gänzlich ausgeklammert. Zur Verdeutlichung sind die Verbindungslinien zwischen Buchstabenpaaren, bei denen (gegebenenfalls unter bestimmten Rahmenbedingungen wie Zeit und Region) ein Austausch in beiden Richtungen möglich ist, dicker gezeichnet. „-“ dient als Platzhalter für die leere Zeichenfolge, um das Entfallen eines Buchstabens darstellen zu können.

Abbildung 3.2 enthält einen ähnlich gestalteten Graphen, der die fakultative Verbindung von Buchstaben miteinander zeigen soll, bei denen einer der beiden (meist der zweite) keinen eigenen Lautwert hat und in der heutigen Orthographie

[507] Vgl. ebd. S. 85 und 106.
[508] Vgl. ebd. S. 102.
[509] Vgl. ebd. S. 43 f. (zu *y* anstelle von *i*), 58 (zu *ey* und *ay*) und 119 (zu *y* anstelle von *j*).
[510] Vgl. ebd. S. 111 (zu *z* anstelle von stimmhaftem *s*), 113 f. (zu *z* anstelle von stimmlosem *s*), 132 (zur Verwendung anstelle von *ts*) und 133 f. (zu Schreibungen für *tsch*).
[511] Inwieweit Diphthonge und Affrikaten als Einzelphoneme zu betrachten sind, wird unterschiedlich gesehen. Vgl. Bussmann 1990 s. v. „Diphthong“ und s. v. „Polyphonem(at)ische Wertung“.

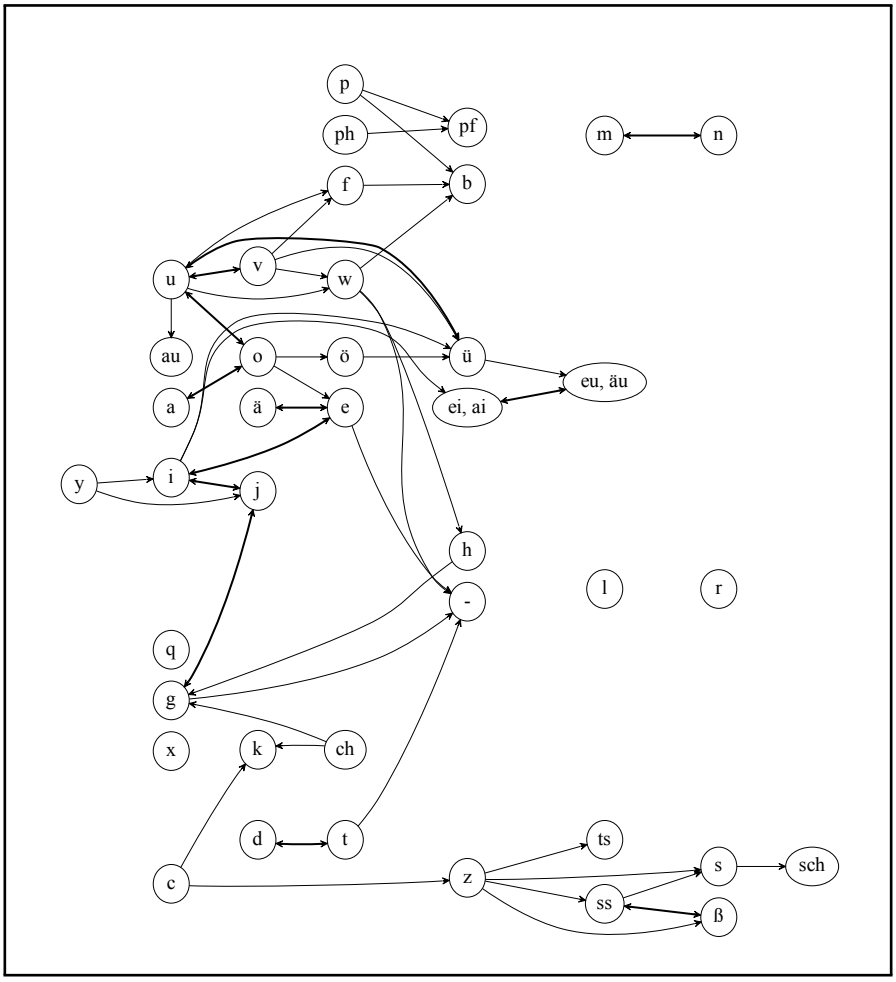


Abb. 3.1: wichtige Buchstabenaustauschmöglichkeiten im Frühneuhochdeutschen

entfällt beziehungsweise – etwa beim *h* als Dehnungszeichen – entfallen kann. Der Pfeil zeigt hier jeweils auf das zweite Glied der Folge. Auch hier geht es nur um Beziehungen zwischen unterschiedlichen Buchstaben, nicht um die auch aus der heutigen Orthographie vertrauten, im Frühneuhochdeutschen allerdings nur zum Teil den heutigen Schreibregeln entsprechenden Konsonantenverdoppelungen, und auch hier soll die Darstellung nicht sämtliche Möglichkeiten aufzeigen, sondern nur eine ohne sprachstatistische Untersuchungen getroffene Auswahl präsentieren.[512]

Anhand des ersten Graphen lassen sich mehrere Buchstabengruppen recht gut erkennen, innerhalb derer es relativ viele Austauschbarkeitsbeziehungen gibt, aber einige Buchstaben stehen als eine Art Bindeglied zwischen verschiedenen Gruppen. Insbesondere sind hier – wie schon dargestellt – *u/v/w* und *i/j/y* zu nennen, die die Vokale mit den Labial- und den Palatallauten verbinden und teilweise auch

[512] Die Graphen wurden mit dem Programm *dot* aus der *Graphviz*-Programmfamilie (vgl. *http://www.graphviz.org/*) generiert. Buchstaben, die für ähnliche Laute stehen, sind darin zwar nach Möglichkeit in Gruppen zusammengefasst, die genaue Positionierung der Knoten und Kanten der Graphen ist aber nur begrenzt aussagekräftig.

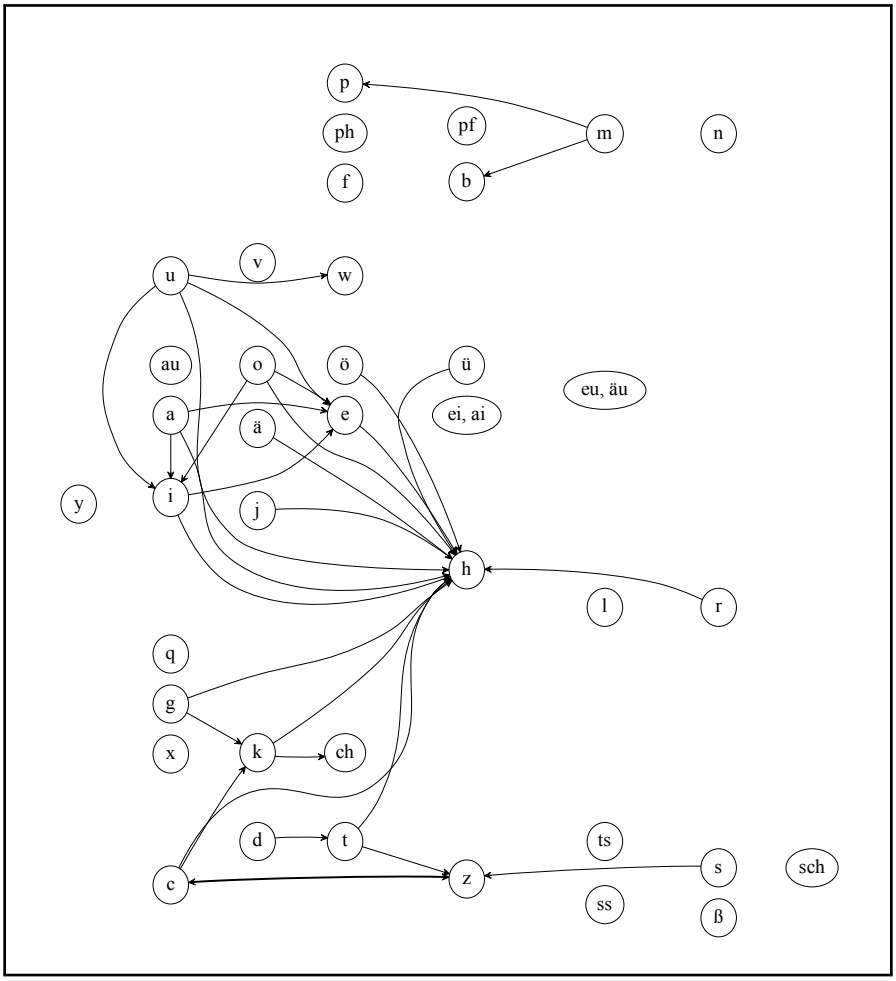


Abb. 3.2: wichtige Buchstabenkombinationen im Frühneuhochdeutschen

im Wortzusammenhang ohne Kenntnis der Schreibgewohnheiten im jeweiligen Text nicht sicher als Vokal oder Konsonant interpretiert werden können. Weniger problematisch ist das *c*, das zwar teils als *k* und teils als *z* zu lesen ist, dabei aber in aller Regel aufgrund des folgenden Buchstabens verlässlich einem der beiden Laute zugeordnet werden kann.

Einige Abgrenzungsprobleme sind abhängig von den zugrunde gelegten Texten. So spielt insbesondere die Verwendung von Frikativen anstelle von Plosiven der heutigen Hochlautung (und umgekehrt bei Hyperkorrektur) vor allem für Texte mit entsprechender dialektaler Prägung eine Rolle. Die Zusammenfassung der betreffenden Schreibungen erhöht dementsprechend zwar wohl den *Recall* gerade für die Erkennung von Beziehungen zwischen solchen Texten und Texten mit einer eher dem heutigen Standard entsprechenden Schreibweise, da davon aber gerade auch die eben genannten Buchstaben betroffen sind, die zwischen Vokalen und Konsonanten stehen, ist auch mit der Gefahr zu rechnen, dass durch die Zusammenfassung zu unspezifische Buchstabengruppen gebildet werden und die *Precision* sinkt.

Ein ähnliches Problem ergibt sich im Hinblick auf die Leerzeichen. Zwar lässt sich für die meisten Stellen auch bei Berücksichtigung aller für das Frühneuhochdeutsche zu konstatierenden Schreibvarianz sicher sagen, ob dort ein Spatium stehen muss oder nicht stehen darf, die Bestandteile von Kompositen und Präfixbildungen sind aber insbesondere in älteren Texten oft durch einen Abstand voneinander getrennt, und je nach zugrunde gelegter Vorlage ist die Setzung von Leerzeichen insbesondere zwischen bestimmten Buchstaben nicht immer sicher zu erkennen. Während also die meisten Wortabgrenzungen bei Textparallelen auch im Frühneuhochdeutschen übereinstimmen, kann ihre Einbeziehung in die varianzreduzierte Textform die Erkennung mancher Textübernahmen erschweren beziehungsweise bei dem hier aus Performanzgründen gewählten Vergleich auf exakte Übereinstimmung hin ganz verhindern.

Eine Aussage darüber, welche Zusammenfassung von Buchstabengruppen (oder auch Buchstabenfolgengruppen) am besten geeignet ist, die Erkennung von Textübereinstimmungen zu ermöglichen, lässt sich wohl nicht rein theoretisch treffen – ganz abgesehen davon, dass die Frage, in welchem Maße eine Steigerung des *Recalls* eine Minderung der *Precision* rechtfertigt (beziehungsweise umgekehrt), nicht allgemein zu entscheiden ist. Für die vorliegende Untersuchung wurden deshalb verschiedene Codierungsmöglichkeiten getestet, die im Hauptbestand der Umwandlungsregeln übereinstimmen, aber vor allem bei den eben beschriebenen Abgrenzungsproblemen unterschiedlich verfahren. Tabelle 3.1 stellt analog zu den Tabellen in Unterkapitel 2.3.4 die dabei angesetzten Buchstabengruppen und die verwendeten Codierungen, gegebenenfalls mit Varianten, zusammen. Dabei werden zur Vereinfachung nur die Grundbuchstaben des Alphabets, *ä*, *ö*, *ü*, *ß* sowie die Ligaturen *æ* und *œ*, die in aufgrund der Herkunft aus dem Lateinischen in Antiqua gesetzten Wörtern vorkommen, aufgeführt.

In der Tabelle sind mit der Streichung von *v*, *w*, *f*, *b* und *p* auch Varianten verzeichnet, die im Hinblick auf die betreffenden Buchstaben radikaler vorgehen als *Soundex* und die *Kölner Phonetik*[513]. Diese Varianten ergeben sich als mögliche Folgerung aus einigen der dargestellten Austauschbarkeitsbeziehungen, nämlich der zwischen *u*, *v* und *w*, der zwischen *u/v/w* (in konsonantischer Funktion) und *f*, der zwischen *w* und *b*, der zwischen *f* und *b/p/pf* und der zwischen *b* und *p*. Sie sind wohl allenfalls dann sinnvoll, wenn außerdem auch die Vokale gestrichen werden.[514]

[513] Vgl. oben Unterkapitel 2.3.4.

[514] Die eben gebotene Zusammenstellung von Beziehungen ist vereinfacht und soll nicht bedeuten, dass in jedem Fall eine Austauschbarkeit in beiden Richtungen gegeben ist. Sie soll vielmehr ähnlich wie Abbildung 3.1 erkennbar machen, dass es Verbindungen zwischen den genannten Buchstaben gibt, auch wenn sie teilweise über Zwischenglieder führen. Wenn *u* gestrichen wird, ist es plausibel, auch *v* und – nicht ganz so zwingend – auch *w* zu streichen. Damit werden dann aber auch Buchstaben eliminiert, denen in anderen Texten möglicherweise ein *f* oder *b*

Code	Buchstaben
1.) A 2.) -	*a*, *e*, *o*, *ä*, *ö*, *æ*, *œ*
1.) B 2.) -	*b*, *p*
C	*tz*, *z*, *x*; *c* vor einem hellen Vokal oder *z*
D	*d*, *t*
1.) F 2.) -	*f*
G	*g*, *k*, *q*; *c* außer vor einem hellen Vokal oder *z*
–	*h*
1.) I 2.) –	*i*, *j*, *y*
L	*l*
M	*m*, *n*
R	*r*
S	*sch*, *s*, *ß*
1.) U 2.) -	*u*, *v*, *w*, *ü*

Tab. 3.1: Codierungsvarianten für die Reduktion von Textvarianz

Neben den Buchstaben sind natürlich noch zahlreiche weitere Zeichen zu berücksichtigen, etwa – soweit sie in der zugrunde gelegten Textrepräsentation nicht schon aufgelöst sind – Kürzungszeichen und Diakritika sowie arabische Ziffern, Leerzeichen und Interpunktion. Für viele dieser Zeichen ist beim Ziel einer varianzreduzierten Form eine Streichung angemessen, allerdings wohl nicht für Ziffern, Kürzungszeichen (hier bietet sich eine Behandlung entsprechend der sicher oder mit der höchsten Wahrscheinlichkeit anzusetzenden ausgeschriebenen Form an) und Wortabstände. Letztere werden hier in verschiedenen Codierungsvarianten unterschiedlich behandelt, nämlich teils beibehalten (unter Entfernung von Trennzeichen mit folgendem Zeilenumbruch und unter Ersetzung der übrigen Zeilenumbrüche durch Spatien), teils ganz gestrichen.

Grundsätzlich werden aufeinander folgende gleiche Codezeichen zu einem einzigen zusammengefasst. Wenn Vokale (oder auch Labiale) eliminiert werden, geschieht dies erst nach dieser Zusammenfassung, so dass in solchen Fällen zwei gleiche Codezeichen nebeneinander stehen können.

Diese Regeln entsprechen den in Unterkapitel 2.3.4 dargestellten aus Codierungsverfahren wie zum Beispiel *Soundex*. Abweichend davon wird in den Codierungsvarianten, die eine solche Eliminierung vorsehen, diese zum Teil auch dann durchgeführt, wenn ein Wort mit einem oder mehreren entsprechenden Zeichen

entspricht. Und wenn auch diese Buchstaben gestrichen werden, kann es sinnvoll sein, auch *p* zu streichen. Ähnlich ließen sich natürlich auch noch weitere Codierungsvarianten bilden, um zum Beispiel den Laut- und Schreibungswechsel zwischen *i/j/y* und *g* zu berücksichtigen, was ebenfalls auch für weitere Buchstaben Konsequenzen hätte.

beginnt.[515] Diese Entscheidung beruht auf dem schon beschriebenen Problem der Wortabgrenzung – eine unterschiedliche Behandlung je nach Position im Wort unterminiert bei den davon betroffenen Wörtern den Versuch, Schreibungsabweichungen hinsichtlich der Leerzeichensetzung zu ignorieren. Die Streichung von Zeichen auch am Wortanfang kann allerdings die *Precision* verringern, insbesondere wenn sich aus den in der codierten Fassung nicht repräsentierten Buchstaben vollständige Präfixe bilden lassen. Soweit also zum Beispiel die Wörter *abschreiben*, *aufschreiben* und *schreiben* unterschieden werden sollen, ist eine Codierung, die neben Vokalen auch *b* und *f* generell eliminiert, problematisch.

Eine Möglichkeit, dieses Problem zu verkleinern, kann in einer kontextabhängigen Differenzierung bestehen. So lässt sich festlegen, dass eine Streichung von *b* beziehungsweise *f* nur dann erfolgen soll, wenn darauf ein als Vokal codierter Buchstabe folgt. Unter dieser Bedingung werden die Präfixe *ab* und *auf* zwar weiterhin getilgt, wenn der Wortstamm mit einem Vokal beginnt, sonst aber nicht.

Während es offenkundig ist, dass mit dem beschriebenen Codierungsverfahren vielfach – und in einem besonders hohen Maße bei den Varianten mit einer kleineren Zahl von Codebuchstaben – auch vom Stamm her ganz unterschiedliche Wörter auf gleiche Zeichenfolgen abgebildet werden, so dass die *Precision* bei der Zusammenfassung von Schreibungsvarianten kontextloser Wortformen (etwa in einem Wortindex) nicht allzu hoch sein dürfte, reduziert sich die Wahrscheinlichkeit solcher nicht dem Erkennungsziel entsprechenden Übereinstimmungen enorm, wenn stattdessen zusammenhängende Wortfolgen zugrunde gelegt werden. Tabelle 3.2 auf S. 127 stellt für das hier ausgewertete Korpus zu verschiedenen Codierungsvarianten und Größen von *n* jeweils vier Werte zusammen: An erster Stelle steht, wie viele unterschiedliche Wort-N-Gramme sich ergeben, an zweiter, wie häufig die Wort-N-Gramme durchschnittlich vorkommen (also die Gesamtzahl der einander überlappenden Folgen von jeweils *n* Wörtern[516] geteilt durch die Zahl unterschiedlicher N-Gramme); drittens ist angegeben, wie groß die Differenz zwischen der Gesamtzahl der laufenden N-Wort-Folgen und der der unterschiedlichen N-Gramme ist (also die Zahl der Wortfolgen, zu denen es Entsprechungen

[515] Eine Ausnahme davon wird hier allerdings dann gemacht, wenn das Wort kein weiteres Zeichen enthält, das durch einen Code repräsentiert wird – in diesem Fall wird ein ansonsten nicht vorkommendes Codezeichen eingefügt. Diese Entscheidung ist nicht inhaltlich begründet, sondern beruht auf einer Performanzsteigerung in der zugrunde gelegten Implementierung, betrifft aber, wenn *b* und *f* nicht einfach gestrichen werden, nur wenige Wörter wie „Ehe“, „je“ oder „wo“ sowie Abkürzungen, aber keine Präfixe mit potentiell unterschiedlicher Getrennt- oder Zusammenschreibung, so dass dadurch keine merkliche Minderung des *Recalls* hervorgerufen werden dürfte.

[516] Diese Gesamtzahl liegt zwischen 6.680.079 bei $n = 1$ und 6.677.571 bei $n = 12$. Die leichte Reduzierung bei steigendem *n* beruht darauf, dass sich das Korpus aus 228 Texten (in 176 Drucken) zusammensetzt und bei den jeweils letzten $n - 1$ Wörtern eines Textes kein vollständiges N-Gramm mehr anfängt.

gibt, ohne das erste Vorkommen zu zählen); der vierte Wert ist die Zahl von Paaren gleicher Textstücke (ein Wert, der nahe beim dritten liegt, wenn Entsprechungen weitestgehend singulär und auch die am häufigsten vorkommenden N-Gramme sehr selten sind, der aber bei höheren Vorkommenszahlen stark ansteigt[517]). Dabei wird neben einigen der dem in Tabelle 3.1 vorgestellten Grundschema entsprechenden Codierungsmöglichkeiten auch eine Abbildung auf eine Textfassung in Kleinschreibung, mit Vereinfachung im Hinblick auf Diakritika[518] und ohne Satzzeichen und Ähnliches berücksichtigt, so dass deutlich wird, in welchem Maße eine Zuordnung von Textstücken zueinander auf dem verwendeten Codierungsverfahren beruht. Die Tabellenspalten sind so geordnet, dass die Varianzreduktion von links nach rechts zunimmt.

Die verschiedenen Codierungen werden hier wie auch im Folgenden unter Benennung der Codezeichen bezeichnet, bei denen es zwischen ihnen Unterschiede gibt, insbesondere der Codes für die Vokale (beziehungsweise für die Buchstaben, die unter anderem zur Bezeichnung eines Vokals dienen können). So bedeutet *mit A/I/U*, dass die Buchstaben *i*, *j* und *y* in den Code *I* umgewandelt werden, die Buchstaben *u*, *v* und *w* in den Code *U* und alle übrigen Vokale in den Code *A*. Die Codierung *ohne A* stimmt damit weitgehend überein, streicht aber den Code *A* ersatzlos. Entsprechend werden in *ohne A/I/U* auch die Codes *I* und *U* entfernt, und in *ohne A/I/U/B/F* werden außerdem noch die Codes *B* und *F*, die anstelle der Buchstaben *b*, *p* und *f* stehen, gestrichen, die in den übrigen Codierungen erhalten bleiben.

Aus der jeweils zweiten Zahl in der Zeile für $n = 1$ lässt sich für die verschiedenen Codierungsmöglichkeiten berechnen, wie viele unterschiedliche Wortformen (ohne Berücksichtigung von Abweichungen in der Groß- und Kleinschreibung) im Durchschnitt auf eine gemeinsame Codefolge abgebildet werden, indem man diese Zahl durch den entsprechenden Wert in der Spalte „klein ohne Codierung“ teilt.

[517] Wenn zwei Textstücke übereinstimmen, bilden sie ein Paar. Jedes weitere hinzukommende Textstück kann mit jedem der schon verzeichneten ein zusätzliches Paar bilden, bei einem dritten Stück kommen somit zwei Paare hinzu (also als Summe 3), bei einem vierten drei (als Summe 6) usw. Es handelt sich also um die mathematische Reihe der natürlichen Zahlen, die durch die Formel $n(n + 1) / 2$ berechnet werden kann (wobei n hier um 1 niedriger ist als die Zahl der übereinstimmenden Textstücke). Diese Formel wird wohl in kaum einer Einführung in mathematische Reihen fehlen. Sie lässt sich leicht durch folgende Überlegung herleiten: Statt die Glieder jeweils in der Reihenfolge der natürlichen Zahlen zu summieren, lassen sie sich zu Paaren zusammenfassen, deren Stücke jeweils gleich weit von der Mitte der betrachteten Zahlenfolge entfernt sind – 1 wird mit n zusammengefasst, 2 mit $n - 1$ usw. Jedes dieser Paare hat die Summe $n + 1$. Bei einer geraden Anzahl von Zahlen gibt es genau $n/2$ solcher Paare (was nach den elementaren Rechenregeln der angegebenen Formel entspricht), bei einer ungeraden Anzahl sind es $(n - 1)/2$ Paare sowie als Rest die mittlere Zahl, deren Wert $(n + 1)/2$ ist. Als Summe aus den Paaren und der mittleren Zahl ergibt sich dann $((n - 1)/2)(n + 1) + (n + 1)/2$, also $(n^2 - 1)/2 + (n + 1)/2$, also $(n^2 + n)/2$ und somit auch in diesem Fall $n(n + 1)/2$.

[518] Dabei wurde $\overset{e}{a}$ in *ä*, $\overset{e}{o}$ in *ö* und $\overset{e}{u}$ in *ü* umgewandelt; sonstige übergeschriebene Zeichen wurden gestrichen.

n	klein ohne Codierung	*mit A/I/U*	*ohne A*	*ohne A/I/U*	*ohne A/I/U/B/F*
1	192.920	103.687	70.818	38.439	22.308
	34,626	64,425	94,327	173,784	299,448
	6.487.159	6.576.392	6.609.261	6.641.640	6.657.771
	111.828.759.755	165.603.096.180	213.244.491.288	371.904.705.381	444.820.841.409
2	2.076.700	1.490.541	1.222.058	849.885	593.690
	3,217	4,481	5,466	7,860	11,251
	4.603.151	5.189.310	5.457.793	5.829.966	6.086.161
	810.431.762	1.444.938.404	1.933.452.520	5.041.256.484	7.852.387.598
3	4.585.469	3.869.025	3.558.178	2.987.081	2.496.165
	1,457	1,726	1,877	2,236	2,676
	2.094.154	2.810.598	3.121.445	3.692.542	4.183.458
	28.494.928	55.501.016	67.528.899	125.440.195	199.882.560
4	5.751.702	5.212.684	5.030.546	4.792.587	4.543.873
	1,161	1,281	1,328	1,394	1,470
	927.693	1.466.711	1.648.849	1.886.808	2.135.522
	3.911.948	8.692.175	9.516.099	11.630.860	14.119.886
5	6.134.834	5.680.549	5.536.360	5.457.264	5.399.670
	1,089	1,176	1,206	1,224	1,237
	544.333	998.618	1.142.807	1.221.903	1.279.497
	1.256.443	2.964.390	3.193.191	3.551.904	3.782.321
6	6.299.527	5.879.079	5.739.102	5.682.027	5.651.741
	1,060	1,136	1,164	1,175	1,182
	379.412	799.860	939.837	996.912	1.027.198
	628.098	1.587.513	1.852.834	2.001.077	2.088.398
7	6.393.739	5.995.921	5.855.236	5.801.776	5.774.828
	1,045	1,114	1,141	1,151	1,157
	284.972	682.790	823.475	876.935	903.883
	399.518	1.115.199	1.389.204	1.500.430	1.564.409
8	6.455.478	6.077.543	5.936.724	5.883.861	5.857.477
	1,035	1,099	1,125	1,135	1,140
	223.005	600.940	741.759	794.622	821.006
	286.320	887.069	1.147.905	1.248.067	1.305.173
9	6.499.553	6.140.410	6.000.375	5.947.313	5.920.810
	1,027	1,088	1,113	1,123	1,128
	178.702	537.845	677.880	730.942	757.445
	215.759	744.667	987.344	1.081.721	1.135.380
10	6.532.227	6.191.211	6.053.138	5.999.743	5.972.940
	1,022	1,079	1,103	1,113	1,118
	145.800	486.816	624.889	678.284	705.087
	167.854	642.599	869.030	959.687	1.011.062
11	6.557.253	6.233.694	6.098.332	6.044.762	6.017.644
	1,018	1,071	1,095	1,105	1,110
	120.546	444.105	579.467	633.037	660.155
	133.252	563.879	776.548	864.034	913.731
12	6.576.753	6.269.967	6.137.835	6.084.197	6.056.843
	1,015	1,065	1,088	1,098	1,102
	100.818	407.604	539.736	593.374	620.728
	108.836	502.300	702.395	786.317	834.537

Tab. 3.2: für unterschiedlichen Codierungen und *n*-Werte jeweils 1.) Anzahl der (unterschiedlichen) Wort-N-Gramme, 2.) durchschnittliche Vorkommenshäufigkeit der N-Gramme, 3.) Differenz zwischen N-Wort-Folgen- und N-Gramm-Zahl und 4.) Anzahl der Paare übereinstimmender N-Wort-Folgen (vgl. S. 125–129)

Die Codierungsvariante, die die drei Codes *A*, *I* und *U* für die Repräsentation von klar als Vokale zu erkennenden Buchstaben sowie *i/j/y* und *u/v/w* vorsieht, fasst also typischerweise weniger als zwei Wortformen mit unterschiedlichem Buchstabenbestand zu einer gleichen Codeform zusammen, die Variante, die eindeutig oder möglicherweise als Vokal zu interpretierende Buchstaben sowie *b* und *f* letztlich streicht, hingegen mehr als acht.

Bei der Betrachtung von zusammenhängenden Wortfolgen schrumpft die Bandbreite aber schnell stark zusammen. Auch bei der am wenigsten differenzierenden Variante *ohne A/I/U/B/F* ist schon bei Wort-Trigrammen die durchschnittliche Vorkommenshäufigkeit mit 2,676 weniger als doppelt so hoch wie beim Vergleich von in Kleinschreibung umgewandelten Textfassungen, bei Vier-Wort-Folgen liegt sie unter 1,5 und nähert sich bei weiterer Erhöhung von *n* immer mehr der 1 an – die Codierung ändert also nur wenig daran, dass zu den meisten etwas längeren Textstücken im Korpus keine genaue Übereinstimmung gefunden wird.

Wesentlich größer sind die Unterschiede bei etwas höheren Werten für *n* allerdings, wenn man die jeweils dritte Zahl betrachtet, also die Differenz zwischen der Zahl der laufenden N-Wort-Folgen und der Zahl der unterschiedlichen N-Gramme, aus der sich besser entnehmen lässt, in welchem Maße überhaupt Übereinstimmungen gefunden werden und wie stark ihre Erkennung von der Reduzierung der Varianz abhängt. Bei der Betrachtung einzelner Wörter sind all diese Werte recht nahe an der Gesamtzahl der laufenden Wortformen, es gibt also auch bei Beibehaltung aller Schreibungsvarianten kaum Formen ohne Parallelen.

Die Spannbreite der Werte geht allerdings schnell auseinander. Absolut sind die Unterschiede am größten bei $n = 3$, was daran liegen dürfte, dass Wortfolgen dieser Länge insbesondere bei wenig differenzierenden Codierungsvarianten noch zu kurz sind, um wirklich aussagekräftige Übereinstimmungen zu liefern, während bei der Fassung in Kleinbuchstaben schon der größte Teil der für Einzelwörter festzustellenden Entsprechungen entfällt. Die relativen Unterschiede der Werte steigen auch bei einer weiteren Erhöhung von *n* an. So ist zum Beispiel bei der am wenigsten vereinfachenden Variante *mit A/I/U* der Wert bei sechs Wörtern schon mehr als doppelt so hoch wie bei der Abbildung in Kleinbuchstaben, bei zwölf Wörtern erreicht er sogar eine Höhe von mehr als dem Vierfachen der Zahl für *klein ohne Codierung*, und für die am stärksten vereinfachende Codierung *ohne A/I/U/B/F* ergibt sich im Vergleich zu *mit A/I/U* nochmals eine Steigerung um fast 30 % bei sechs und um mehr als 50 % bei zwölf Wörtern.

Noch stärkere Unterschiede treten beim vierten Wert, also der Zählung der übereinstimmenden Wortfolgenpaare, zutage, und zwar auch und im Hinblick auf die absoluten Differenzen sogar am stärksten bei $n = 1$. Auch hier zeigt sich, dass Wort-Trigramme jedenfalls bei Anwendung einer stark vereinfachenden Codierung noch sehr wenig spezifisch sind, dass die Trefferzahl dann aber zunächst noch schnell sinkt. Ab $n = 6$ bleibt in den codierten Fassungen im Vergleich zu einem

um 1 niedrigeren *n* jeweils mehr als die Hälfte der Zuordnungen erhalten, bei Erhöhung von *n* von 11 auf 12 in den meisten Varianten sogar mehr als 90 %. Bei der Fassung in Kleinschreibung vermindern sich die Werte prozentual etwas stärker, wie es ja auch sachlich naheliegt – mit zunehmender Größe der N-Gramme steigt die Wahrscheinlichkeit, dass im Hinblick auf die Formulierung übereinstimmende Wortfolgen aufgrund eines Unterschiedes in der Schreibung einander in dieser Fassung nicht mehr zugeordnet werden.[519]

Auch wenn sich anhand dieser Werte wohl vermuten lässt, dass die Zahl der Zuordnungen durch das vorgestellte Codierungsverfahren bei einem nicht zu klein gewählten *n* nur in einem Maße steigt, das in Anbetracht der Schreibungsvarianz im Frühneuhochdeutschen einigermaßen plausibel ist, ist eine genauere Einschätzung auf dieser Datenbasis schwierig. Deshalb soll in Tabelle 3.3 auf S. 130 zusätzlich noch dokumentiert werden, wie entsprechende Daten für eine ähnlich große Textmenge in vermutlich weitgehend einheitlicher Orthographie aussehen. Die dabei ausgewerteten Texte[520] sind weder inhaltlich noch im Hinblick auf wörtliche Übernahmen irgendwie vergleichbar mit dem in dieser Arbeit untersuchten Korpus von Rechtstexten. Grund für die Zusammenstellung für diese Vergleichsuntersuchung war vor allem, dass es sich überwiegend um Werke der zweiten Hälfte des 19. und des frühen 20. Jahrhunderts handelt. Insbesondere für die vor der Orthographischen Konferenz von 1901[521] publizierten Texte muss zwar mit gewissen Abweichungen von der heutigen Rechtschreibung gerechnet werden, sie fallen aber im Vergleich zur Varianz in frühneuhochdeutschen Texten kaum ins Gewicht.

Dass die Zahlen deutlich abweichen und insbesondere die Häufigkeit etwas längerer Übereinstimmungen viel geringer ist als in den Texten aus dem DRQEdit-Korpus, ist in Anbetracht dieser Unterschiede nicht weiter erstaunlich.[522] Hervorzuheben

[519] Allerdings ist bei der Codierung *ohne A/I/U/B/F* der relative Unterschied zur Fassung in Kleinschreibung für den vierten Wert im hier betrachteten Wertebereich ($n = 1$ bis $n = 12$) bei $n = 2$ am größten, und auch das Verhältnis des vierten Werts für die Codierung *ohne A/I/U* zu dem für die Fassung in Kleinschreibung hat hier ein lokales Maximum. Dies lässt sich wohl so erklären, dass bei diesen Codierungen bei niedrigen Werten für *n* vielfach auch solche N-Gramme einander zugeordnet werden, bei denen gar keine oder keine vollständige Wortgleichheit (abgesehen von möglichen Schreibungsunterschieden) vorliegt. Bei höheren *n*-Werten sinkt die Wahrscheinlichkeit solcher Fehlzuordnungen enorm.

[520] Es handelt sich um die im *TextGrid Repository* (*https://textgridrep.org/repository.html*) zur Verfügung gestellten Werke von Hedwig Dohm, Marie von Ebner-Eschenbach, Gustav Freytag, Carl Hauptmann, Hugo von Hofmannsthal, Franz Kafka, Fanny Lewald, Conrad Ferdinand Meyer, Christian Morgenstern, Wilhelm Raabe, Theodor Storm und Ludwig Thoma. Die annotierten Daten stehen unter der *Creative Commons Namensnennung 3.0 Deutschland Lizenz* (*http://creativecommons.org/licenses/by/3.0/de/legalcode*), die Texte selbst sind gemeinfrei. Es handelt sich laut Vermerk in den Dateien um „eine Abwandlung des Datenbestandes von www.editura.de durch TextGrid“ (Förderkennzeichen des Bundesministeriums für Bildung und Forschung: 01UG1203A; vgl. *http://www.textgrid.de/Digitale-Bibliothek*).

[521] Vgl. zum Beispiel *http://de.wikipedia.org/wiki/Orthographische_Konferenz_von_1901*.

[522] Die längeren Übereinstimmungen in den Texten des literarischen Korpus dürften zum größten Teil auf Textparallelen innerhalb des jeweiligen Gesamtwerks eines Autors beruhen. Das ist

n	klein ohne Codierung	*mit A/I/U*	*ohne A*	*ohne A/I/U*	*ohne A/I/U/B/F*
1	199.064	165.774	128.792	82.971	51.433
	34,194	41,060	52,850	82,037	132,341
	6.607.642	6.640.932	6.677.914	6.723.735	6.755.273
	125.482.848.484	163.047.745.434	204.778.398.822	424.259.857.443	528.953.099.472
2	2.276.677	2.035.771	1.719.807	1.162.196	810.439
	2,990	3,344	3,958	5,857	8,399
	4.530.016	4.770.922	5.086.886	5.644.497	5.996.254
	1.306.943.574	1.782.674.630	2.298.062.774	7.626.987.552	12.190.417.353
3	5.211.969	5.029.251	4.735.929	3.764.415	3.034.156
	1,306	1,353	1,437	1,808	2,243
	1.594.711	1.777.429	2.070.751	3.042.265	3.772.524
	20.025.374	24.798.375	33.970.744	132.553.151	262.184.264
4	6.512.044	6.477.901	6.419.282	6.030.909	5.587.892
	1,045	1,051	1,060	1,129	1,218
	294.623	328.766	387.385	775.758	1.218.775
	803.242	915.330	1.119.249	3.116.665	6.498.751
5	6.747.047	6.743.441	6.737.124	6.701.392	6.640.479
	1,009	1,009	1,010	1,016	1,025
	59.607	63.213	69.530	105.262	166.175
	89.789	97.060	108.408	161.553	255.791
6	6.781.825	6.781.017	6.779.804	6.777.290	6.774.114
	1,004	1,004	1,004	1,004	1,005
	24.816	25.624	26.837	29.351	32.527
	28.608	30.574	32.276	35.298	38.965
7	6.789.082	6.788.674	6.788.137	6.787.740	6.787.436
	1,003	1,003	1,003	1,003	1,003
	17.546	17.954	18.491	18.888	19.192
	18.631	19.398	20.019	20.471	20.814
8	6.791.523	6.791.215	6.790.808	6.790.645	6.790.565
	1,002	1,002	1,002	1,002	1,002
	15.092	15.400	15.807	15.970	16.050
	15.723	16.225	16.646	16.820	16.907
9	6.792.843	6.792.571	6.792.191	6.792.063	6.792.026
	1,002	1,002	1,002	1,002	1,002
	13.759	14.031	14.411	14.539	14.576
	14.223	14.595	14.983	15.118	15.158
10	6.793.769	6.793.517	6.793.141	6.793.029	6.793.000
	1,002	1,002	1,002	1,002	1,002
	12.820	13.072	13.448	13.560	13.589
	13.175	13.448	13.829	13.948	13.979
11	6.794.508	6.794.279	6.793.898	6.793.788	6.793.760
	1,002	1,002	1,002	1,002	1,002
	12.068	12.297	12.678	12.788	12.816
	12.344	12.587	12.970	13.087	13.115
12	6.795.154	6.794.941	6.794.552	6.794.442	6.794.416
	1,002	1,002	1,002	1,002	1,002
	11.409	11.622	12.011	12.121	12.147
	11.616	11.840	12.230	12.347	12.373

Tab. 3.3: Daten wie in Tabelle 3.2 für ein Korpus von Texten in vermutlich weitgehend einheitlicher Orthographie (vgl. S. 129–129

ist aber, dass die Reduzierung der Varianz durch Codierung in den Texten mit weitgehend standardisierter Orthographie jedenfalls bei höheren *n*-Werten nur einen vergleichsweise kleinen Einfluss auf die Zahl ermittelter Entsprechungen hat. Auch daraus lässt sich zwar kein genauer Wert für die jeweilige *Precision* ableiten (unter anderem weil dieser auch davon abhängt, wie hoch überhaupt der Anteil tatsächlicher Entsprechungen ist, weil bei einer niedrigen Quote falsche Zuordnungen natürlich umso mehr ins Gewicht fallen), aber es sollte doch ersichtlich sein, dass die Wahrscheinlichkeit relativ gering ist, dass das vorgestellte Codierungsverfahren zu Zuordnungen führt, die nicht auf einer tatsächlichen Übereinstimmung im Wortlaut beruhen.

3.1.3 Positionsspeicherung

Wenn die in dem varianzreduzierten Textmaterial gefundenen Übereinstimmungen nicht nur irgendwie statistisch ausgewertet, sondern konkreten Textpassagen zugeordnet werden sollen, müssen die entsprechenden Positionen in den eigentlichen Texten rekonstruiert werden. Das ist noch relativ einfach zu bewerkstelligen, allerdings mit einem nicht zu vernachlässigenden Aufwand, wenn bei der Erstellung der vereinfachten Textfassung Leerzeichen erhalten bleiben – dann lässt sich jeweils abzählen, beim wievielten Wort eine Entsprechung beginnt, und diese Zählung in den unveränderten Texten wiederholen. Sofern die Texttransformation aber auch die Entfernung von Leerzeichen beinhaltet, fehlen in der varianzreduzierten Fassung die Informationen, aus denen sich die Originalpositionen erschließen ließen. Deshalb wird bei dem für diese Untersuchung entwickelten Verfahren bei der Transformation für jedes Wort die Position im Ausgangstext und in der vereinfachten Textform protokolliert.

Dass sich die Positionsangaben auf Wörter und nicht auf Buchstaben oder zum Beispiel Sätze beziehen, ist für die hier untersuchte Fragestellung wohl sachlich angemessen. Eine Zuordnung von Einzelbuchstaben beziehungsweise Buchstabengruppen zu Codes, um bei gefundenen Übereinstimmungen unterschiedliche Schreibweisen direkt aufeinander beziehen zu können, lässt sich allerdings mit dem hier eingesetzten Verfahren zumindest nicht ohne Anpassungsmaßnahmen bewerkstelligen und wäre im Hinblick auf Laufzeit und Speicheraufwand sicherlich erheblich aufwendiger.

Grundannahme ist hier, dass die zu speichernden Informationen jeweils Textstücke betreffen, deren Transformation unabhängig voneinander erfolgen kann.[523] Dann können diese Textstücke jeweils nacheinander mit dem Transformationsver-

jedenfalls der Befund bei einer Ermittlung von *maximal exact matches* bei Ansetzung einer größeren Mindestlänge (vgl. unten Kapitel 3.2).

fahren bearbeitet werden, wobei anschließend anhand der Länge von Ausgangstextstück und Resultat berechnet wird, welche Position im Originaltext und in der codierten Fassung erreicht ist.

Die Speicherung dieser Positionen als 32-Bit-*Integer*-Daten reicht für Texte in allen tatsächlich vorkommenden Größen aus.[524] Da für jedes Wort die Endposition in beiden Fassungen verzeichnet werden muss (die Anfangsposition ergibt sich aus dem jeweils vorangehenden Wert), sind dafür acht Byte erforderlich – ein Wert, der etwas höher ist als der durchschnittliche Speicherbedarf für den originalen Text im untersuchten Korpus.[525]

Die Ermittlung des originalen Textstücks zu einer bestimmten Position im codierten Text (oder auch umgekehrt eines codierten Textstücks zu einer Position im eigentlichen Text) kann über eine binäre Suche erfolgen. Damit ist der Zeitaufwand nur logarithmisch von der Textlänge abhängig.[526]

Nach demselben Verfahren lässt sich eine Zuordnung zwischen *XPath*-Ausdrücken und Textstücken herstellen. Der Speicheraufwand ist hier abhängig davon, wie feingliedrig das *Tagging* ist und bis zu welcher Stufe es bei der Textextraktion entsprechend Unterkapitel 3.1.1 protokolliert wurde.

Es ist aber natürlich nicht sinnvoll, hier Zuordnungen zu verzeichnen, die den Text in mehr Segmente unterteilen, als bei der Erstellung einer varianzreduzierten Fassung entsprechend Unterkapitel 3.1.2 protokolliert werden, also nach dem hier vorgesehenen Verfahren Einzelwörter. Der Speicherbedarf kann also bei einer sachgerechten Vorgehensweise maximal denselben Wert erreichen wie bei der

[523] Das ist bei Wörtern (worunter hier Buchstabenfolgen verstanden werden, die durch Leerraum voneinander getrennt sind) oder auch Wortketten jedenfalls für das hier untersuchte Sprachmaterial wohl weitestgehend unproblematisch – soweit Ersetzungsregeln eine differenzierte Behandlung je nach Position im Wort vorsehen, wird hier in Kauf genommen, dass unterschiedliche Wortabgrenzungen zu unterschiedlichen Ergebnissen führen können. Bei einer Transformation auf Buchstabenebene wäre es erforderlich, Buchstabengruppen, die als Einheit codiert werden sollen, vorab zusammenzufassen; zudem müsste das Verfahren um eine Auswertung von Informationen über die Zeichen davor und danach erweitert werden, um gegebenenfalls davon abhängige Regeln anwenden zu können.

[524] Es wird damit – bei *unsigned integers* – der Zahlenbereich bis 4.294.967.295 abgedeckt, also könnten theoretisch Textdateien mit einer Größe bis zu etwa vier GB verarbeitet werden.

[525] Bei Speicherung in UTF-8 sind 47.303.440 Byte für die 6.680.079 laufenden Wortformen des Untersuchungskorpus erforderlich, also durchschnittlich etwa 7,1 Byte je Wort. Dieser Wert dürfte auch innerhalb einer Sprache – abhängig von Textgrundlage und Zeichensatz – leicht schwanken, zum einen aufgrund der je nach Schreibgewohnheiten beziehungsweise orthographischen Konventionen unterschiedlichen Beziehung zwischen Lauten und Buchstaben, zum anderen aufgrund der Abhängigkeit vom gewählten Vokabular, zum Beispiel im Hinblick auf den Anteil längerer Komposita. Auch mit diesen Vorbehalten lässt sich aber verallgemeinern, dass sich der Speicherbedarf in einer ähnlichen Größenordnung bewegt wie der für die zu verarbeitenden Texte.

[526] Vgl. zum Beispiel Ottmann/Widmayer 1996, S. 154–156. Die binäre Suche wird oben auf S. 112 kurz beschrieben.

Positionsverzeichnung für die Textcodierung. Wenn statt Wörtern Sätze, Absätze oder Textabschnitte mit kanonischer Referenz verzeichnet werden, reduziert sich der Aufwand erheblich.

Da sich die Positionsdaten problemlos serialisieren lassen[527], brauchen sie nicht permanent im Arbeitsspeicher gehalten zu werden, so dass das Verfahren auch für die Verarbeitung großer Datenmengen geeignet ist.

3.2 MEM-Ermittlung

In diesem Kapitel geht es um die Ermittlung von *maximal exact matches (MEMs)* in Textdaten. Diese Aufgabe deckt sich aufgrund des unterschiedlichen Zeichenbestandes nicht völlig mit der MEM-Ermittlung im Rahmen der Bioinformatik, die oben in Unterkapitel 2.2.1 vorgestellt wurde.

Unterkapitel 3.2.1 beschreibt deshalb für vier Programme, die sich mit relativ geringem Aufwand an die Verarbeitung von Texten anpassen lassen, welche Änderungen dafür durchzuführen sind und an welchen Stellen mit gewissen Einschränkungen zu rechnen ist.

Unterkapitel 3.2.2 stellt Messergebnisse zum Zeit- und Speicherbedarf der MEM-Ermittlung mit diesen Programmen vor. Im Hinblick auf den Vergleich der Programme und Aufrufvarianten zeigen sich dabei teilweise deutliche Unterschiede zu ähnlichen Messungen, die anhand von DNA-Sequenzen durchgeführt wurden.

Unterkapitel 3.2.3 untersucht die MEMs im hier ausgewerteten Korpus in verschiedenen Codierungen und bei Ansetzung verschiedener Mindestlängen im Hinblick auf quantitative Merkmale, wobei es auch um erste Anhaltspunkte dafür geht, wie aussagekräftig die Funde für die Ermittlung textueller Beziehungen sind.

3.2.1 Programmanpassung

Oben in Unterkapitel 2.2.1 wurden bereits verschiedene Programme vorgestellt, die im Rahmen der Bioinformatik für die MEM-Ermittlung entwickelt wurden. Zusammenfassend lässt sich feststellen, dass mehrere Programme, nämlich *MUMmer*, *sparseMEM* und das darauf aufbauende *essaMEM* sowie *backwardMEM*, auch für den Vergleich von Textdaten in einem 8-Bit-Zeichensatz gut geeignet sind, wenn kleinere Anpassungen vorgenommen werden und wenn die Daten entsprechend bestimmten Vorgaben des *multi-FASTA*-Datenformats gespeichert werden können.

[527] In der für diese Untersuchung entwickelten Implementierung in *Perl* werden die Positionen ohnehin aus datentechnischen Gründen als Gesamtstring gespeichert, in dem jeweils acht Byte einem Positionspaar entsprechen; die Konvertierung von und in Zahlen erfolgt über die Funktionen *pack* und *unpack*.

Wie oben schon beschrieben, sieht das *multi-FASTA*-Format vor, dass zunächst eine mit „>“ beginnende Zeile einen Identifikator für die Sequenz enthält und alle folgenden Zeilen bis zur nächsten mit „>“ beginnenden Zeile beziehungsweise bis zum Ende der Datei die eigentlichen Sequenzdaten. Zeilenumbrüche werden beim Einlesen der Sequenzdaten übersprungen, ebenso sind Leerzeichen irrelevant. Zwischen Groß- und Kleinbuchstaben wird nicht unterschieden.[528]

Bei der Berücksichtigung dieser Regeln gibt es kleinere Unterschiede. So sieht *MUMmer* eine Verarbeitung Zeichen für Zeichen vor, berücksichtigt die Zeilenstruktur nur beim Umschalten von der Verarbeitung der Daten in der Kopfzeile zum Einlesen der Sequenzdaten und überspringt in den Sequenzdaten alle *Whitespace*-Zeichen. Demgegenüber verarbeiten *sparseMEM*, *essaMEM* und *backwardMEM* die Dateien zeilenweise. „>“ muss dementsprechend tatsächlich am Zeilenanfang stehen. Leerzeichen werden nur am Anfang und Ende von Zeilen entfernt, wobei offenbar die stillschweigende Annahme ist, dass sie innerhalb von Zeilen ohnehin nicht vorkommen.

Die Einbeziehung von Leerzeichen in den Vergleich lässt sich in diesen Programmen mit geringen Anpassungsmaßnahmen erreichen.[529] Die generelle Nichtberücksichtigung von *Whitespace* innerhalb der eigentlichen Sequenzdaten ergibt sich in *MUMmer* aus der in *scanmultiplefastafile* in der Datei *maxmatinp.c* gesetzten Bedingung „if (!isspace ((Ctypeargumenttype) tmpchar))“ , die zum Beispiel durch „if (tmpchar != '\n')“ ersetzt werden kann, um alle Zeichen bis auf den Zeilenumbruch zu berücksichtigen.[530] In den drei anderen hier betrachteten Programmen kann der Aufruf der *trim*-Funktion abgeschaltet werden.[531]

Wenn auch Unterschiede zwischen Groß- und Kleinbuchstaben berücksichtigt werden sollen, lässt sich das leicht erreichen, indem die Bearbeitung der eingelesenen Daten durch die Funktion *tolower* ausgeschaltet wird.[532]

Bei *sparseMEM* und *essaMEM* kann es außerdem ein Problem bei der Verarbei-

[528] Diese Beschreibung berücksichtigt nicht alle Details, sondern fasst die für die Verarbeitung in den hier untersuchten Programmen wichtigen Punkte zusammen. Vgl. oben S. 65, Anm. 256.

[529] Das ist nicht unbedingt auf Zeilenumbrüche zu übertragen, da diese im *multi-FASTA*-Format auch eine Steuerfunktion für die Verarbeitung haben. Wie auch sie in den Vergleich einbezogen werden könnten, wurde hier nicht untersucht, da es hier nicht um den Vergleich verschiedener Ausgaben eines Textes mit nahezu identischem Druckbild geht, sondern um den Vergleich verschiedener Texte, bei denen ein gleicher Zeilenumbruch im Bereich einer Textübereinstimmung wohl als in aller Regel zufällig zu bezeichnen ist. Dementsprechend enthalten die codierten Daten anstelle von Zeilenumbrüchen Leerzeichen, und bei einem Zeilenumbruch mitten im Wort entfällt der Umbruch (und gegebenenfalls das Trennzeichen).

[530] Auf den Ausschluss des Zeilenumbruchs aus den zu vergleichenden Zeichen kann anscheinend ohne weitere Anpassungsmaßnahmen nicht verzichtet werden.

[531] *trim* kommt sowohl in der Datei *fasta.cpp* vor (wobei *backwardMEM* die Datei aus dem parallelen *sparseMEM*-Verzeichnis verwendet) als auch in einer weiteren Datei (*mummer.cpp* in *sparseMEM* und *essaMEM*, *backwardMEM.cpp* in *backwardMEM*).

[532] Auch diese Funktion kommt wie *trim* sowohl in *fasta.cpp* als auch in einer weiteren Datei vor, vgl. die Angaben in Anm. 531.

tung von Zeichen geben, die nicht zum eigentlichen ASCII-Code gehören. Es lässt sich beheben, wenn im *Makefile* der Wert von „FLAGS“ um „ -funsigned-char“ erweitert wird.[533]

Allerdings zeigten sich in Testläufen für *sparseMEM* und *essaMEM* bei Ansetzung einer vergleichsweise kurzen MEM-Mindestlänge von 18 oder weniger Zeichen bei der Untersuchung des Korpus auf der Basis einer Umwandlung in Kleinbuchstaben und Leerzeichen[534] auch nach Durchführung dieser Anpassungsmaßnahmen in Einzelfällen kleinere Ungenauigkeiten in den ausgegebenen MEM-Informationen. Die Ursache ließ sich im Rahmen der vorliegenden Untersuchung nicht klären.[535]

In *sparseMEM* und *essaMEM* haben mehrere Zeichen eine besondere Funktion, nämlich (soweit im Quellcode keine Anpassungen vorgenommen werden) „$“, „`“ und – wenn über einen Parameter angegeben wird, dass nur die Zeichen *a*, *c*, *g* und *t* (also die Codes für die DNA-Bestandteile) berücksichtigt werden sollen – „~“. Dementsprechend sind für beide Programme weitere Anpassungen erforderlich, falls diese Zeichen auch in den zu verarbeitenden Daten vorkommen.

Außerdem ist darauf hinzuweisen, dass die hier beschriebenen Quellcodeänderungen nichts daran ändern, dass von allen in diesem Unterkapitel betrachteten Programmen Zeilenumbrüche in den zu vergleichenden Daten nicht berücksichtigt werden und das Zeichen „>“ (zumindest am Zeilenanfang) als Code für die Abgrenzung zwischen zwei Sequenzen beziehungsweise Strings dient. Dies spielt für die vorliegende Untersuchung keine Rolle, gegebenenfalls wären aber weitere Änderungen im Quellcode oder Umcodierungen in den Untersuchungsdaten erforderlich, um auch diese Zeichen in der gewünschten Weise verarbeiten zu können.

Die mit den beschriebenen Änderungen kompilierten Programme haben in Tests mit dem Untersuchungskorpus für etwas größere Mindestlängen die gleichen MEMs ermittelt;[536] bei *sparseMEM* und *essaMEM* gab es allerdings – wie eben schon etwas

[533] Die Festlegung ist relevant, weil der Wertebereich des Datentyps *char* andernfalls (maschinenabhängig) auch der von *signed char* sein kann (vgl. *https://gcc.gnu.org/onlinedocs/gcc-4.0.4/gcc/C-Dialect-Options.html#index-funsigned_002dchar-114*) und in einer Codetransformation in *sparseSA::sparseSA* in der Datei *sparseSA.cpp* keine Zeichen berücksichtigt werden, denen ein Wert unter 0 zugeordnet ist.

[534] Dabei wurden Satz- und Sonderzeichen weitgehend eliminiert, aber zum Beispiel Umlaute unverändert gelassen.

[535] Teilweise sind die Abweichungen möglicherweise darauf zurückzuführen, dass das Zeichen „$“ in einer besonderen Funktion verwendet und dabei vorausgesetzt wird, dass es – im Hinblick auf den zugeordneten numerischen Wert – kleiner als alle tatsächlich in den Daten vorkommenden Zeichen ist (im Quellcode der beiden Programme steht in der Datei *sparseSA.cpp* der Kommentar: „It must be lexicographically less“). Das ist aber zum Beispiel dann nicht der Fall, wenn der zu verarbeitende String Leerzeichen enthält. Die Ersetzung von '$' (einschließlich der einfachen Anführungszeichen) durch das durch die Zahl 31 (ohne Anführungszeichen) repräsentierte Zeichen führte dazu, dass die MEM-Ermittlung für die Mindestlänge 18 und $K = 1$ dieselben Ergebnisse wie zum Beispiel *MUMmer* lieferte. Allerdings ergaben sich auch nach dieser Änderung Abweichungen, wenn K einen höheren Wert hatte oder die Mindestlänge auf 17 reduziert wurde.

näher beschrieben – für kürzere Längen in bestimmten Fällen Unstimmigkeiten. Ob es auch noch andere problematische Konstellationen gibt, kann hier nicht geklärt werden.[537] Jedenfalls ist festzuhalten, dass die in diesem Unterkapitel betrachteten Programme voraussetzen, dass die Zeichen in einem 8-Bit-Zeichensatz codiert werden können.

Der Vollständigkeit halber soll noch kurz betrachtet werden, inwieweit die hier untersuchten Programme mit den beschriebenen Änderungen für die MEM-Ermittlung in Texten mit *Unicode*-Zeichen verwendet werden können.

Natürlich ist es prinzipiell möglich, *Unicode*-Daten als Folgen von Bytes zu verarbeiten und in dieser Form zu vergleichen. Wenn dabei eine *Unicode*-Codierung zugrunde gelegt wird, die für jedes Zeichen dieselbe Zahl an Bytes vorsieht (also zum Beispiel nicht UTF-8), ist es anschließend mit einem geringen Aufwand möglich, die ermittelten Bytepositionen in Zeichenpositionen umzurechnen.[538]

Allerdings ist damit zu rechnen, dass die Daten auch Bytes enthalten können, denen im ASCII-Code eines der eben als problematisch beschriebenen Zeichen zugeordnet ist, oder auch Bytes mit dem numerischen Wert 0, also dem Code, der in C (und bei einer String-Repräsentation im C-Stil auch in C++) zur Markierung eines String-Endes dient. Dementsprechend ist eine solche Vorgehensweise fehleranfällig.

Eine Alternative besteht darin, nicht die *Unicode*-Zeichen zu vergleichen, sondern codierte Textfassungen, die nur 8-Bit-Zeichen enthalten. Das entspricht dem in dieser Untersuchung gewählten Ansatz eines Vergleichs auf der Basis einer Codierung, wobei möglicherweise ein passender Standard für die Zeichenzuordnung zur Verfügung steht.[539] Und auch wenn ein 8-Bit-Zeichensatz nicht ausreicht, um eine 1:1-Abbildung der *Unicode*-Zeichen vorzunehmen, und wenn es tatsächlich um exakte Übereinstimmungen geht, eignet sich eine 8-Bit-Codierung, bei der unterschiedliche Zeichen zusammengefasst werden, vermutlich als Basis für einen ersten Schritt bei der MEM-Ermittlung, um anschließend die gefundenen Stellen in den Originaltexten zu vergleichen, da die Wahrscheinlichkeit zufälliger Übereinstimmungen jedenfalls für nicht sehr kurze MEMs wohl gering sein dürfte.

[536] Dies wurde jedenfalls im Hinblick auf die Größe der Ausgabedateien überprüft.

[537] Es wäre etwa daran zu denken, dass möglicherweise noch weitere Zeichen von einem Programm mit einer besonderen Bedeutung belegt werden und dabei vorausgesetzt wird, dass sie in den untersuchten Strings nicht vorkommen. Eine Suche im Quellcode der Programme hat über die genannten Fälle hinaus keine Hinweise darauf ergeben, aber es wäre wohl nichts Ungewöhnliches. So ist es, wie oben bereits erwähnt, ein gängiges Verfahren, in Suffixbäumen das String-Ende durch ein besonderes, im String selbst nicht vorkommendes Zeichen (zum Beispiel „$“) zu codieren.

[538] Dabei ist natürlich auf die korrekte Anpassung an die Zeichengrenzen zu achten, und im theoretisch denkbaren, aber praktisch wohl extrem unwahrscheinlichen Fall, dass ein Match ermittelt wird, das in den beiden Strings nicht an derselben Stelle eines Zeichens beginnt (also zum Beispiel im ersten String beim ersten Byte eines Zeichens und im zweiten String beim zweiten Byte), ist das Match ganz zu streichen.

[539] Vgl. zum Beispiel zur Gruppe der ISO-8859-Zeichensätze *https://de.wikipedia.org/wiki/ISO_8859*.

Die Angaben zu Matches, die einer bestimmten ID der (mit der Referenzdatei verglichenen) Abfragedatei zuzuordnen sind, werden in den von den vier genannten Programmen ausgegebenen Daten jeweils durch eine entsprechende ID-Zeile eingeleitet. Die zugehörigen Funde sind offenbar nach den Positionen im Abfragestring geordnet, wobei *backwardMEM* absteigend sortiert, die anderen drei Programme aufsteigend. Die Sortierung der MEMs, die an der gleichen Position im Abfragestring beginnen, weist keine unmittelbar erkennbaren Regeln auf. Sie ist bei *sparseMEM* und *essaMEM* – soweit sich das aufgrund von Stichproben verallgemeinern lässt – gleich, stimmt aber ansonsten nur partiell überein.

Die beschriebenen Sortierkriterien sind vermutlich nicht auf inhaltliche Überlegungen zurückzuführen, sondern einfach technisch bedingt – diese Annahme erklärt jedenfalls die absteigende Sortierung durch *backwardMEM*. Sie erleichtern einerseits die Prüfung, ob eine bestimmte Textstelle eine Vielzahl an Entsprechungen (mit einem oder mehreren Texten) aufweist, also zum Beispiel möglicherweise auf ein festes sprachliches Muster zurückzuführen ist. Andererseits ist aber unmittelbar nur schwer zu erkennen, ob ein Textpaar eine Vielzahl an Entsprechungen aufweist und wie gut sich mehrere Matches zu einem Bereich mit einem hohen Entsprechungsgrad zusammenfassen lassen.

Dementsprechend kann es sinnvoll sein, die Matches so zu sortieren, dass die zu einem Textpaar gehörenden Übereinstimmungen jeweils zusammengefasst sind. Dabei legt es sich nahe, als zusätzliches Sortierkriterium die Position in einem der Texte zugrunde zu legen. Zum einen führen häufig vorkommende Formulierungen oft auch innerhalb eines Textpaars zu einer Vielzahl an Entsprechungen, was sich in einer sortierten Liste schneller erkennen lässt. Zum anderen erleichtert die Positionssortierung die Prüfung, ob Matches im Umfeld anderer Matches liegen und sich mit diesen zu größeren Bereichen mit (mehr oder weniger) hoher Ähnlichkeit zusammenfassen lassen. Wie sich eine nach Positionen sortierte Liste für eine Matchbewertung nutzen lässt, wird unten am Ende von Unterkapitel 3.3.2 näher betrachtet.

3.2.2 Programmvergleich

Die Frage des Zeit- und insbesondere auch des Speicherbedarfs der Programme zur Ermittlung von *maximal exact matches* wurde in Unterkapitel 2.2.1 bereits thematisiert, und sie kann für die praktische Verwendbarkeit entscheidend sein. Deshalb finden sich in mehreren der dort angeführten Publikationen Informationen zu Messungen bei Testläufen. Diese Daten sind aber nicht unbedingt direkt auf den Vergleich von Textdaten, wie sie hier untersucht werden, übertragbar, und sie beziehen sich auch im Hinblick auf den Genomvergleich nur auf bestimmte Konstellationen. Deshalb sollen hier entsprechende Angaben folgen, die für die MEM-Ermittlung im Testkorpus in zwei Codierungen und für verschiedene Min-

destlängen erhoben wurden. Dabei wurden die Fassung in Kleinbuchstaben mit Leerzeichen und die Codierung *ohne A/I/U/B/F* ohne Leerzeichen zugrunde gelegt und damit von den in dieser Untersuchung betrachteten Codierungsvarianten die beiden Formen, die sich im Hinblick auf Differenziertheit und Textmenge am stärksten unterscheiden.

Zusammenfassend ist zunächst einmal festzuhalten, dass der maximale Speicherbedarf der hier betrachteten Programme und Aufrufvarianten nicht von der für die MEMs geforderten Mindestlänge abhängt, sondern vom Umfang der zu vergleichenden Daten, der allerdings je nach gewählter Codierung recht unterschiedlich sein kann, bei *sparseMEM*, *essaMEM* und *backwardMEM* außerdem von der Größe des Ausdünnungsfaktors K beziehungsweise k, bei *essaMEM* und *backwardMEM* auch davon, ob der für die MEM-Ermittlung erforderliche Index schon existiert oder noch erstellt werden muss, und bei *essaMEM* außerdem auch noch davon, ob eine Kombination von *sparse child array* und Suffixlinks verwendet wird oder nur eine der beiden Erschließungsstrukturen[540] und ob bei der Verarbeitung des Abfragestrings ebenfalls eine Ausdünnung erfolgt. Auch der Zeitbedarf wird von diesen Faktoren teilweise sehr stark beeinflusst. Dementsprechend werden für die genannten Programme verschiedene Varianten untersucht. Es wird jeweils eine Programmversion zugrunde gelegt, in der die in Unterkapitel 3.2.1 beschriebenen Änderungen (einschließlich der in Anmerkung 535 genannten) vorgenommen wurden.

Die Auswahl der dokumentierten Konfigurationsvarianten ist für die verschiedenen Programme unterschiedlich und auch davon beeinflusst, dass sie teilweise zu einem im Vergleich sehr hohen Zeitbedarf führen und deshalb praktisch wohl kaum von Interesse sind. Es sollte sich aber die Entwicklungstendenz von Laufzeit und Speicherbedarf bei Erhöhung des Ausdünnungsfaktors für die verschiedenen Programme und sonstigen Konfigurationseinstellungen erkennen lassen.

Die Tabellen 3.4 und 3.5 stellen die für die beiden genannten Codierungen ermittelten Werte zusammen. Die Messung wurde für den Aufruf eines Programms mit den Parametern, die sich aus den Angaben zur Codierung, zur Mindestlänge, gegebenenfalls zum Ausdünnungsfaktor und bei *essaMEM* auch zu weiteren Einstellungen ergeben, jeweils nur einmal durchgeführt; dementsprechend sollte bei den Zeitangaben ein gewisser Schwankungsbereich berücksichtigt und das Verhältnis der Zahlen zueinander nicht überbewertet werden. Die Angaben zum Arbeitsspeicher beziehen sich auf den jeweils für einen Programmlauf ermittelten Maximalwert; die Messungen wurden im Abstand von 0,1 Sekunden durchgeführt.[541]

[540] Prinzipiell ist es auch möglich, auf beides zu verzichten, dies führt allerdings zu einer massiven Verlangsamung.

[541] Die Messung erfolgte über ein *Perl*-Skript unter Verwendung des Moduls *Win32::OLE* auf einem Rechner mit einem 2,9-GHz-Prozessor und 8 GB Arbeitsspeicher.

Programm / Konfiguration	Arbeitsspeicher (MB) Indexaufbau	Vergleich	Zeitbedarf (Sek.) für Mindestlänge 18	24	30	36	42	48	54	60	66
MUMmer:											
	800 – 925		292	86	61	57	49	46	46	46	45
sparseMEM:											
K=1	430		187	51	41	41	42	37	34	42	33
K=2	239		212 (≠)	69	59	75	55	55	52	56	51
K=3	176		249 (≠)	113	111	133	112	103	97	102	104
K=4	144		325 (≠)	157	157	159	167	161	157	170	159
K=5	125 – 126		697 (≠)	243	249	266	241	235	240	212	206
K=6	112		2185 (≠)	300	268	265	265	260	282	261	271
essaMEM:											
K=1, Konfig. 1	597	430	177	55	55	52	47	38	39	43	38
K=1, Konfig. 3	766	600	176	64	47	62	61	44	44	45	45
K=2, Konfig. 1	321	239	198 (≠)	73	60	73	67	54	54	55	54
K=2, Konfig. 3	406	324	176 (≠)	57	45	49	58	42	43	41	43
K=3, Konfig. 1	229	176	247 (≠)	123	110	121	108	101	98	106	102
K=3, Konfig. 3	286	232	197 (≠)	70	58	81	64	55	54	54	54
K=4, Konfig. 1	183	144	340 (≠)	172	158	181	154	154	154	155	162
K=4, Konfig. 2	183 – 184	144	6911	4567	2869	2379	1745	1556	1280	1220	1077
K=4, Konfig. 3	226	186	239 (≠)	84	73	99	77	69	67	68	68
K=5, Konfig. 1	156	125	658 (≠)	229	217	213	227	231	220	210	214
K=5, Konfig. 3	190	159	397 (≠)	99	85	82	82	80	81	81	80
K=6, Konfig. 1	111 – 138	112	2134 (≠)	277	270	242	264	261	264	264	261
K=6, Konfig. 2	138	112	9274	4657	3138	2292	1880	1596	1343	1195	1002
K=6, Konfig. 3	166	141	1414 (≠)	117	96	95	93	92	93	105	92
K=8, Konfig. 1	115	96	9536	518	343	366	426	403	353	355	334
K=8, Konfig. 2	115	96	7366	6794	3533	2368	1722	1765	1386	1156	1005
K=8, Konfig. 3	136	118	6754	235	120	108	121	112	117	117	106
backwardMEM:											
k=1	701 – 742	452	350	207	178	166	166	162	163	180	161
k=2	616 – 659	367	581	292	243	254	232	224	223	211	226
k=4	574 – 613	325	763	359	307	316	284	293	278	282	279
k=8	553 – 595	304	1342	553	469	413	416	414	384	390	384

Tab. 3.4: Zeit- und Speicherbedarf der verschiedenen Programme für die MEM-Ermittlung in der Textfassung in Kleinbuchstaben (42.229.648 Zeichen). Das Ungleichheitszeichen markiert Unstimmigkeiten. Konfig. 1: -suflink 1 -child 0 -skip 1 (Standard-Konfiguration von *essaMEM* bis $K = 3$), Konfig. 2: -suflink 0 -child 1 (Standard-Konfiguration von *essaMEM* ab $K = 4$), Konfig. 3: -suflink 1 -child 1 -skip 1

Zunächst soll der Speicherbedarf etwas näher betrachtet werden. Dazu eine Vorbemerkung: Für *MUMmer* sowie für *essaMEM* und *backwardMEM* mit Neuerstellung des jeweils benötigten Indexes wurden teilweise entgegen dem, was ohne detaillierte Kenntnis wohl zu erwarten wäre, bei Zugrundelegung der gleichen Codierung, aber unterschiedlicher Mindestlänge deutlich unterschiedliche Angaben zum Speicherbedarf protokolliert, wobei aber nach den Messungen kein Abhängigkeitsverhältnis zwischen der Mindestlänge und dem belegten Speicher erkennbar ist. Die Schwankung ist allem Anschein nach darauf zurückzuführen, dass die Programme während der Konstruktion der Indexstruktur für sehr kurze Zeit deutlich mehr Speicher benötigen und der verzeichnete Wert davon abhängt, zu welchen Zeitpunkten

Programm / Konfiguration	Arbeitsspeicher (MB) Indexaufbau	Vergleich	Zeitbedarf (Sek.) für Mindestlänge 15	18	24	30	36	42	48	54	60	66
MUMmer:												
	274 – 421		35	28	26	27	26	23	23	22	21	21
sparseMEM:												
K=1	183		19	17	14	16	15	14	13	14	14	14
K=2	101		28	26	24	25	23	24	23	23	23	24
K=3	75		65	62	63	63	62	60	59	60	62	64
K=4	61		107	120	98	103	111	111	103	104	105	107
K=5	53 – 54		773	145	128	133	142	129	131	123	119	120
K=6	48		2150	156	126	126	125	125	126	125	125	129
essaMEM:												
K=1, Konfig. 1	251	183	20	17	17	16	15	15	15	15	16	16
K=1, Konfig. 3	320	253	22	22	18	18	25	18	19	18	18	17
K=2, Konfig. 1	135	101	29	26	26	26	26	26	23	24	24	25
K=2, Konfig. 3	169	136	23	21	19	18	24	18	19	18	18	18
K=3, Konfig. 1	96	75	68	63	63	71	63	61	59	61	63	62
K=3, Konfig. 3	119 – 120	98	32	32	28	27	38	30	29	27	27	27
K=4, Konfig. 1	62 – 77	61	108	103	101	100	99	101	100	99	110	110
K=4, Konfig. 2	62 – 77	61	2369	1159	751	484	389	292	250	213	189	172
K=4, Konfig. 3	94	79	42	38	35	35	48	34	36	34	34	34
K=5, Konfig. 1	54 – 60	54	793	124	122	122	122	131	128	132	126	130
K=5, Konfig. 3	67 – 80	67	478	40	39	37	37	37	37	37	37	37
K=6, Konfig. 1	48 – 58	48	1573	183	125	126	124	128	125	124	135	138
K=6, Konfig. 2	48 – 58	48		1509	803	503	382	297	253	219	192	170
K=6, Konfig. 3	60 – 70	60	1426	72	40	38	39	38	38	39	38	37
K=8, Konfig. 1	42 – 45	42	1730	1792	129	128	126	124	120	119	120	121
K=8, Konfig. 2	42 – 45	42		1143	1144	572	389	306	302	238	191	162
K=8, Konfig. 3	51 – 58	50	1196	1178	48	44	40	45	40	33	33	33
backwardMEM:												
k=1	285 – 292	182	63	62	58	66	59	57	58	61	57	58
k=2	253 – 265	147	84	80	77	78	75	74	74	77	72	80
k=4	236 – 239	130	107	101	100	98	94	90	91	96	92	90
k=8	227 – 239	121	154	140	143	147	125	133	123	131	120	118

Tab. 3.5: Zeit- und Speicherbedarf der verschiedenen Programme für die MEM-Ermittlung in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (17.240.460 Zeichen). Konfig. 1–3: vgl. Legende zu Tab. 3.4

genau die Messungen durchgeführt wurden.[542] Dementsprechend ist davon auszugehen, dass der insgesamt für die verschiedenen Mindestlängen gemessene höchste Wert dem tatsächlichen Maximum am nächsten kommt; um aber den ermittelten Befund nicht zu verfälschen, wird in der Tabelle der Wertebereich der gemessenen Maximalwerte verzeichnet.

[542] Bei *essaMEM* wurden nur bei der Codierung *ohne A/I/U/B/F* ohne Leerzeichen und nur bei höheren *K*-Werten unterschiedliche Werte gemessen. Das ist wohl darauf zurückzuführen, dass die Indexerstellung in den betreffenden Fällen sehr wenig Zeit benötigte. Da die Indexerstellung nicht unabhängig von der eigentlichen MEM-Ermittlung erfolgt, sondern ihr im Bedarfsfall (nämlich wenn der benötigte Index noch nicht vorhanden ist) automatisch vorangeht, bezieht sich die Angabe zum Speicherbedarf in der Spalte „Indexaufbau“ *de facto* auf den höchsten Speicherbedarf, der insgesamt während eines Programmlaufs mit Indexerstellung gemessen wurde.

Der Speicherbedarf für ein bestimmtes Programm beziehungsweise eine bestimmte Programmvariante hängt offenbar in etwa linear von der als Referenzstring eingelesenen Datenmenge ab.[543] Dabei gibt es kleinere Unterschiede. Die eingelesene Textmenge ist für die Fassung in Kleinbuchstaben mit Leerzeichen etwa 2,45mal so groß wie für die Fassung *ohne A/I/U/B/F* ohne Leerzeichen, das entsprechende Verhältnis liegt beim Speicherbedarf für die verschiedenen Programme und Programmvarianten zwischen etwa 2,2 für *MUMmer* (wobei wie gesagt mit gewissen Messungenauigkeiten zu rechnen ist) und etwa 2,5 für *backwardMEM* nach der Indexerstellung.

Wenn man die absoluten Zahlen vergleicht, lässt sich feststellen, dass – wie zu erwarten war – *MUMmer* den höchsten Speicherbedarf hat. Für *sparseMEM* und *essaMEM* nach der Indexerstellung werden – wie es den Angaben in der Literatur entspricht[544] – jeweils gleiche Werte verzeichnet, sofern die vorgesehene Konfiguration von *essaMEM* nicht durch bestimmte Aufrufparameter verändert wird. Die Indexerstellung erhöht den Speicherbedarf von *essaMEM* bei Verwendung nur einer der beiden Erschließungsstrukturen auf einen nach den hier dokumentierten Messungen maximal etwas weniger als 1,4fachen Wert;[545] wenn die Erschließungsstrukturen kombiniert werden, verdoppelt das die Zunahme in etwa. Noch höhere Werte erreicht *backwardMEM*, jedenfalls wenn die Indexdaten neu erstellt werden müssen oder wenn k beziehungsweise K größer als 1 ist.[546]

Auch im Hinblick auf den Zeitbedarf soll auf einige Punkte hingewiesen werden. Zunächst auch hier eine Vorbemerkung: Allgemeine Aussagen zum Zeitaufwand zu treffen, wird dadurch erschwert, dass er auch von der angesetzten Mindestlänge beeinflusst wird, wobei kleine Änderungen in bestimmten Konstellationen sehr stark ins Gewicht fallen, nämlich bei sehr kurzen Mindestlängen und bei höheren Ausdünnungsfaktoren, wobei die Unterschiede zwischen den Programmen und Aufrufvarianten sowohl absolut als auch relativ recht groß sind.

[543] Für diese Messungen wie auch für die weiteren MEM-Ermittlungen wurde das ganze Untersuchungskorpus sowohl als Referenz- als auch als Abfragestring zugrunde gelegt, es wurde also alles mit allem verglichen und auch textinterne Vergleiche wurden mit einbezogen. Dass der Speicherbedarf vom Umfang des Referenzstrings abhängt, wird für *MUMmer* explizit konstatiert (*http://mummer.sourceforge.net/manual/#mummer*), und auch im Hinblick auf Suffixarrays und davon abgeleitete Datenstrukturen geht es in der Literatur regelmäßig um den dafür erforderlichen Speicherbedarf (vgl. die oben in Unterkapitel 2.2.1 aufgeführte Literatur). In den Messdaten zeigt sich die Abhängigkeit von der Größe des Referenzstrings bei *essaMEM* und *backwardMEM* darin, dass der Speicherbedarf bei der Indexerstellung, die ja für den Referenzstring erfolgt, am größten ist.

[544] Vgl. Vyverman u. a. 2013b, S. 6.

[545] Je größer der Ausdünnungsfaktor ist, desto geringer die Indexgröße und damit auch der Speicherbedarf.

[546] Das ist ein deutlich anderer Befund, als nach den in Ohlebusch/Gog/Kügel 2010, S. 357 dokumentierten Messungen zu erwarten wäre. Danach sollte der Speicherbedarf für *backwardMEM* bei einem niedrigen k-Wert deutlich unter dem von *sparseMEM* und *essaMEM* (in der Standardkonfiguration) liegen, wenn man die Indexerstellung ausklammert.

Dass die Verarbeitungszeit steigt, wenn die geforderte Mindestlänge reduziert wird, lässt sich darauf zurückführen, dass die Anzahl der Matches umso größer ist, je kürzer diese Länge ist. Wie unten noch ausführlicher dargestellt werden soll, unterscheiden sich die Fallzahlen bei etwas größeren Mindestlängen aber nur vergleichsweise wenig, bei den kürzesten hier betrachteten Längen aber in einem sehr starken Maße.

Das erklärt auch, warum die Verwendung von höheren Ausdünnungsfaktoren bei niedrigen Mindestlängen zu einer viel stärkeren Erhöhung des Zeitbedarfs führt: Über die Indexstruktur werden zunächst einmal kürzere Matches ermittelt, die dann jeweils daraufhin überprüft werden müssen, ob sie die geforderte Mindestlänge erreichen. Die Zahl dieser kürzeren Matches ist bei niedrigen Mindestlängen und höheren Ausdünnungsfaktoren aber noch einmal viel höher als die der Matches, die die letztlich geforderte Länge erreichen.

Wie stark dies den Zeitaufwand beeinflusst, hängt vom jeweiligen Programm ab. Offenbar ist der Effekt bei *sparseMEM* recht ausgeprägt. Bei *essaMEM* in den hier als „Konfig. 1“ und „Konfig. 3“ bezeichneten Aufrufvarianten wirkt sich die Ausdünnung nicht ganz so stark aus und noch weniger bei *backwardMEM*.[547]

Für *essaMEM* zeigte sich, dass die automatische Konfiguration der MEM-Ermittlung auf der Basis des K-Werts für den Vergleich von Textdaten sehr ungünstig sein kann. Im Programmcode ist vorgesehen, dass ab $K = 4$ zur Beschleunigung der MEM-Ermittlung ein *sparse child array* verwendet wird, bei einem kleineren K-Wert hingegen Suffixlinks.[548] Die Tests zeigen hingegen, dass eine allein auf ein *sparse child array* gestützte MEM-Ermittlung (in den Tabellen „Konfig. 2“) in den Untersuchungsdaten insbesondere in der Fassung in Kleinbuchstaben sowie bei sehr kurzen Mindestlängen in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen ausgesprochen zeitaufwendig ist,[549] dass auch bei $K = 4$ und $K = 6$ sowie überwiegend auch bei $K = 8$ Suffixlinks anstelle eines *sparse child array* demgegenüber zu einer deutlichen Beschleunigung führen und dass jedenfalls für die hier dokumentierten K-Werte zwischen 2 und 8 die in den Tabellen als „Konfig. 3“ bezeichnete Kombination von Suffixlinks mit einem *sparse child array* und der Abschaltung der Ausdünnung für die Verarbeitung des Abfragestrings im Hinblick auf den Zeitbedarf die günstigste Konfiguration ist.[550] Sie geht zwar mit der Erstellung einer zusätzlichen Indexdatei und einer Erhöhung des Speicherbedarfs offenbar

[547] Ebd. S. 357 wird für die MEM-Erkennung in DNA-Daten festgestellt, dass sich die Erhöhung des Ausdünnungsfaktors bei *backwardMEM* weniger stark auf die Laufzeit auswirke als bei *sparseMEM* und nicht in jedem Fall zu einer Verlangsamung führen müsse.

[548] Die automatische Konfiguration findet sich in der *main*-Funktion in der Datei *mummer.cpp* im Bedingungsblock „if(automatic)“. Nach der Beschreibung in Vyverman u. a. 2013b, S. 4 wäre eigentlich zu vermuten, dass auch bei kleinen K-Werten in der Standardkonfiguration ein *sparse child array* eingesetzt wird.

[549] Die Datenlücken in Tabelle 3.5 in der Spalte für die Mindestlänge 15 beruhen darauf, dass die Ausgabedatei bei $K = 6$ und $K = 8$ in Kombination mit der Konfiguration 2 auch nach mehreren Stunden Laufzeit noch gänzlich leer war; daraufhin wurde der Testlauf abgebrochen.

entsprechend der Größe dieser Indexdatei einher, die Indexdateien werden aber mit zunehmender Größe von K immer kleiner und zugleich steigt der relative Zeitgewinn gegenüber der Konfiguration 1. Schon bei $K = 3$ in Kombination mit der Konfiguration 3 werden im Hinblick auf Zeit- und Speicherbedarf in etwa ähnliche Werte erreicht wie bei einem um 1 geringeren K-Wert und der Konfiguration 1, und bei $K = 6$ in der Konfiguration 3 kann die MEM-Ermittlung – jedenfalls für die hier untersuchten Zeichenfolgen und ohne Berücksichtigung von sehr kurzen Mindestlängen – im Vergleich zu $K = 4$ in der Konfiguration 1 mit weniger Arbeitsspeicher und zugleich schneller durchgeführt werden.

Allerdings ist zu betonen, dass diese Konfiguration auch den Parameter „-skip 1“ beinhaltet. Wenn er weggelassen wird, führt das bei Verwendung der Parameter „-suflink 1 -child 1“ zu einer massiven Verlangsamung der MEM-Ermittlung. Ähnliches würde auch für „-suflink 1 -child 0“ gelten; allerdings sieht der Quellcode in diesem Fall schon eine entsprechende automatische Setzung vor.[551]

Wenn auf eine Ausdünnung verzichtet wird (und teilweise auch bei $K = 2$), ist *essaMEM* (in einer günstigen Konfiguration) am schnellsten, dicht gefolgt von *sparseMEM*, während *MUMmer* und insbesondere *backwardMEM* deutlich mehr Zeit benötigen. Sofern der verfügbare Arbeitsspeicher ausreicht, liegt es also nahe, *essaMEM* zu verwenden.

Allerdings sei noch einmal darauf hingewiesen, dass die oben in Unterkapitel 3.2.1 beschriebenen Quellcodeänderungen nicht ausreichen, um mit *essaMEM* (oder auch mit *sparseMEM*) in jedem Fall ganz exakte Angaben zu den in den untersuchten Texten enthaltenen MEMs zu erhalten. In den Tests zeigten sich zwar für die Codierung *ohne A/I/U/B/F* ohne Leerzeichen keine Unstimmigkeiten, wohl aber – wenn auch in sehr geringem Umfang – für die Fassung in Kleinbuchstaben. Wie in Tabelle 3.4 durch Zusatz des Ungleichheitszeichens vermerkt, betraf dieses Problem in den hier dokumentierten Konstellationen nur die Mindestlänge 18 und trat bei $K = 1$, $K = 8$ und in der Konfiguration 2 auch bei anderen K-Werten nicht auf. Bei der Ermittlung von MEMs mit der Mindestlänge 17 gab es allerdings auch bei $K = 1$ Unstimmigkeiten.[552] Verschiedene K-Werte führten jeweils zu etwas unterschiedlichen Ausgabedaten.

Um sicherzustellen, dass die im weiteren Verlauf dieser Untersuchung ausgewerteten MEM-Daten exakt sind, wurde für ihre Ermittlung *MUMmer* verwendet.

[550] Diese Kombination ist für den Vergleich von DNA-Sequenzen offenbar nicht sinnvoll (so der Befund in Vyverman u. a. 2013a, S. 803 und Vyverman u. a. 2013b, S. 4; vgl. die Untersuchungsdaten und Diagramme in Vyverman u. a. 2013b, S. 14 f., 17, 19 und 21).

[551] Das Verhalten wird über die Bedingung „if(automaticSkip)“ in der Datei *mummer.cpp* gesteuert. Die darin enthaltene Bedingung „if(suflink && !child)“ deckt die Parameterkombination „-suflink 1 -child 1“ nicht ab. Das ist aber wohl dadurch zu erklären, dass sich diese Kombination für den Vergleich von DNA-Sequenzen als ungünstig erwiesen hat und von den Entwicklern deshalb nicht weiter in Betracht gezogen wurde.

[552] Weitere Aufrufvarianten wurden für diese Mindestlänge nicht untersucht.

3.2.3 MEMs im Untersuchungskorpus

Die Daten im letzten Unterkapitel haben schon gezeigt, dass der Zeitbedarf für die MEM-Ermittlung auch von der angesetzten Mindestlänge beeinflusst wird: Die Messungen für eine Programmkonfiguration liegen für unterschiedliche größere Mindestlängen zwar nahe beieinander, bei den hier betrachteten niedrigeren Mindestlängen gibt es aber zum Teil deutliche und bei manchen Programmen und Aufrufvarianten sogar sehr erhebliche Unterschiede.

Dies lässt sich darauf zurückführen, dass die Anzahl der MEMs bei sehr niedrigen Mindestlängen teilweise enorm hoch ist, die Zahlen für größere Mindestlängen aber viel niedriger sind und sich vergleichsweise nur wenig voneinander unterscheiden.[553] Da eine sehr hohe Zahl von MEMs also den Zeitbedarf sehr nachteilig beeinflussen kann und vor allem da sich die Frage stellt, welche Mindestlänge bei der MEM-Ermittlung angesetzt werden sollte, um zu möglichst aussagekräftigen Ergebnissen zu kommen, sollen in diesem Unterkapitel entsprechende Untersuchungsdaten vorgestellt werden.

Die eigentlichen Messdaten für verschiedene Codierungen und Mindestlängen finden sich in den Tabellen 3.6 und 3.7 auf S. 156 und 157. Aus Platzgründen enthalten die Tabellen nur Angaben zu den durch 6 teilbaren Mindestlängen zwischen 18 und 66 und außerdem teilweise für die Mindestlänge 12 oder 15. Die Abweichungen am unteren Rand der Messreihen erklären sich durch den sehr starken, aber je nach Codierung unterschiedlich verlaufenden Anstieg der Zahlen bei sehr niedrigen Mindestlängen.

Bei der Ermittlung der MEMs wird fast jedes Match doppelt verzeichnet, nämlich als Übereinstimmung einer Textstelle *a* mit einer Textstelle *b* sowie als Übereinstimmung von *b* mit *a*. Die einzige Ausnahme bilden diejenigen Matches, die sich daraus ergeben, dass jeder Text mit sich selbst natürlich vollständig übereinstimmt. Hier wie auch im Folgenden sind diese textinternen Gesamtmatches nicht berücksichtigt, die übrigen Matches werden aber auch hier doppelt gezählt. Dies hat praktische Gründe – die doppelte Verzeichnung der Matchpositionen hat den Vorteil, dass sich für jeden Text leicht ein Überblick über die ihm zugeordneten Matches gewinnen lässt, und es wäre fehleranfällig, bei Auswertungen zu den Matchdaten jeweils die zweite Auflistung eines Matchs zu übergehen.

[553] Dass sich sehr kurze Mindestlängen auf die Laufzeit (und zugleich auf die Aussagekraft der Matches aufgrund einer hohen Zahl irrelevanter Übereinstimmungen) auswirken, wird schon im Handbuch von *MUMmer* erwähnt – für DNA-Daten sind Werte unter etwa 15 in dieser Hinsicht problematisch (*http://mummer.sourceforge.net/manual/#mummer*). Der starke Anstieg des Zeitbedarfs vor allem bei *essaMEM*, aber auch bei *sparseMEM* bei einem größeren Ausdünnungsfaktor und zugleich einer niedrigen Mindestlänge im Vergleich zu anderen Einstellungen erklärt sich in ähnlicher Weise dadurch, dass bei Zugrundelegung einer ausgedünnten Datenstruktur zunächst einmal auch Übereinstimmungen ermittelt werden, die noch kürzer sind als die eigentlich angesetzte Mindestlänge und dementsprechend noch wesentlich zahlreicher, und anschließend das zu ihnen gehörende Umfeld überprüft werden muss.

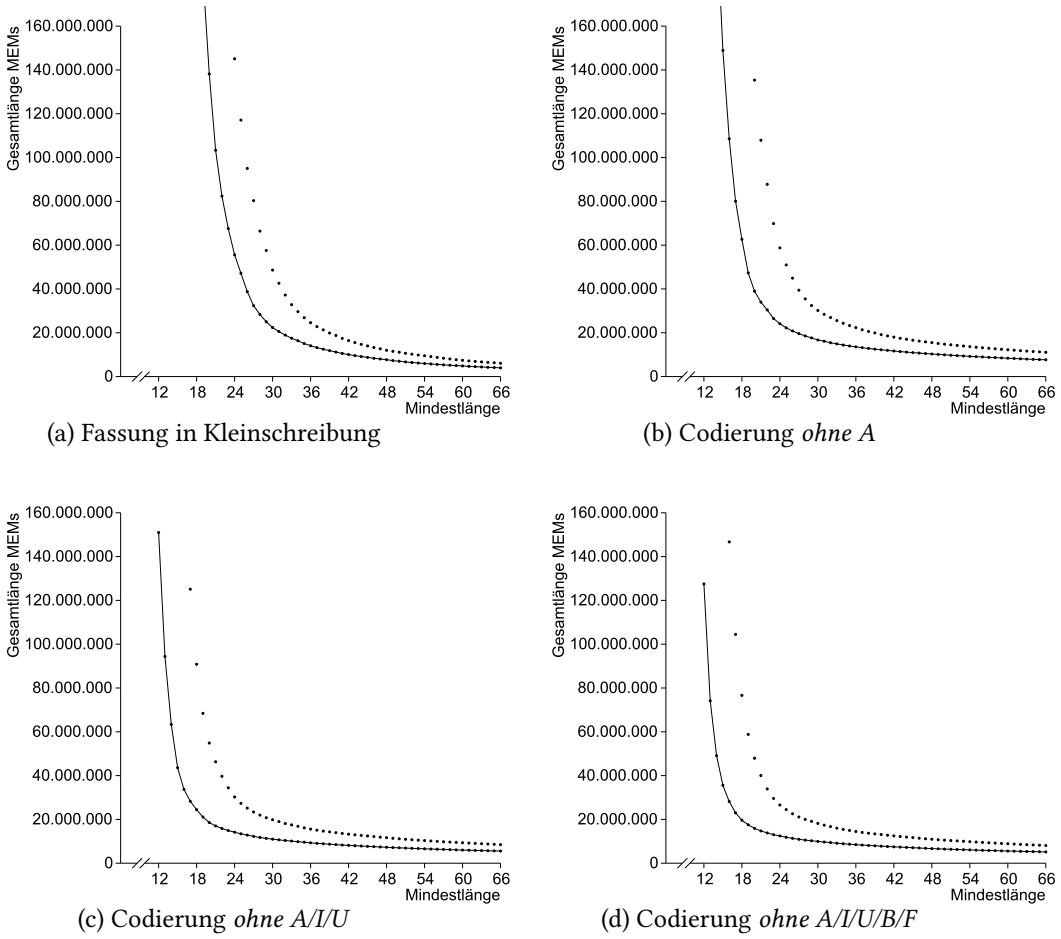
Da die in den Tabellen enthaltenen Daten nicht allzu gut zu überblicken und in ihren Abhängigkeitsbeziehungen zu durchschauen sind, sollen hier einige Beziehungen in Diagrammform präsentiert werden. Den Diagrammen liegen dabei, um ein präziseres Bild zu liefern, zusätzlich auch Messdaten für die übrigen ganzen Zahlen zwischen der für eine Codierung angesetzten niedrigsten Mindestlänge und der Obergrenze 66 zugrunde. Die Varianten mit und ohne Leerzeichen sind jeweils in einem Diagramm zusammengefasst, wobei die durch eine Linie verbundenen Punkte für die Fassung ohne Leerzeichen stehen und die nicht verbundenen Punkte für die Fassung mit Leerzeichen.[554]

Zunächst soll die Beziehung zwischen der Mindestmatchlänge und der Gesamtlänge der MEMs näher betrachtet werden. In Abbildung 3.3 wird auf der y-Achse der Diagramme nicht der gesamte Wertebereich abgebildet, sondern nur ein Ausschnitt gezeigt, um in diesem Bereich eine differenziertere Darstellung zu ermöglichen. Es ist deutlich erkennbar, dass die Werteentwicklung in den Grundzügen jeweils sehr ähnlich verläuft: Die Verbindungslinie zwischen den Messpunkten für Codierungen ohne Leerzeichen weist bei den niedrigsten Mindestlängen ein sehr starkes Gefälle auf und flacht dann immer weiter ab. Entsprechendes gilt auch für eine gedachte Verbindungslinie zwischen den Messpunkten für die Codierungen mit Leerzeichen.

Unterschiede zwischen den verschiedenen Codierungen zeigen sich zum Beispiel daran, welchen Bereich auf der x-Achse die gezeigten Werte überhaupt abdecken. Bei der Mehrzahl der Codierungen ist der niedrigste x-Wert nicht der in den Tabellen dokumentierte, da der entsprechende y-Wert jenseits der Obergrenze des Ausschnitts liegt. Nur bei *ohne A/I/U* ohne Leerzeichen und *ohne A/I/U/B/F* ohne Leerzeichen setzen die Datenreihen erst bei einer Mindestlänge ein, für die die Gesamtlänge der MEMs unter dieser Obergrenze liegt. Auch wenn die Daten also nicht vollständig vergleichbar sind, ist doch deutlich zu erkennen, dass sich in in dieser Weise codierten Texten kürzere MEMs ermitteln lassen als bei Zugrundelegung einer anderen Codierung, ohne den hier gezeigten Wertebereich auf der y-Achse zu überschreiten, und dass die Fassung in Kleinbuchstaben mit Leerzeichen beim im Vergleich höchsten x-Wert diese Grenze erreicht.

Weiter lässt sich erkennen, dass die eine Codierungsvariante mit Leerzeichen repräsentierenden Einzelpunkte jeweils vollständig oberhalb der Linie für die entsprechende Codierungsvariante ohne Leerzeichen liegen. Bei einer Zusammenfassung der verschiedenen Codierungsvarianten mit Leerzeichen in einem Diagramm und Hinzufügung von Verbindungslinien zwischen den zu jeweils einer Variante gehörenden Punkten würde es hingegen zu Linienüberschneidungen kommen. Insbesondere ist die Gesamtlänge der MEMs in der Fassung in Kleinbuchstaben mit

[554] Dass trotz der Darstellung in Linienform auch bei der Fassung ohne Leerzeichen zwischen den ganzzahligen Mindestlängen keine Zwischenwerte zu bilden sind, ist wohl offensichtlich.



(a) Fassung in Kleinschreibung

(b) Codierung *ohne A*

(c) Codierung *ohne A/I/U*

(d) Codierung *ohne A/I/U/B/F*

Abb. 3.3: MEM-Gesamtlänge bei verschiedenen Codierungen und Mindestlängen (durch eine Linie verbundene Punkte: Codierungen ohne Leerzeichen, unverbundene Punkte: Codierungen mit Leerzeichen)

Leerzeichen bei den niedrigeren hier untersuchten Mindestlängen jeweils am größten, bei den größeren Mindestlängen hingegen (wenn man nur die Fassungen mit Leerzeichen betrachtet) am niedrigsten.[555] Entsprechendes gilt für einen Vergleich der verschiedenen Varianten ohne Leerzeichen.[556]

Der zuletzt beschriebene Befund mag zunächst irritieren, denn das Codierungsverfahren soll ja auch für niedrigere Mindestlängen zu einer Ermittlung von Übereinstimmungen führen, die bei einem Vergleich auf der Basis der originalen bezie-

[555] Nach den den Diagrammen zugrunde liegenden Daten überschreitet der Wert der Fassung in Kleinbuchstaben mit Leerzeichen den der Codierung *ohne A* mit Leerzeichen ab der Mindestlänge 39, den der Codierung *ohne A/I/U* mit Leerzeichen ab der Mindestlänge 50 und den der Codierung *ohne A/I/U/B/F* mit Leerzeichen ab der Mindestlänge 53.

[556] Hier sind die Mindestlängen, ab denen die MEM-Gesamtlänge der Fassung in Kleinbuchstaben den entsprechenden Wert für eine codierte Textfassung übersteigt, teilweise etwas größer, nämlich 38, 50 beziehungsweise 54.

hungsweise in Kleinbuchstaben umgewandelten Texte nicht gefunden würden. Er erklärt sich dadurch, dass die codierten Texte wesentlich kürzer sind, so dass eine bestimmte Mindestlänge einer größeren Textmenge des Originals entspricht als dieselbe Länge in der Fassung in Kleinbuchstaben.

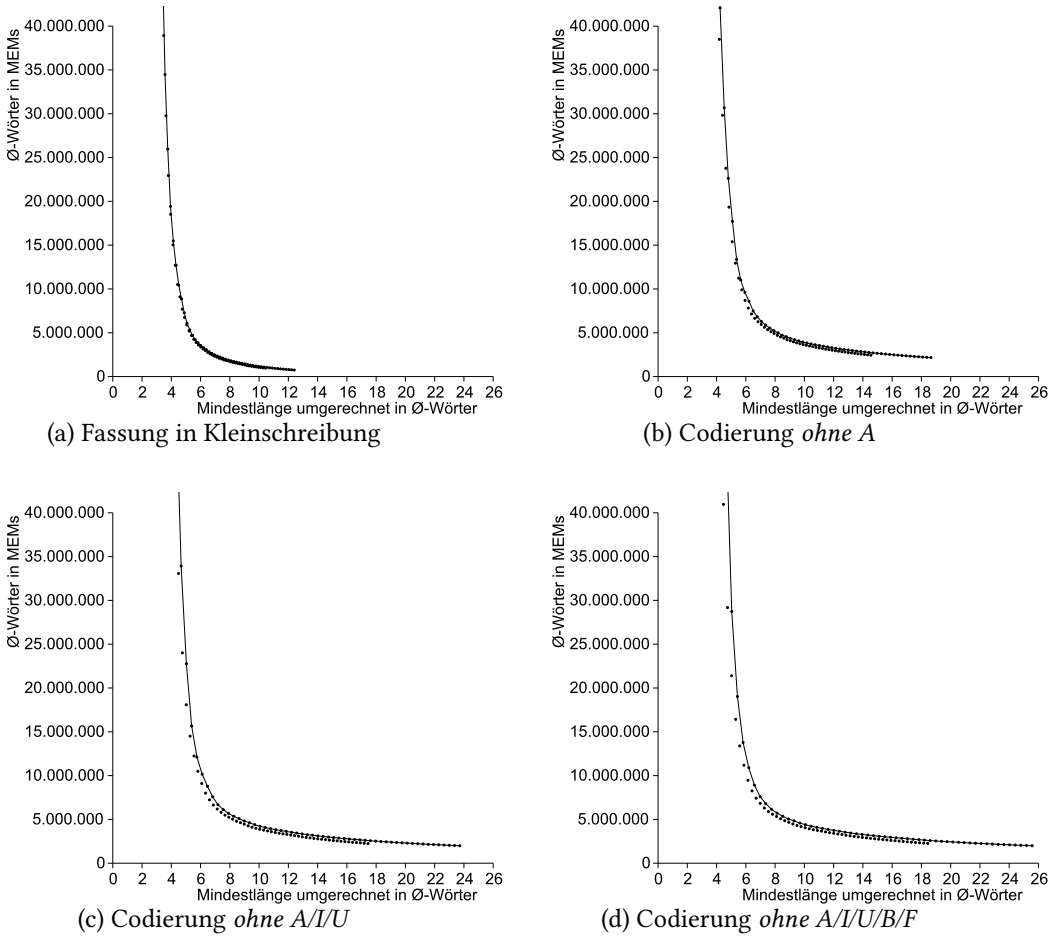
Um die Leistungsfähigkeit der verschiedenen Codierungsvarianten unter diesem Aspekt besser miteinander vergleichen zu können, lässt sich eine Umrechnung vornehmen, die das unterschiedliche Maß an Textverdichtung zugrunde legt. Abbildung 3.4 bereitet die in Abbildung 3.3 gezeigten Daten entsprechend auf. Hier wie auch im Folgenden wird dabei als fiktive, aber hoffentlich einigermaßen anschauliche Maßeinheit ein „Wort durchschnittlicher Länge“ (oder vereinfacht „Durchschnittswort“) gewählt, das dem Zahlenwert entspricht, der sich ergibt, wenn man die Gesamtzahl der Zeichen in der jeweiligen Codierung durch die Zahl der in den zugrunde liegenden Texten enthaltenen laufenden Wortformen teilt. Wie bereits in Abbildung 3.3 wird nicht der gesamte Wertebereich auf der y-Achse gezeigt, um für die niedrigeren Werte eine differenziertere Darstellung zu ermöglichen.

Auch hier sollen einige Beobachtungen und Überlegungen festgehalten werden. In allen vier Diagrammen liegt die gedachte Verbindungslinie der Messpunkte für die Codierungsvariante mit Leerzeichen etwas unterhalb der Linie für die Variante ohne Leerzeichen – für die Fassungen in Kleinschreibung mit und ohne Leerzeichen sind die Unterschiede der Werte allerdings so gering, dass sich das im Diagramm kaum erkennen lässt.

Ob eine Codierungsvariante mit oder ohne Leerzeichen zugrunde gelegt wird, wirkt sich offenbar vor allem bei den Codierungen *ohne A/I/U/B/F* und *ohne A/I/U* aus, aber auch bei ihnen sind die relativen Unterschiede bei Mindestlängen im mittleren Bereich nicht sehr groß. Je größer die einander entsprechenden Mindestlängen sind, desto geringer fällt die absolute Differenz aus, bei größeren Mindestlängen nimmt die in Durchschnittswörter umgerechnete Gesamtlänge der MEMs aber bei Zugrundelegung einer Codierungsvariante mit Leerzeichen etwas stärker ab als in der Variante ohne Leerzeichen. Bei Ansetzung von sehr niedrigen Mindestlängen sind die relativen Unterschiede ebenfalls größer.[557]

Da die Wortabgrenzung im hier untersuchten Korpus weitgehend einheitlich ist und in der Regel nur in bestimmten Fällen (insbesondere bei Komposita und präfigierten Wörtern) Unterschiede aufweist, ist die Wahrscheinlichkeit, dass eine Übereinstimmung bei Verwendung einer Codierungsvariante mit Leerzeichen aufgrund einer solchen Abweichung nicht gefunden wird, auch bei etwas längeren

[557] Zum Beispiel entspricht die Gesamtmatchlänge bei Ansetzung einer Mindestlänge von umgerechnet etwa fünf Durchschnittswörtern in der Codierung *ohne A/I/U/B/F* mit Leerzeichen etwas weniger als 75 % des in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen erreichten Werts. Bei Mindestlängen, die in etwa sieben Wörtern durchschnittlicher Länge entsprechen, haben die Werte dieser beiden Codierungsvarianten ein Verhältnis von fast 90 %, und bei knapp 35 Durchschnittswörtern liegt es bei etwa 78 %.



(a) Fassung in Kleinschreibung

(b) Codierung *ohne A*

(c) Codierung *ohne A/I/U*

(d) Codierung *ohne A/I/U/B/F*

Abb. 3.4: MEM-Gesamtlänge umgerechnet in Zahl von Wörtern mit durchschnittlicher Länge (durch eine Linie verbundene Punkte: Codierungen ohne Leerzeichen, unverbundene Punkte: Codierungen mit Leerzeichen)

Wortfolgen nicht sehr groß und nimmt nur allmählich zu – das passt gut zum hier gezeigten Befund. Dass es für die Codierungen *ohne A/I/U* und *ohne A/I/U/B/F* für einander nach der Umrechnung in Durchschnittswörter entsprechende niedrige Mindestlängen erhebliche Unterschiede zwischen den Varianten mit und ohne Leerzeichen gibt, deutet wohl darauf hin, dass es insbesondere in den Varianten ohne Leerzeichen bei niedrigen Mindestlängen auch zu einem nicht ganz geringen Anteil von Fehlzuordnungen kommt, weil auch ganz unterschiedliche Wörter und kurze Wortfolgen auf übereinstimmende Codezeichenfolgen abgebildet werden. Die Zugrundelegung einer Codierung ohne Leerzeichen kann also zwar den *Recall* des Vergleichsergebnisses verbessern, das geht aber für kürzere Übereinstimmungen wohl mit einer merklichen Verringerung der *Precision* einher.[558]

[558] Eine Möglichkeit, wie die Wortabgrenzung in einem Nachbearbeitungsschritt zur Aussortierung zumindest eines Teils der falschen Zuordnungen genutzt werden kann, wird unten in Unterkapitel 3.3.1 vorgestellt.

Wenn man die vier Diagramme in Abbildung 3.4 miteinander vergleicht, ist deutlich zu erkennen, dass die in Diagramm 3.4a verzeichnete Gesamtlänge der Matches bei Umrechnung in Wörter durchschnittlicher Länge jedenfalls für größere Mindestlängen deutlich niedriger ausfällt als in den Diagrammen 3.4b–3.4d. Hier zeigt sich der große Einfluss, den eine Codierung entsprechend Unterkapitel 3.1.2 auf die Ermittlung von Übereinstimmungen im Untersuchungskorpus hat.

Natürlich lassen sich auch für sehr kurze Wortfolgen auf der Basis einer solchen Codierung mehr Matches finden als auf der Basis einer Fassung in Kleinbuchstaben. Da hier aber nur ein Ausschnitt der Daten gezeigt wird, sind in den Diagrammen 3.4b–3.4d nur die Werte für Mindestlängen verzeichnet, die zumindest etwas größer sind als vier Wörter durchschnittlicher Länge. In Diagramm 3.4a finden sich hingegen auch Werte für eine Länge von etwas weniger als vier Durchschnittswörtern. Entsprechendes gilt auch bei Einbeziehung aller Daten, die in den Tabellen 3.6 und 3.7 dokumentiert sind, da auch darin die Datenreihen für die Codierungen *ohne A*, *ohne A/I/U* und *ohne A/I/U/B/F* mit und ohne Leerzeichen mit Mindestlängen beginnen, die einer größeren Zahl von Durchschnittswörtern entsprechen als die dort für die Fassungen in Kleinbuchstaben mit und ohne Leerzeichen angesetzten niedrigsten Mindestlängen.[559]

Die für eine Codierung in den Tabellen verzeichneten Daten setzen jeweils bei einer Mindestlänge ein, bei der die Gesamtlänge der MEMs viel größer als die Gesamtlänge der zugrunde liegenden codierten Texte ist. Das dabei erreichte Verhältnis dieser beiden Werte ist zwar je nach Codierung unterschiedlich,[560] es sollte aber wohl ersichtlich sein, dass dabei jeweils ein Punkt erreicht ist, bei dem ein Großteil der Matches jedenfalls für die hier untersuchte Fragestellung wenig aussagekräftig ist und allenfalls in Verbindung mit zusätzlichen Verarbeitungskriterien sinnvoll ausgewertet werden kann. Das gilt nach den im Folgenden noch vorgebrachten Überlegungen auch für die niedrigsten Mindestlängen, die in den Diagrammen der Abbildungen 3.3 und 3.4 berücksichtigt sind.

Die Tabellen 3.6 und 3.7 enthalten auch Angaben zur Anzahl der MEMs sowie zur Anzahl und Gesamtlänge der textinternen MEMs, also der Übereinstimmungen, bei denen beide einander entsprechenden Stellen im selben Text zu finden sind. Die Angaben zu textinternen MEMs haben zum einen dokumentarische Funktion: Es soll verdeutlicht werden, dass auch solche Übereinstimmungen hier mit eingerechnet sind, obwohl diese für die Frage nach textuellen Beziehungen in aller Regel selbst nicht aussagekräftig sind,[561] und in welchem Maße sich die Zahlen reduzieren, wenn man solche Übereinstimmungen nicht einbezieht. Sie werden

[559] Diese Zahl liegt für die jeweils niedrigsten in der Tabelle für eine Codierung verzeichneten Mindestlängen bei 2,85 beziehungsweise 2,82 für die Fassungen in Kleinbuchstaben mit beziehungsweise ohne Leerzeichen und ansonsten zwischen 3,3 (für die Codierung *ohne A* mit Leerzeichen) und 4,65 (für die Codierung *ohne A/I/U/B/F* ohne Leerzeichen).

[560] Es ist in einer Spalte der Tabellen 3.6 und 3.7 dokumentiert.

hier nicht einfach übergangen, weil sie möglicherweise für andere Fragestellungen von Interesse sein könnten.

Zum anderen hat sich bei der Datenanalyse gezeigt, dass es eine auffällige Werteentwicklung des Anteils solcher textinternen MEMs an der Gesamtzahl der MEMs in Abhängigkeit von der angesetzten Mindestlänge gibt. Dieser Anteil ist nämlich in allen hier zugrunde gelegten Codierungen bei den niedrigsten untersuchten Mindestlängen vergleichsweise gering, steigt dann bei Erhöhung der Mindestlänge bis zu einem bestimmten Punkt zunächst recht stark an und sinkt bei weiterer Erhöhung der Mindestlänge wieder ab. Abbildung 3.5 enthält die entsprechenden Diagramme für die verschiedenen Codierungen.

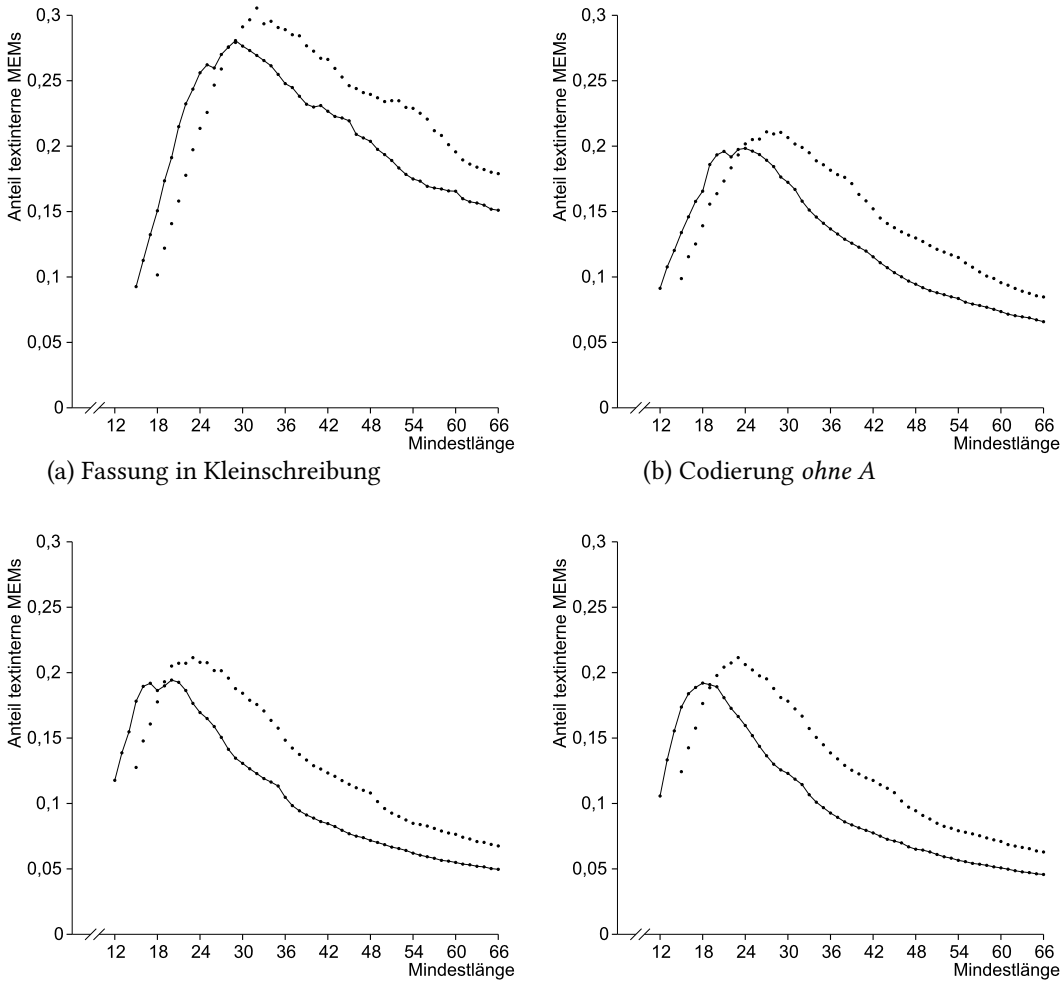
Dass sich in den Diagrammen kleinere Abweichungen von der eben beschriebenen Grundtendenz feststellen lassen, braucht als normales statistisches Phänomen wohl nicht weiter untersucht zu werden. Die Grundtendenz lässt sich vermutlich dadurch erklären, dass verschiedene Arten der Übereinstimmung je nach angesetzter Mindestlänge unterschiedlich häufig vorkommen und dass bei einer mehrfach vorkommenden Zeichenfolgen jede beliebige Kombination zweier Vorkommen als Match zählt und deshalb die Gesamtzahl der Matches von sehr häufigen Zeichenfolgen besonders stark beeinflusst wird.

Sowohl bei sehr kurzen als auch bei sehr langen Übereinstimmungen gibt es gute Gründe, warum der Anteil der textinternen Matches an der Gesamtzahl relativ gering ist: Kurze Übereinstimmungen beruhen vielfach auf Wortfolgen, die sehr häufig vorkommen. Dabei handelt es sich insbesondere um feste sprachliche Muster, aber auch – vor allem bei sehr kurzen Matches – um Folgen von häufigen Wörtern (beziehungsweise von den entsprechenden Codezeichenfolgen). Selbst wenn solche insgesamt sehr häufigen Wortfolgen in einem Einzeltext überdurchschnittlich oft vertreten sind, führt die Tatsache, dass jede beliebige Kombination zweier Vorkommen ein Match bildet, dazu, dass die textinternen Matches mit hoher Wahrscheinlichkeit nur einen kleinen Teil ausmachen.

Bei sehr langen Übereinstimmungen hingegen ist im Regelfall davon auszugehen, dass sie auf der Verwendung von Vorlagen beruhen. Auch innerhalb der einzelnen Texte des hier untersuchten Textkorpus finden sich (insbesondere bei Zugrundelegung einer codierten Textfassung, aber auch in einer Fassung in Kleinbuchstaben) erstaunlich häufig wiederholte Vorkommen längerer Zeichenfolgen, aber gegenüber Matches, die auf umfangreiche textuelle Übernahmen zurückzuführen sind, fallen sie jedenfalls in diesem Korpus nicht allzu sehr ins Gewicht.

Dass der Anteil textinterner MEMs bei Mindestmatchlängen, die weder sehr niedrig noch sehr hoch sind, größer ist, lässt sich vermutlich dadurch erklären, dass nur wenige Formulierungen mit einer solchen Länge sehr häufig vorkommen,

[561] Eine Ausnahme kann bei Texten vorliegen, die selbst verschiedene Einzeltexte enthalten, da diese Einzeltexte jeweils für sich genommen auch zu anderen Texten innerhalb dieser Sammlung in einer Traditionsbeziehung stehen können.



(a) Fassung in Kleinschreibung

(b) Codierung *ohne A*

(c) Codierung *ohne A/I/U*

(d) Codierung *ohne A/I/U/B/F*

Abb. 3.5: Anteil textinterner Matches an der Gesamtzahl der Matches (durch eine Linie verbundene Punkte: Codierungen ohne Leerzeichen, unverbundene Punkte: Codierungen mit Leerzeichen)

aber sprachliche Präferenzen des jeweiligen Verfassers und der jeweils behandelte Gegenstand zur wiederholten Verwendung übereinstimmender Wortfolgen führen können. Nach dieser Erklärung können Übereinstimmungen mittlerer Länge also mit einer gewissen Wahrscheinlichkeit auf sprachliche Eigentümlichkeiten und inhaltliche Beziehungen zurückgeführt werden.

Man kann daraus wohl die Folgerung ziehen, dass bei Unterschreiten der Mindestlängen, bei denen der Anteil textinterner MEMs besonders hoch ist, in einem zunehmenden Maße mit Matches gerechnet werden muss, die auf sehr häufigen Formulierungen beziehungsweise auf sehr häufigen Folgen von Codezeichen beruhen und deshalb wenig aussagekräftig sind.

Wie gut sich der Befund auf andere Korpora übertragen lässt, kann hier nicht in größerem Rahmen untersucht, aber doch anhand der Daten für ein ganz anders

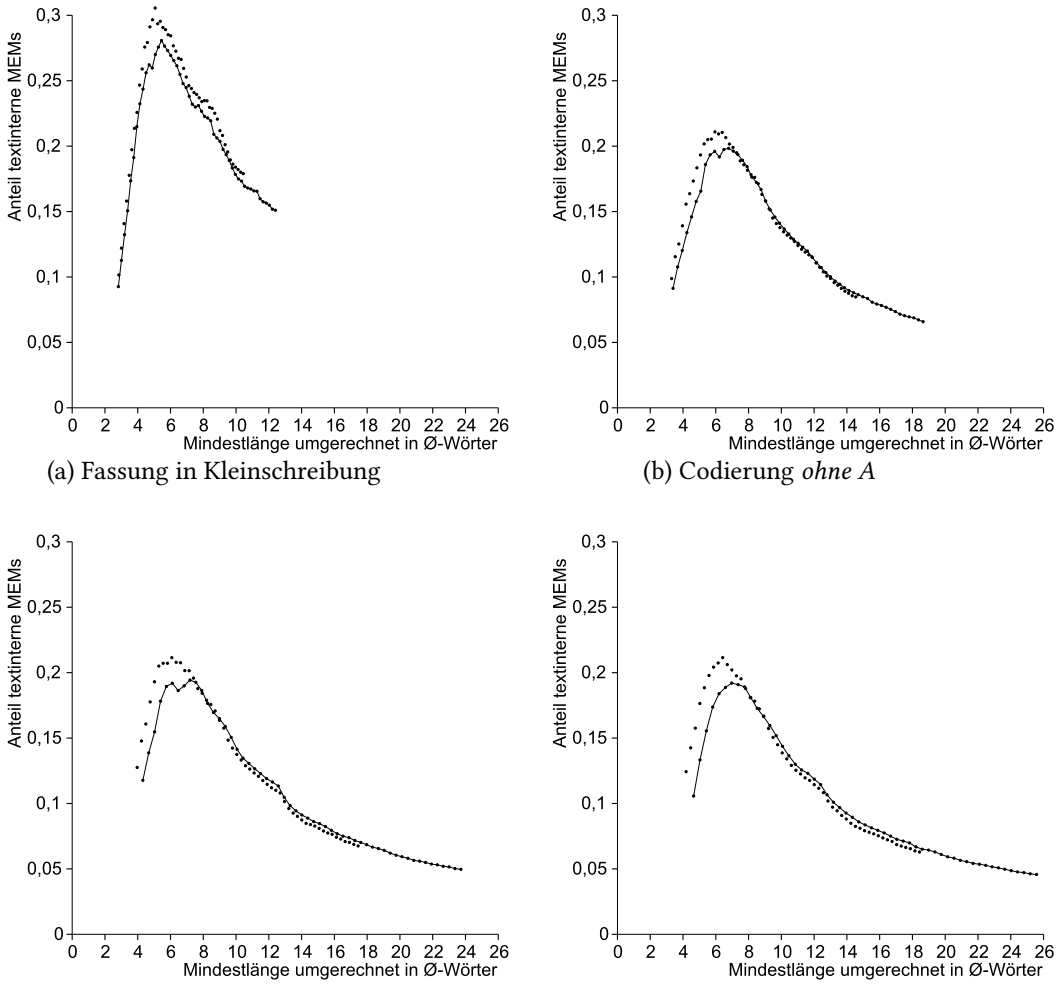
zusammengesetztes Korpus überlegt werden. Voraussetzung für eine Entwicklung des Anteils textinterner MEMs von relativ niedrigen zu deutlich höheren und dann wieder niedrigeren Werten ist natürlich, dass auch bei Ansetzung einer höheren Mindestmatchlänge ein zumindest erheblicher Teil der Matches auf Übereinstimmungen zwischen unterschiedlichen Texten beruht. Wenn das nicht der Fall ist, es aber etwas längere Matches innerhalb der einzelnen Texte gibt (was eine vermutlich häufige Konstellation ist), ergibt sich natürlich keine Werteverteilung wie in den Diagrammen der Abbildung 3.5, da ab einer bestimmten Mindestlänge der Anteil der textinternen MEMs mehr oder weniger konstant hoch bleibt. Auch in einem solchen Fall ist der Anteil textinterner Übereinstimmungen für niedrige Mindestlängen aber vermutlich deutlich niedriger, sofern das Korpus eine Reihe von sprachlich einander einigermaßen nahestehenden Texten enthält. Das ist jedenfalls der Befund für das Auswahlkorpus literarischer Texte, das oben schon als Grundlage für die Tabelle 3.3 gedient hat,[562] wobei hier allerdings alle darin enthaltenen Werke eines Autors zusammen als textuelle Einheit behandelt wurden; es wurde also nicht der Anteil der im engeren Sinne textinternen MEMs untersucht, sondern der Anteil der dateiinternen MEMs.[563]

Neben der allgemein ähnlichen Grundtendenz der Werteentwicklung ist beim Vergleich der Diagramme wohl insbesondere auffällig, dass der Anteil textinterner MEMs bei den Fassungen in Kleinbuchstaben mit und ohne Leerzeichen mit Abstand am höchsten ist. Das ist leicht erklärlich: Innerhalb eines Textes sind die Schreibungen natürlich wesentlich weniger unterschiedlich als im gesamten Korpus, und dementsprechend ist die Wahrscheinlichkeit geringer, dass eine wörtliche Übereinstimmung aufgrund eines Schreibungsunterschieds nicht gefunden wird. Bei Zugrundelegung einer Codierung, die solche Unterschiede nivelliert, fallen abweichende Schreibungen viel weniger ins Gewicht, und dementsprechend sinkt der Anteil der textinternen Matches.

Auch im Hinblick auf den Anteil textinterner MEMs ist es vielleicht von Interesse, die Mindestlängen umzurechnen in Wörter durchschnittlicher Länge, um einen

[562] Als Beispiel sollen die Fassung in Kleinbuchstaben mit Leerzeichen und die Codierung *ohne A/I/U/B/F* ohne Leerzeichen dienen. In beiden Fällen ergibt sich mit zunehmender Mindestlänge ein zunächst starker Anstieg der Messwerte, der im weiteren Verlauf immer schwächer ausfällt. Für die Codierung *ohne A/I/U/B/F* ohne Leerzeichen ist der Anstieg wesentlich stärker ausgeprägt; bei ihr liegt der Anteil der MEMs innerhalb der Werke der einzelnen Autoren schon bei einer Mindestlänge von 19 Zeichen über 90 % der Gesamtzahl, bei der Fassung in Kleinbuchstaben wird ein Anteil dieser Höhe bei 39 Zeichen erreicht. Ab einer Länge von 28 beziehungsweise 75 Zeichen finden sich in diesem Auswahlkorpus keine MEMs, bei denen die zugehörigen Stellen aus Werken unterschiedlicher Verfasser stammen.

[563] Dies ergab sich zunächst einmal daraus, dass die in den Dateien enthaltenen Texte hier nicht näher untersucht, sondern nur zu Vergleichszwecken herangezogen wurden. Die Frage, was als textuelle Einheit zu behandeln ist, stellt sich aber generell bei Textsammlungen und kann je nach Fragestellung unterschiedlich beantwortet werden.



(a) Fassung in Kleinschreibung

(b) Codierung *ohne A*

(c) Codierung *ohne A/I/U*

(d) Codierung *ohne A/I/U/B/F*

Abb. 3.6: Anteil textinterner Matches an der Gesamtzahl der Matches (durch eine Linie verbundene Punkte: Codierungen ohne Leerzeichen, unverbundene Punkte: Codierungen mit Leerzeichen)

einheitlichen Maßstab für den Vergleich der verschiedenen Codierungen zu erhalten. Die entsprechenden Diagramme finden sich in Abbildung 3.6.

Hier lässt sich erkennen, dass die Mindestlängen, bei denen die textinternen MEMs den höchsten Anteil haben, Werten zwischen etwa fünf Durchschnittswörtern (in der Fassung in Kleinbuchstaben mit Leerzeichen) und etwas mehr als sieben Durchschnittswörtern (in der Fassung *ohne A/I/U* ohne Leerzeichen) entsprechen. Da sich in den Diagrammen 3.6b und 3.6c (und entsprechend auch in den Diagrammen 3.5b und 3.5c) für die Codierungsvarianten ohne Leerzeichen links neben den Maximalwerten etwas niedrigere Werte finden und links davon wiederum Werte, die dem jeweiligen Maximalwert nahekommen, und da der Maximalwert bei der Codierung *ohne A/I/U/B/F* ohne Leerzeichen bei einer geringfügig niedrigeren Zahl von Wörtern durchschnittlicher Länge liegt als bei der Codierung *ohne A/I/U* ohne

Leerzeichen, obwohl erstere zu einer noch stärkeren Vereinheitlichung der Zeichenfolgen führt, sei nochmals betont, dass bei Messungen dieser Art eine gewisse Schwankungsbreite der Daten zu berücksichtigen ist.

Jedenfalls kann festgehalten werden, dass der Wert für die Fassung in Kleinbuchstaben mit Leerzeichen zwar am niedrigsten ist, aber auch für Codierungen mit starker Reduzierung der Varianz nicht allzu viel höher. Wenn die Überlegung stimmt, dass der Übergang von einer steigenden zu einer fallenden Entwicklung des Anteils textinterner MEMs ein Indiz für einen bei niedrigeren Mindestlängen stark zunehmenden Anteil von MEMs ist, die im Sinne der hier untersuchten Fragestellung unergiebig sind, kann wohl gefolgert werden, dass der Vergleich auf der Basis einer Textfassung in Kleinbuchstaben für Übereinstimmungen von etwa fünf Wörtern durchschnittlicher Länge deutliche Vorteile im Hinblick auf die Aussagekraft der Matches hat, dass ab Mindestlängen von je nach Codierung etwa sechs bis sieben Wörtern durchschnittlicher Länge aber auch codierte Textfassungen gut als Vergleichsgrundlage genutzt werden können.

Die Zeilen der Tabellen 3.6 und 3.7 enthalten in den letzten beiden Spalten auch zwei Werte, die die MEMs in Beziehung zum Gesamtumfang der codierten Texte setzen, nämlich zum einen das Verhältnis von MEM-Gesamtlänge und Textgesamtlänge, zum anderen den Anteil der codierten Textzeichen, die einem oder auch mehren MEMs zugeordnet sind (in den Tabellen als „durch MEMs abgedeckter Textanteil“ bezeichnet). Diese Zahlen sollen einen deutlicheren Eindruck vom Ausmaß der Matches und davon vermitteln, dass die MEM-Gesamtlänge (und -anzahl) bei Einbeziehung kürzerer Matchlängen deshalb so stark ansteigt, weil die Zeichenfolgen dieser Matches vielfach eine höhere Vorkommenshäufigkeit haben.

Wenn man den für eine bestimmte Codierung und Mindestlänge verzeichneten Wert der vorletzten Spalte durch den der letzten Spalte teilt, ergibt sich, einer wie großen Zahl von Matches ein Zeichen, das im Bereich zumindest eines Matchs liegt, im Durchschnitt zugeordnet ist.

Es ist nicht erstaunlich, dass das Verhältnis von MEM-Gesamtlänge zu Textgesamtlänge auch bei den größten hier untersuchten Mindestlängen deutlich über dem Anteil der den MEMs zugeordneten Zeichen liegt, da das Korpus Gruppen von recht eng verwandten Texten enthält und deshalb auch etwas längere Textstücke nicht ganz selten mehrere Parallelen haben. Dass die Werte in den letzten beiden Spalten in den Fassungen in Kleinbuchstaben mit und ohne Leerzeichen bei größeren Mindestlängen wesentlich näher beieinander liegen als bei den die Schreibungen vereinheitlichenden Codierungen und jeweils für sich hinter den entsprechenden Werten für die Codierungen zurückbleiben, deutet – wie viele andere Befunde in den Untersuchungsdaten – darauf hin, dass bei Verzicht auf eine Schreibungsvereinheitlichung zahlreiche längere Übereinstimmungen nicht

gefunden werden, und diese sind im Sinne der hier untersuchten Fragestellung von besonderem Interesse.

Bei den niedrigsten untersuchten Mindestlängen hingegen erreichen die beiden Werte sowohl für sich als auch im Verhältnis zueinander jeweils eine Höhe, die die Vermutung nahelegt, dass Matches dieser Länge zu einem großen Teil nicht auf textuelle Abhängigkeitsverhältnisse zurückzuführen sind.[564]

[564] Diese Aussage lässt sich natürlich nicht einfach verallgemeinern, da die Werte vom Ausmaß der Textbeziehungen im untersuchten Korpus abhängen. Wenn ein Korpus zum Beispiel aus Texten einer bestimmten Textfamilie bestünde, wären hohe Werte durchaus plausibel. In einem solchen Fall würden aber wohl auch bei größeren Mindestlängen wesentlich höhere Werte verzeichnet als hier in den entsprechenden Tabellenzeilen.

Mindestlänge	Anzahl MEMs	Anzahl text-interne MEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
18	44.390.820	4.508.266	905.793.168	21,449	0,728
24	4.994.522	1.066.752	145.095.350	3,436	0,453
30	1.239.914	361.022	48.618.026	1,151	0,315
36	485.516	140.342	24.597.266	0,582	0,244
42	271.054	72.178	16.417.356	0,389	0,200
48	170.996	40.964	11.991.338	0,284	0,169
54	121.146	27.730	9.484.186	0,225	0,145
60	84.606	16.550	7.426.938	0,176	0,125
66	62.978	11.272	6.080.320	0,144	0,109

(a) Fassung in Kleinbuchstaben mit Leerzeichen

Mindestlänge	Anzahl MEMs	Anzahl text-interneMEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
15	35.807.526	3.536.872	627.028.246	20,681	0,772
18	10.706.430	1.489.638	234.333.590	7,729	0,594
24	1.694.160	341.714	58.748.330	1,938	0,413
30	584.106	120.642	30.180.140	0,995	0,346
36	339.630	61.676	22.342.458	0,737	0,311
42	226.358	34.424	18.025.704	0,595	0,287
48	169.096	21.948	15.495.470	0,511	0,269
54	132.524	15.228	13.656.852	0,450	0,252
60	107.428	10.274	12.243.288	0,404	0,238
66	88.770	7.522	11.079.852	0,365	0,225

(b) Codierung *ohne A* mit Leerzeichen

Mindestlänge	Anzahl MEMs	Anzahl text-interne MEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
15	14.008.564	1.786.870	249.656.724	9,880	0,661
18	3.879.586	689.206	90.832.962	3,595	0,494
24	757.376	157.472	30.254.442	1,197	0,369
30	353.280	65.096	19.845.200	0,785	0,322
36	220.670	32.744	15.576.310	0,616	0,295
42	160.002	19.730	13.259.238	0,525	0,275
48	124.280	13.424	11.676.360	0,462	0,257
54	97.800	8.296	10.348.278	0,410	0,242
60	79.958	6.112	9.344.228	0,370	0,229
66	66.066	4.462	8.476.886	0,335	0,217

(c) Codierung *ohne A/I/U* mit Leerzeichen

Mindestlänge	Anzahl MEMs	Anzahl text-interne MEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
15	12.304.942	1.529.614	218.857.208	9,149	0,662
18	3.217.714	567.656	76.634.502	3,204	0,482
24	641.304	132.226	26.583.288	1,111	0,362
30	316.180	56.338	18.180.622	0,760	0,319
36	200.910	27.856	14.480.892	0,605	0,293
42	148.600	17.480	12.481.092	0,522	0,274
48	114.204	10.784	10.956.514	0,458	0,257
54	91.626	7.252	9.821.386	0,411	0,242
60	75.246	5.340	8.898.798	0,372	0,229
66	62.172	3.912	8.083.820	0,338	0,218

(d) Codierung *ohne A/I/U/B/F* mit Leerzeichen

Tab. 3.6: MEMs in verschiedenen Codierungen mit Leerzeichen

Mindestlänge	Anzahl MEMs	Anzahl text-interne MEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
15	49.363.480	4.572.062	837.483.720	23,558	0,723
18	12.304.686	1.853.224	258.490.862	7,271	0,550
24	1.817.710	465.456	55.563.978	1,563	0,347
30	527.444	145.830	22.396.026	0,630	0,256
36	268.358	66.500	14.086.350	0,396	0,205
42	162.906	36.928	10.063.232	0,283	0,170
48	109.448	22.294	7.701.070	0,217	0,143
54	74.622	13.052	5.956.064	0,168	0,122
60	54.858	9.082	4.843.174	0,136	0,105
66	41.064	6.202	3.984.276	0,112	0,091

(a) Fassung in Kleinbuchstaben ohne Leerzeichen

Mindestlänge	Anzahl MEMs	Anzahl text-interneMEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
12	33.581.850	3.067.364	470.773.874	19,916	0,751
18	2.563.252	424.408	62.684.446	2,652	0,440
24	569.416	112.928	24.114.382	1,020	0,351
30	284.262	48.996	16.712.702	0,707	0,313
36	185.962	25.424	13.551.554	0,573	0,288
42	136.930	15.802	11.675.186	0,494	0,269
48	106.006	10.016	10.306.198	0,436	0,252
54	84.760	7.082	9.237.184	0,391	0,238
60	69.532	5.114	8.378.856	0,354	0,225
66	57.948	3.814	7.657.020	0,324	0,214

(b) Codierung *ohne A* ohne Leerzeichen

Mindestlänge	Anzahl MEMs	Anzahl text-interne MEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
12	10.562.160	1.243.218	150.978.038	8,122	0,645
18	823.538	153.460	24.441.994	1,315	0,376
24	296.204	50.212	14.162.496	0,762	0,323
30	173.648	22.690	10.977.370	0,591	0,295
36	121.894	12.764	9.306.002	0,501	0,273
42	91.572	7.748	8.147.312	0,438	0,256
48	71.988	5.162	7.279.728	0,392	0,241
54	57.964	3.596	6.574.776	0,354	0,228
60	47.842	2.628	6.003.672	0,323	0,216
66	40.538	2.014	5.547.850	0,298	0,206

(c) Codierung *ohne A/I/U* ohne Leerzeichen

Mindestlänge	Anzahl MEMs	Anzahl text-interne MEMs	MEM-Gesamtlänge	MEM-Gesamtlänge/ Text-Gesamtlänge	durch MEMs abgedeckter Textanteil
12	8.970.758	948.298	127.506.364	7,396	0,686
18	616.076	118.308	19.625.678	1,138	0,366
24	251.620	40.160	12.466.552	0,723	0,318
30	154.360	18.986	9.944.194	0,577	0,291
36	108.742	10.080	8.473.562	0,491	0,271
42	82.782	6.414	7.478.996	0,434	0,254
48	64.504	4.196	6.666.460	0,387	0,239
54	52.846	2.988	6.079.678	0,353	0,227
60	43.612	2.214	5.559.482	0,322	0,216
66	37.002	1.692	5.146.758	0,299	0,206

(d) Codierung *ohne A/I/U/B/F* ohne Leerzeichen

Tab. 3.7: MEMs in verschiedenen Codierungen ohne Leerzeichen

3.3 Überarbeitung und Bewertung von Matchdaten

Übereinstimmungslisten, die einfach sämtliche *maximal exact matches* verzeichnen, lassen sich ohne weitere Analyseschritte nur mit Einschränkungen nutzen und sind auch statistisch nur begrenzt aussagekräftig. Zwar zeigt sich – jedenfalls wenn die geforderte Mindestlänge nicht zu niedrig angesetzt wird – recht deutlich, bei welchen Texten beziehungsweise Textpaaren gehäuft Entsprechungen auftreten, die die angegebenen Kriterien erfüllen, aber bei relativ kurzen Matches auf der Basis einer Codierung mit starker Reduzierung der Varianz ist nicht unmittelbar ersichtlich, ob sie überhaupt auf einer Formulierungsübereinstimmung beruhen, und auch bei Verzicht auf eine Codierung umfassen die ermittelten Matches zunächst einmal in vielen Fällen im Randbereich auch Zeichen, die mit der wörtlichen Übereinstimmung nichts zu tun haben.

Schon aus diesem Grund liegt es nahe, die Matchdaten nicht einfach ohne weitere Aufbereitung zu verwenden, sondern jedenfalls die Informationen zur Wortabgrenzung mit einfließen zu lassen. Darüber hinaus kann man natürlich auch die Relevanz von Übereinstimmungen mehrerer vollständiger Wörter hinterfragen. Aufgrund der großen Zahl an Matches lässt sich eine Überprüfung aber sicherlich nur stichprobenartig beziehungsweise für Textpaare oder Textstellen, die von besonderem Interesse sind, durchführen. Deshalb scheint es sinnvoll, formale Kriterien zu entwickeln, anhand derer sich eine Bewertung vornehmen lässt, um eine weitere Filterung der Matchdaten vorzunehmen oder ihre Aussagekraft kritisch zu bewerten.

Unterkapitel 3.3.1 behandelt insbesondere die Frage, wie sich die Wortabgrenzung für eine Überarbeitung der Matchdaten nutzen lässt und was sich daraus für die Funde im Untersuchungskorpus ergibt, und geht daneben auch auf einige weitere Möglichkeiten einer Filterung anhand struktureller oder sprachlicher Kriterien ein. Unterkapitel 3.3.2 betrachtet verschiedene Möglichkeiten einer Bewertung von Matches auf der Basis der zugrunde liegenden Textpositionen.

3.3.1 Matchdatenüberarbeitung

Das dieser Untersuchung zugrunde liegende Verfahren einer Ermittlung von *maximal exact matches* hat zwangsläufig zur Folge, dass die gefundenen Übereinstimmungen in vielen Fällen nicht an Wortgrenzen beginnen und enden. Dass an den betreffenden Stellen in den Originaltexten keine Leerzeichen beziehungsweise Zeilenumbrüche stehen, ist im hier ausgewerteten Korpus in einem kleinen Teil der Fälle darauf zurückzuführen, dass die Wortabgrenzung durch Leerraum nicht ganz einheitlich ist beziehungsweise nicht zweifelsfrei festgestellt werden kann[565] – wegen solcher Abweichungen werden hier auch Codierungsvarianten

[565] Zum einen ist bei Komposita und Präfixbildungen mit Abweichungen bei der Leerzeichensetzung zu rechnen, zum anderen sind die Abstände zwischen Wörtern teilweise sehr gering und die

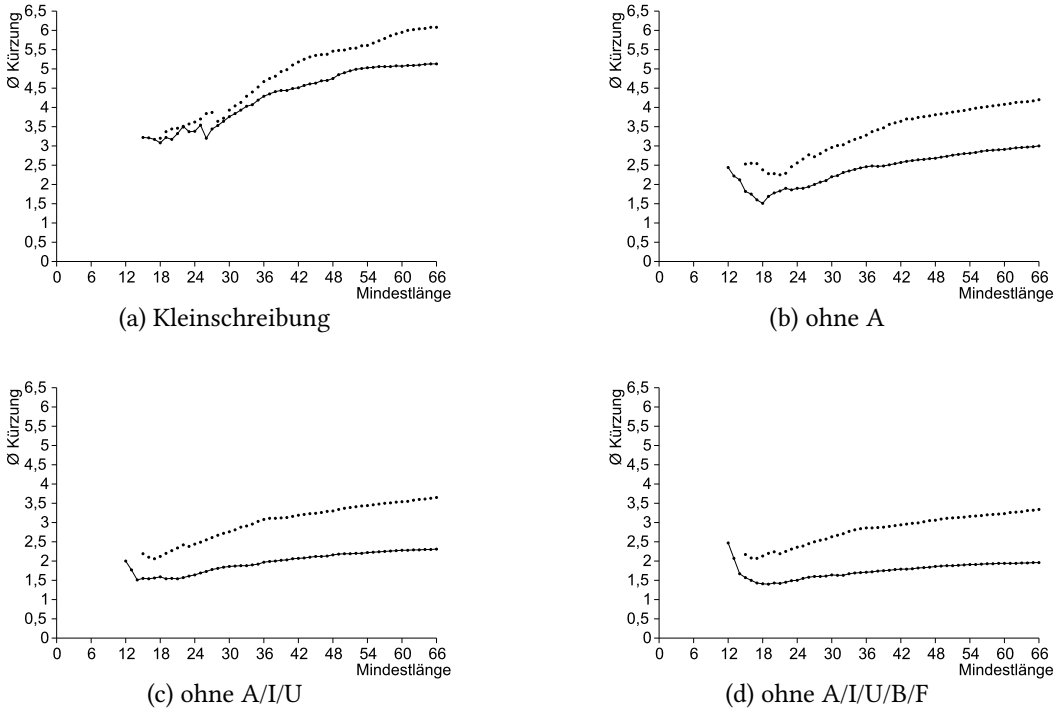
ohne Leerzeichen betrachtet. Viel häufiger als auf einer unterschiedlichen oder unklaren Leerzeichensetzung in eigentlich übereinstimmenden Wortfolgen beruhen Abweichungen zwischen Match- und Wortgrenzen aber darauf, dass ein Match im Randbereich auch Zeichen oder Zeichenfolgen umfasst, denen mit Sicherheit kein vollständiges Wort zuzuordnen ist.

Um möglichst aussagekräftige Daten zu wörtlichen Übereinstimmungen zu erhalten, ist es also insbesondere sinnvoll, von den Rändern jedes Matchbereichs aus zu prüfen, wie weit dieser Bereich verkleinert werden muss, damit die Endpunkte in beiden jeweils betrachteten Texten mit einer Wortgrenze zusammenfallen. Wie oben auf S. 132 schon dargestellt, lassen sich bei entsprechender Protokollierung der Wortpositionen die Wortgrenzen für eine bestimmte Textposition mit überschaubarem Aufwand ermitteln, und sofern die Codierung nicht dazu führt, dass einem Match zwei im Original ganz unterschiedliche Textstücke zugeordnet sind, reicht es jedenfalls im hier untersuchten Korpus meist aus, in jedem Text die Wortgrenzen des ersten und des letzten Worts im Matchbereich festzustellen, um eine solche Verkleinerung durchführen zu können.

Eine Kürzung der Matches nach diesem Kriterium wirkt sich je nach Codierung und Mindestlänge in unterschiedlichem Maße auf die Länge der einzelnen Matches aus. Offensichtlich gilt das im Hinblick auf den Kürzungsanteil: Bei einem Match, das fünf vollständige Wörter umfasst, fällt die Eliminierung von zufällig gleichen Zeichenfolgen davor und danach prozentual stärker ins Gewicht als bei einem Match von zehn vollständigen Wörtern. Übereinstimmungen mit im Hinblick auf die jeweilige Codierung großer Länge verlieren auch nur wenig von ihrer Signifikanz, wenn im Randbereich einige Zeichen gestrichen werden, da sie keine vollständigen Wörter bilden; wenn aber nach Übereinstimmungen mit nur relativ geringer Mindestlänge gesucht wird, kann sich der Matchbereich durch die Anpassung an die Wortgrenzen so sehr verkleinern, dass die Aussagekraft der Übereinstimmung, die ja ohnehin bei kurzen Matches eher gering ist, noch einmal zusätzlich erheblich eingeschränkt wird, und in manchen Fällen wird sogar überhaupt kein vollständiges Wort mit übereinstimmender Abgrenzung in beiden Texten gefunden.

Daneben lässt sich auch feststellen, dass der absolute Wert der durchschnittlichen Kürzung bei einer Anpassung an die Wortgrenzen je nach Codierung und Mindestlänge unterschiedlich hoch ausfällt, und zwar, wenn man sehr niedrige Mindestlängen ausklammert und über kleinere Schwankungen in den Messdaten hinwegsieht, um so höher, je größer die Mindestlänge ist. Bei den niedrigsten hier

Abstände zwischen manchen Buchstaben innerhalb von Wörtern in manchen Fällen recht groß, so dass die Wortabgrenzung nach rein optischen Kriterien nicht immer klar zu erkennen ist. Ein weiteres Problem ergibt sich daraus, dass bei einer Worttrennung aufgrund eines Zeilenumbruchs insbesondere in den älteren Texten des Untersuchungskorpus vielfach kein Trennzeichen gesetzt ist.



(a) Kleinschreibung (b) ohne A

(c) ohne A/I/U (d) ohne A/I/U/B/F

Abb. 3.7: bei Anpassung an Wortgrenzen durchschnittlich entfallende Zeichen (durch eine Linie verbundene Punkte: Codierungen ohne Leerzeichen, gestrichelte Linien: Codierungen mit Leerzeichen)

untersuchten Mindestlängen hingegen lässt sich teilweise eine umgekehrte Entwicklungsrichtung erkennen, die Absenkung der Mindestlänge unter eine gewisse Grenze führt also nicht zu einer Verringerung, sondern zu einer Steigerung der Kürzung.

Abbildung 3.7 veranschaulicht den Zusammenhang in Diagrammform. Dabei lassen sich gewisse Unterschiede zwischen den verschiedenen Codierungen erkennen, wobei zu betonen ist, dass die Datenreihen jeweils bei einer bestimmten Mindestlänge einsetzen und sich deshalb für die niedrigsten Längen kein gänzlich vergleichbares Bild ergibt.[566]

Dass die Länge der durchschnittlichen Kürzung bei größeren Mindestlängen zunimmt, kann wohl darauf zurückgeführt werden, dass die dabei herausgenommenen Matchbereiche teilweise auf zufälligen Übereinstimmungen einzelner Zeichen und teilweise auf einer Entsprechung oder Ähnlichkeit im Wortlaut beruhen. Vor allem sehr kurze wörtliche Entsprechungen sind vielfach nicht durch direkte oder indirekte textuelle Beziehungen zu erklären, sondern vielmehr durch übliche Wortver-

[566] Der Abbildung liegen die Angaben zur Zahl und Gesamtlänge der MEMs aus den Tabellen 3.6 und 3.7 und die Angaben zur Gesamlänge nach dem Filterschritt 1 (der Anpassung an die Wortgrenzen) aus den Tabellen 3.8 und 3.9 sowie entsprechende Messwerte für die übrigen Mindestlängen im untersuchten Wertebereich zugrunde.

bindungen oder auch durch zufällige Gleichheit in der Wortwahl, und entsprechend häufig kommt es vor, dass davor und danach keine Übereinstimmung oder Ähnlichkeit in der Formulierung festzustellen ist. Bei Matches mit einer recht großen Länge hingegen muss man davon ausgehen, dass sie in irgendeiner Weise auf der Verwendung von Vorlagen beruhen und damit möglicherweise eine unmittelbare Verwandtschaft zwischen den beiden betrachteten Texten, zumindest aber eine gemeinsame Abhängigkeit von einer bestimmten (möglicherweise ausdifferenzierten) Texttradition besteht. Und je länger die gefundenen Übereinstimmungen sind, desto eher ist eine enge Verwandtschaft zu vermuten. Damit steigt aber auch die Wahrscheinlichkeit, dass vor beziehungsweise nach einem Match jeweils eine ähnliche Wortform steht, so dass im Durchschnitt eine längere Zeichenfolge übereinstimmt als bei rein zufälliger Wortwahl. Dass diese Wortformen nicht völlig gleich sind, kann dabei an einer tatsächlichen Änderung im Wortlaut liegen oder aber an einer durch das Codierungsverfahren nicht aufgefangenen Schreibungsvarianz.

Dass sich die Messwerte zu den niedrigsten Mindestlängen teilweise nicht in die beschriebene Entwicklungslinie einfügen, lässt sich wohl ebenfalls dadurch erklären, dass die Entsprechungen teilweise zufällig sind. Zugleich zeigt sich dabei der Effekt der Codierung: Während die Messdaten für die Fassungen in Kleinschreibung mit und ohne Leerzeichen bei den niedrigeren Mindestlängen keine ganz klare Entwicklungstendenz zeigen, lässt sich vor allem bei den Codierungen, die auch die Leerzeichen eliminieren, feststellen, dass die Kurve der Messwerte deutlich umschlägt von einer ziemlich stark fallenden zu einer weniger stark steigenden Linie. Das kann man wohl darauf zurückführen, dass den ermittelteten MEMs bei den niedrigsten Mindestlängen tatsächlich in nicht wenigen Fällen keine übereinstimmende Formulierung zugeordnet werden kann, sondern dass ganz unterschiedliche Wortfolgen zufällig auf gleiche Folgen von Codezeichen abgebildet werden. Eine Anpassung an die äußersten gemeinsamen Wortgrenzen behält bei Codierungen, die die Leerzeichen unverändert lassen, die vollständigen Wörter im Kernbereich der Matches bei; bei Codierungen ohne Leerzeichen kann es hingegen durchaus sein, dass die Wortgrenzen innerhalb der Matches in den beiden verglichenen Texten nicht an den einander entsprechenden Positionen liegen. Es kann sich dabei trotzdem um einander entsprechende Wortfolgen handeln – der Verzicht auf eine Repräsentation der Leerzeichen beruht ja gerade darauf, dass die Leerzeichensetzung in den zu vergleichenden Texten auch bei einer im Prinzip wörtlichen Übereinstimmung nicht unbedingt gleich sein muss. Wenn enthaltene Wortgrenzen nicht übereinstimmen, kann das aber auch daran liegen, dass die codierten Textstücke nur zufällig übereinstimmen.

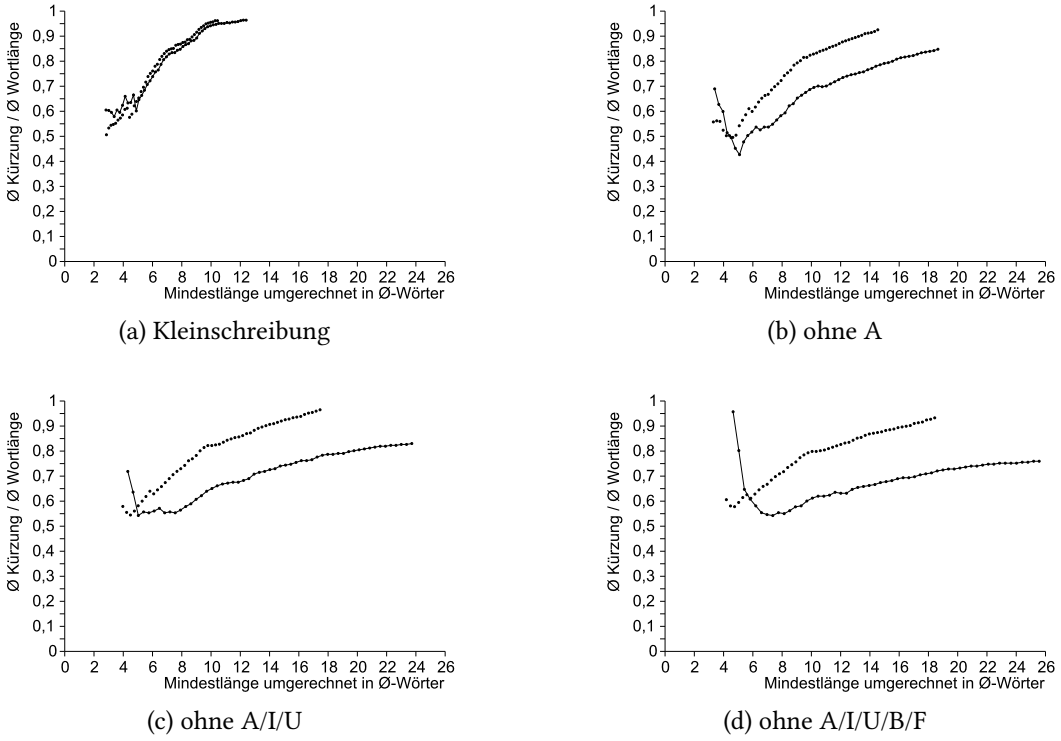
Ein Anhaltspunkt dafür ist der Anteil der Matches, die bei einer Anpassung an die äußersten gemeinsamen Wortgrenzen wegfallen, also kein vollständiges Wort enthalten. Bei den kürzeren hier untersuchten Mindestlängen lässt sich nämlich

teilweise keine oder nur eine einzige gemeinsame Wortgrenze finden. Das kann zwei Ursachen haben: Ein *maximal exact match* kann aus dem Ende eines Worts und dem Anfang eines weiteren Worts oder auch überhaupt nur aus einem unvollständigen Wort bestehen, oder es sind zwar Wortgrenzen enthalten, die aber nicht an den einander entsprechenden Positionen liegen. Die zuletzt genannte Fallkonstellation kommt nur dann in Betracht, wenn die zugrunde gelegten Textfassungen keine Leerzeichen enthalten.

Beim Vergleich von Codierungen ohne Leerzeichen und bemerkenswerterweise auch beim Vergleich auf der Basis einer Umwandlung in Kleinbuchstaben ohne Leerzeichen entfallen im Untersuchungskorpus bei den kürzesten ausgewerteten Mindestlängen zwischen etwa 1,6 % und 4,09 % der Matches, wenn unvollständige Wörter im Randbereich gestrichen werden. Der Fall, dass ein MEM in keinem der Texte ein vollständiges Wort umfasst, ist demgegenüber offenbar viel seltener – bei der Wortgrenzenanpassung der Matches in Textfassungen, die Leerzeichen enthalten, entfällt auch bei den kürzesten untersuchten Mindestlängen nur ein Anteil, der zwischen weniger als 0,0017 % und etwas mehr als 0,46 % liegt.[567]

Um die Daten der verschiedenen Codierungen besser vergleichen zu können, ist auch für die bei der Anpassung an die Wortgrenzen durchschnittlich entfallenden Zeichen eine Umrechnung der Mindestlängen und der Kürzungen in Wörter durchschnittlicher Länge sinnvoll. Abbildung 3.8 veranschaulicht die sich daraus ergebenden Daten. Hier zeigt sich, dass der Übergang von fallenden zu steigenden Werten für die Codierungsvarianten *ohne A* ohne Leerzeichen und *ohne A/I/U* ohne Leerzeichen jeweils etwa bei fünf Wörtern durchschnittlicher Länge erfolgt, bei der Codierung *ohne A/I/U/B/F* ohne Leerzeichen hingegen bei etwa sieben Wörtern durchschnittlicher Länge. Nach der eben angestellten Überlegung ist die zuletzt genannte Codierung also für die Suche nach Übereinstimmungen, die zum Beispiel nur fünf Wörter durchschnittlicher Länge umfassen, nicht allzu gut geeignet, vielmehr empfiehlt sich dafür eine stärker differenzierende Codierung, oder es muss mit einer niedrigeren *Precision* gerechnet werden, so dass möglicherweise zusätzliche Schritte zur Überprüfung der Matches sinnvoll sein könnten. Inwieweit das auch auf die Codierungsvarianten, die Leerzeichen beibehalten, übertragen werden kann, soll hier nicht weiter untersucht werden. Die Messdaten sind hierzu wohl nicht unbedingt aussagekräftig, da eine Kürzung ja nur die Zeichen vor dem ersten und nach dem letzten Leerzeichen umfassen kann. Zwar wird auch hier bei den niedrigsten Mindestlängen eine etwas stärkere Kürzung verzeichnet als bei etwas höheren Werten, aber das könnte sich durch normale Schwankungen in statistischen Daten erklären lassen. Prinzipiell ist jedenfalls auch bei Codierungen

[567] Den Prozentangaben liegen zum einen die Matchzahlen in den Tabellen 3.6 und 3.7 auf S. 156 und 157, zum anderen die Zahlen zu den nach dem Filterschritt 1 verbleibenden Matches in den Tabellen 3.8 und 3.9 auf S. 166 und 167 zugrunde.



(a) Kleinschreibung (b) ohne A

(c) ohne A/I/U (d) ohne A/I/U/B/F

Abb. 3.8: bei Anpassung an Wortgrenzen durchschnittlich entfallende Wortanteile (durch eine Linie verbundene Punkte: Codierungen ohne Leerzeichen, gestrichelte Linien: Codierungen mit Leerzeichen)

mit Leerzeichen bei sehr niedrigen Mindestlängen mit Matches zu rechnen, die auf zufällig gleichen Folgen von Codezeichen beruhen, denen keine Übereinstimmung im Wortlaut zugrunde liegt.

Das Problem, dass bei sehr kurzen Matches mit einer nicht ganz unerheblichen Wahrscheinlichkeit keine weitgehend übereinstimmende Wortfolge zugrunde liegt, ist durch die Anpassung der Matches an die Wortgrenzen zwar reduziert, aber nicht behoben. Eine einfache Möglichkei, die Matchqualität weiter zu verbessern, besteht bei Textfassungen ohne Leerzeichen darin, eine Plausibilitätskontrolle auf der Basis eines Vergleichs der nach der Wortgrenzenanpassung innerhalb des Matchbereichs verbliebenen inneren Wortgrenzen vorzunehmen. In Tabelle 3.9 auf S. 167 ist in den Spalten zum „Filterschritt 2“ dokumentiert, wie sich eine Auswahl auf der Basis eines Vergleichs dieser Wortgrenzen auswirkt. Die hierbei zugrunde gelegten Auswahlregeln sind relativ großzügig gestaltet mit der Intention, möglichst wenige tatsächliche Übereinstimmungen auszuschließen.[568]

[568] Ein Match wird im Filterschritt 2 übernommen, wenn die darin in einem der Texte enthaltenen Wortgrenzen zu mindestens 70 % mit Wortgrenzen im anderen Text übereinstimmen oder (um möglicherweise innerhalb eines zu weit gefassten Matchs enthaltene tatsächliche Übereinstimmungen nicht auszuschließen) an drei aufeinander folgenden Stellen. Wenn ein Match in beiden

Wie sich der Tabelle entnehmen lässt, gibt es bei sehr niedrigen Mindestlängen einen gewissen, allerdings kleinen Anteil von Matches, bei denen die Leerzeichenpositionen in den beiden Texten stärker voneinander abweichen, als es diesen Regeln entspricht. Auch dies kann wohl als Indikator dafür dienen, dass bei diesen Mindestlängen mit Matches zu rechnen ist, denen keine Entsprechung im Wortlaut zugeordnet werden kann.[569] Das bedeutet nicht automatisch, dass alle verbleibenden Matches in dieser Hinsicht unproblematisch sind. Wenn über diesen Filterschritt aber keine oder fast keine Matches eliminiert werden, scheint dies ein Anhaltspunkt dafür zu sein, dass den Matches bis auf wenige Ausnahmen tatsächlich im Kernbereich gleiche Formulierungen zugrunde liegen. So zeigt sich bei einer Überprüfung der 129 Matches der Mindestlänge 18 in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen, die mit dem Filterschritt 2 eliminiert werden, dass sie zum größten Teil immerhin partiell auf gleichen Wörtern beruhen.[570]

Dieser Befund passt gut zu den Beobachtungen im Hinblick auf die durchschnittliche Kürzung bei der Anpassung der Matches an die Wortgrenzen: Bei den Codierungen mit Eliminierung von Leerzeichen gibt es bei zunehmender Reduzierung der Matchlänge bei Längen unterhalb des Werts, bei dem die Kürzung für die jeweilige Codierung den niedrigsten Wert erreicht, immer mehr Matches, die den Bedingungen des zweiten Filterschritts nicht entsprechen und denen vielfach tatsächlich keine wörtliche Entsprechung zugrunde liegt. Deshalb scheint es plausibel, die Daten zur Matchkürzung jedenfalls für Codierungen ohne Leerzeichen als Anhaltspunkt dafür zu nehmen, ab welcher Mindestlänge mit einigermaßen verlässlichen Ergebnissen in dem Sinne zu rechnen ist, dass den ermittelten Matches tatsächlich zumindest weitgehend übereinstimmende Wortfolgen zugrunde liegen.[571]

Texten jeweils vier Wörter umfasst, reicht es also nicht aus, wenn von den jeweils drei enthaltenen Leerzeichenpositionen nur zwei an den einander entsprechenden Stellen liegen – wenn das Match aber vier Wörter in einem der Texte umfasst und drei Wörter im anderen (weil zum Beispiel ein Kompositum in einem Text getrennt geschrieben ist), ist das Auswahlkriterium erfüllt, wenn zwei Wortgrenzen einander zugeordnet werden können.

[569] Auch bei einem Vergleich auf der Basis einer Umwandlung in Kleinbuchstaben ohne Leerzeichen entfallen im Untersuchungskorpus beim Filterschritt 2 einige tausend Matches, der Anteil ist aber verschwindend gering und viel niedriger als bei den Textrepräsentationen auf der Basis einer Codierung. Es handelt sich dabei teilweise tatsächlich um unterschiedliche Wortfolgen (zum Beispiel „Testaments / oder “ und „testament / so der“) und teilweise um im Prinzip gleiche Wortfolgen mit unterschiedlicher Leerzeichensetzung (zum Beispiel „als dañ zusehẽ“ und „alsdañ zu sehen“). Bei der noch relativ stark differenzierenden Codierung *ohne A* ohne Leerzeichen weisen von den Übereinstimmungen, die beim Filterschritt 2 herausgenommen werden, auch bei der Mindestlänge 12 viele immerhin eine partielle Übereinstimmung im Wortlaut auf (zum Beispiel „vnd procediert werden. Wann“ und „vnd procediert würt. Nůn“). Bei der Codierung *ohne A/I/U/B/F* ohne Leerzeichen und der Mindestlänge 12 werden mit dem Filterschritt 2 sehr viele Matches eliminiert, bei denen tatsächlich keine Ähnlichkeit im Wortlaut zu erkennen ist (zum Beispiel „der Appellation nach“ und „ayd erhalten mag“, beides repräsentiert durch die Codefolge „DRLDMMG“).

[570] Ein Beispiel ist ein Match, dem die Wortfolgen „der Kriegsbefestigung ist ein“ und „der Kriegsbefestigung / so etwan“ zugrunde liegen.

Die Wortabgrenzung lässt sich auch noch für einen weiteren Filterschritt nutzen. Bei der Anpassung der Matches an die Wortgrenzen hat sich schon gezeigt, dass sehr kurze Matches teilweise überhaupt keine vollständigen Wörter enthalten. Aber auch wenn bei dieser Anpassung ein oder zwei Wörter übrig bleiben, kann eine solche Übereinstimmung für sich genommen – außer vielleicht in ganz besonderen Ausnahmefällen – kaum als Anzeichen für eine textuelle Beziehung gewertet werden. Auch Matches, die nur drei (oder auch noch etwas mehr) Wörter umfassen, lassen sich wohl in vielen Fällen als zufällig erklären. Da Häufungen von übereinstimmenden Wort-Trigrammen aber als Indikator für Textabhängigkeiten genutzt werden können,[572] könnte es sinnvoll sein, sie in der weiteren Analyse der Daten zu berücksichtigen. Deshalb sollen hier in dem in den Tabellen 3.8 und 3.9 als „Filterschritt 3“ bezeichneten Verarbeitungsschritt nur solche Matches gestrichen werden, die in beiden Texten maximal zwei Wörter umfassen.

Eine Filterung anhand dieses Kriteriums führt, wie in den beiden Tabellen zu sehen ist, bei sehr kurzen Mindestlängen teilweise zu einer starken Reduzierung der Gesamtzahl der Matches. Das gilt insbesondere für die Textfassungen in Kleinbuchstaben mit und ohne Leerzeichen, bei denen die niedrigste untersuchte Mindestlänge nur etwa 2,85 beziehungsweise 2,82 Wörtern durchschnittlicher Länge entspricht. Schon diese Angabe lässt vermuten, dass MEMs dieser Länge wenig aussagekräftig sind, denn wenn sie drei vollständige Wörter enthalten, hat davon mit großer Wahrscheinlichkeit höchstens eines eine etwas größere Wortlänge. Aber auch bei den codierten Textfassungen erfüllt ein teilweise nicht unerheblicher Teil der MEMs bei den niedrigsten betrachteten Mindestlängen dieses Auswahlkriterium nicht.

Während sich die Anpassung an Wortgrenzen – soweit es sich um Texte mit entsprechend abgegrenzten Einheiten handelt – ohne eigene Interpretationsleistung durchführen lässt, ist bei weitergehenden Filtermaßnahmen mit Fehlentscheidungen zu rechnen, wobei freilich eine Bewertung in Form einer Bezifferung von *Precision* und *Recall* mangels einer objektiv richtigen Auswahl kaum praktikabel und sachgerecht sein dürfte. Grundsätzlich ist aber zu berücksichtigen, dass bei einer Erhöhung der *Precision* mit einer Verringerung des *Recalls* zu rechnen ist und umgekehrt. Das gilt auch für die schon beschriebenen Filterschritte 2 und 3, also die Entfernung von Matches, bei denen die Wortgrenzen in den beiden Texten zu stark voneinander abweichen und die zu wenige Wörter enthalten. Die hier angesetzten Auswahlkriterien sind sehr großzügig gestaltet, so dass ihre Anwendung wohl kaum zu einer Eliminierung von Matches führen dürfte, die auch ohne weite-

[571] Inwieweit das auf andere Sprachen und Codierungen übertragen werden kann, soll hier nicht weiter untersucht werden. Wenn die Wortlängen in einer Sprache beziehungsweise in einer Codierung weniger stark variieren als im Untersuchungskorpus, eignet sich die Übereinstimmung bei der Wortabgrenzung sicherlich weniger gut als Bewertungskriterium.

[572] Vgl. oben S. 79, Anm. 320, S. 86, Anm. 350 und S. 95 f.

Mindest-länge	Anzahl nach Filterschritt 1	Gesamtlänge nach Filterschritt 1	Anzahl nach Filterschritt 2	Gesamtlänge nach Filterschritt 2	Anzahl nach Filterschritt 3	Gesamtlänge nach Filterschritt 3
18	44.185.656	763.805.482	44.185.656	763.805.482	16.732.406	337.699.348
24	4.987.050	127.017.884	4.987.050	127.017.884	4.145.710	111.506.242
30	1.239.870	43.751.014	1.239.870	43.751.014	1.199.352	42.797.782
36	485.516	22.330.830	485.516	22.330.830	483.052	22.264.988
42	271.054	15.013.168	271.054	15.013.168	270.854	15.007.692
48	170.996	11.057.276	170.996	11.057.276	170.980	11.056.748
54	121.146	8.804.236	121.146	8.804.236	121.146	8.804.236
60	84.606	6.923.602	84.606	6.923.602	84.606	6.923.602
66	62.978	5.697.532	62.978	5.697.532	62.978	5.697.532

(a) Umwandlung in Kleinbuchstaben, mit Leerzeichen

15	35.759.140	536.364.472	35.759.140	536.364.472	23.453.132	384.537.004
18	10.704.900	208.832.492	10.704.900	208.832.492	8.758.328	180.396.480
24	1.694.094	54.408.390	1.694.094	54.408.390	1.685.284	54.250.654
30	584.106	28.453.302	584.106	28.453.302	583.860	28.447.288
36	339.630	21.229.204	339.630	21.229.204	339.622	21.228.980
42	226.358	17.200.732	226.358	17.200.732	226.350	17.200.508
48	169.096	14.851.238	169.096	14.851.238	169.096	14.851.238
54	132.524	13.133.376	132.524	13.133.376	132.524	13.133.376
60	107.428	11.804.830	107.428	11.804.830	107.428	11.804.830
66	88.770	10.707.132	88.770	10.707.132	88.770	10.707.132

(b) Codierung *ohne A* mit Leerzeichen

15	14.008.132	219.019.456	14.008.132	219.019.456	11.870.118	193.325.560
18	3.879.422	82.617.672	3.879.422	82.617.672	3.795.390	81.398.864
24	757.376	28.404.896	757.376	28.404.896	756.380	28.382.676
30	353.280	18.871.876	353.280	18.871.876	353.268	18.871.612
36	220.670	14.897.680	220.670	14.897.680	220.670	14.897.680
42	160.002	12.749.312	160.002	12.749.312	160.002	12.749.312
48	124.280	11.266.712	124.280	11.266.712	124.280	11.266.712
54	97.800	10.011.652	97.800	10.011.652	97.800	10.011.652
60	79.958	9.061.296	79.958	9.061.296	79.958	9.061.296
66	66.066	8.235.700	66.066	8.235.700	66.066	8.235.700

(c) Codierung *ohne A/I/U* mit Leerzeichen

15	12.304.738	192.138.288	12.304.738	192.138.288	11.300.016	180.846.284
18	3.217.550	69.780.470	3.217.550	69.780.470	3.163.116	68.975.448
24	641.304	25.069.728	641.304	25.069.728	640.360	25.048.504
30	316.180	17.349.822	316.180	17.349.822	316.168	17.349.570
36	200.910	13.905.634	200.910	13.905.634	200.910	13.905.634
42	148.600	12.043.820	148.600	12.043.820	148.600	12.043.820
48	114.204	10.606.508	114.204	10.606.508	114.204	10.606.508
54	91.626	9.531.906	91.626	9.531.906	91.626	9.531.906
60	75.246	8.655.396	75.246	8.655.396	75.246	8.655.396
66	62.172	7.876.416	62.172	7.876.416	62.172	7.876.416

(d) *ohne A/I/U/B/F* mit Leerzeichen

Tab. 3.8: Daten zur Filterung der Matches anhand der Wortgrenzen bei verschiedenen Codierungen und Mindestlängen (Filterschritt 1: Eliminierung unvollständiger Wörter; Filterschritt 2: Eliminierung von Matches mit deutlich abweichenden enthaltenen Wortgrenzen; Filterschritt 3: Eliminierung von Matches mit weniger als 3 Wörtern; vgl. S. 159, 163 (Anm. 568) und 165)

Mindest-länge	Anzahl nach Filterschritt 1	Gesamtlänge nach Filterschritt 1	Anzahl nach Filterschritt 2	Gesamtlänge nach Filterschritt 2	Anzahl nach Filterschritt 3	Gesamtlänge nach Filterschritt 3
15	47.343.556	678.766.568	47.339.490	678.700.693	13.978.464	248.956.723
18	12.172.500	220.533.154	12.170.976	220.504.285	6.844.436	142.412.343
24	1.816.790	49.425.818	1.816.662	49.422.679	1.559.578	44.892.419
30	527.402	20.410.698	527.382	20.410.115	519.438	20.239.703
36	268.358	12.934.322	268.357	12.934.284	267.507	12.909.896
42	162.906	9.328.686	162.906	9.328.686	162.884	9.327.992
48	109.448	7.181.358	109.448	7.181.358	109.440	7.181.158
54	74.622	5.580.472	74.622	5.580.472	74.622	5.580.472
60	54.858	4.565.020	54.858	4.565.020	54.858	4.565.020
66	41.064	3.773.544	41.064	3.773.544	41.064	3.773.544

(a) Umwandlung in Kleinbuchstaben ohne Leerzeichen

12	32.554.840	388.695.088	32.512.005	388.231.210	16.191.833	229.967.586
18	2.562.470	58.819.168	2.561.766	58.807.024	2.424.322	56.913.842
24	569.416	23.034.440	569.414	23.034.397	567.680	23.002.459
30	284.262	16.087.736	284.262	16.087.736	284.198	16.086.202
36	185.962	13.093.942	185.962	13.093.942	185.962	13.093.942
42	136.930	11.323.362	136.930	11.323.362	136.930	11.323.362
48	106.006	10.021.576	106.006	10.021.576	106.006	10.021.576
54	84.760	8.998.804	84.760	8.998.804	84.760	8.998.804
60	69.532	8.176.492	69.532	8.176.492	69.532	8.176.492
66	57.948	7.483.426	57.948	7.483.426	57.948	7.483.426

(b) Codierung *ohne A* ohne Leerzeichen

12	10.392.966	129.874.098	10.273.738	128.776.506	7.349.290	102.226.156
18	823.456	23.135.230	823.336	23.133.269	817.850	23.058.977
24	296.204	13.676.834	296.203	13.676.811	296.121	13.675.205
30	173.648	10.654.376	173.648	10.654.376	173.648	10.654.376
36	121.894	9.066.462	121.894	9.066.462	121.894	9.066.462
42	91.572	7.957.944	91.572	7.957.944	91.572	7.957.944
48	71.988	7.124.478	71.988	7.124.478	71.988	7.124.478
54	57.964	6.446.328	57.964	6.446.328	57.964	6.446.328
60	47.842	5.894.774	47.842	5.894.774	47.842	5.894.774
66	40.538	5.454.356	40.538	5.454.356	40.538	5.454.356

(c) Codierung *ohne A/I/U* ohne Leerzeichen

12	8.664.926	105.357.620	8.398.140	103.055.648	6.601.498	89.163.714
18	615.990	18.754.556	615.861	18.752.510	612.565	18.705.264
24	251.620	12.088.272	251.620	12.088.272	251.570	12.087.258
30	154.360	9.691.528	154.360	9.691.528	154.360	9.691.528
36	108.742	8.287.518	108.742	8.287.518	108.742	8.287.518
42	82.782	7.330.936	82.782	7.330.936	82.782	7.330.936
48	64.504	6.546.418	64.504	6.546.418	64.504	6.546.418
54	52.846	5.979.006	52.846	5.979.006	52.846	5.979.006
60	43.612	5.474.722	43.612	5.474.722	43.612	5.474.722
66	37.002	5.074.382	37.002	5.074.382	37.002	5.074.382

(d) Codierung *ohne A/I/U/B/F* ohne Leerzeichen

Tab. 3.9: Daten zur Filterung der Matches anhand der Wortgrenzen bei verschiedenen Codierungen und Mindestlängen (Filterschritt 1: Eliminierung unvollständiger Wörter; Filterschritt 2: Eliminierung von Matches mit deutlich abweichenden enthaltenen Wortgrenzen; Filterschritt 3: Eliminierung von Matches mit weniger als 3 Wörtern; vgl. S. 159, 163 (Anm. 568) und 165)

re Übereinstimmungen im unmittelbaren Umfeld von Interesse sind – der *Recall* wird also wohl allenfalls unwesentlich verringert, die *Precision* aber jedenfalls für sehr kurze Matches erhöht. Deshalb werden diese Filterschritte bei den weiteren Untersuchungen vorausgesetzt.

Darüber hinaus gibt es noch vielfältige weitere Möglichkeiten einer Matchfilterung, von denen im Folgenden einige beschrieben, allerdings nicht näher untersucht werden sollen, da ihre konkrete Ausgestaltung Detailentscheidungen erfordert und da sie teilweise von zusätzlichen Informationen über die Textstruktur beziehungsweise die Sprache abhängen, so dass eine Dokumentation der jeweiligen Auswirkung auf den Umfang der Matches wohl wenig aussagekräftig wäre. Je nach Erkenntnisinteresse beziehungsweise Vorannahmen dürften sie unterschiedlich sinnvoll sein. Insbesondere ist dabei zu differenzieren zwischen Fragestellungen, die eher statistischer Natur sind und solchen, bei denen es um die Gegenüberstellung von einander entsprechenden Passagen für den Vergleich geht. Die große Zahl an jedenfalls für das hier untersuchte Korpus festgestellten Übereinstimmungen legt es nahe, Letzteres nur für einen vergleichsweise kleinen Teil durchzuführen, wobei die Auswahl freilich nicht auf den hier genannten Filterkriterien beruhen muss. Wenn etwa die Annahme zugrunde liegt, dass zwei Texte in einem engen Abhängigkeitsverhältnis zueinander stehen, kann es angemessen sein, auch Übereinstimmungen zu berücksichtigen, die bestimmten Filterkriterien nicht genügen. Das Gleiche kann gelten, wenn die Rezeption oder die Vorlagen einer bestimmten Textstelle betrachtet werden sollen: Hier ist vermutlich nicht ausschlaggebend, eine möglichst hohe *Precision* zu erreichen, sondern vielmehr (soweit sich die *Precision* noch in einem sinnvollen Rahmen bewegt), nichts Relevantes zu übersehen.

Prinzipiell bietet es sich an an, eine Filterung anhand syntaktischer Strukturen vorzunehmen. Eine ermittelte Übereinstimmung kann zum Beispiel mit einer codierten Wortform enden, die das erste Wort eines neuen Satzes repräsentiert. Selbst wenn hier tatsächlich in den originalen Texten jeweils das gleiche Wort steht, ist die Wahrscheinlichkeit wohl sehr gering, dass eine solche Übereinstimmung von Interesse ist, jedenfalls wenn es sich um ein Stoppwort im weiteren Sinne handelt (was am Anfang eines Satzes im Deutschen in aller Regel der Fall ist). Und dass es sich tatsächlich um das gleiche Wort handelt, ist in solch einem Fall auch weniger wahrscheinlich als bei der Einbettung in eine Wortfolge, da hier die einzige Information zum syntaktischen Kontext in der Positionierung am Satzanfang besteht.

Ein solcher Filterschritt setzt allerdings voraus, dass sich die Satzgrenzen mit zumindest einigermaßen großer Sicherheit ermitteln lassen. Während das für Texte in heutigem Deutsch meist gut machbar sein dürfte, erweist sich die Syntax frühneuhochdeutscher Texte als in recht vielen Fällen auch für den sprachkundigen Leser schwer durchschaubar. Insbesondere ist eine Untergliederung in Sätze keineswegs immer eindeutig möglich, und erst recht gilt das für automatisierte

Verfahren, da sich die Zeichensetzung erst im Laufe der frühneuhochdeutschen Periode entwickelte und vielfach nicht den heutigen Kriterien entspricht.[573]

Wenn aber eine zumindest einigermaßen sinnvolle Satzgliederung – oder auch eine Einteilung in Absätze oder in Haupt- und Nebensätze – vorliegt, kann sie als Indikator für die Bewertung beziehungsweise Eingrenzung von Matches genutzt werden. So sind kleinere Randbereiche von Matches, die jenseits von Satzgrenzen liegen, mit relativ hoher Wahrscheinlichkeit irrelevant. Eine genaue Anzahl von Wörtern, bis zu der ein solcher Randbereich für die weitere Untersuchung nicht mehr berücksichtigt werden soll, lässt sich wohl nicht allgemein angeben und hängt von der jeweiligen Gewichtung von *Precision* und *Recall* ab. Bei einem Feinvergleich des Umfelds von exakten Übereinstimmungen können auch kurze Stücke zu Beginn oder Ende eines Satzes von Interesse sein, weil dabei die Grundannahme ist, dass trotz einer Textabweichung im Umfeld eine Ähnlichkeit vorliegen könnte. Auch in diesem Fall dürften aber Einzelwörter, die zugleich Stoppwörter sind, in aller Regel kein Hinweis auf eine textuelle Beziehung sein.

Natürlich ist auch innerhalb von Sätzen damit zu rechnen, dass ermittelte Matches Randwörter umfassen, die auf zufälligen Übereinstimmungen beruhen. Mit wenig Aufwand lässt sich eine Filterung anhand der Wortart (*part of speech*, POS) vornehmen, wenn diese Information schon in der Auszeichnung der Originaldaten enthalten ist. Während dies für sprachwissenschaftliche Korpora zu modernen Sprachen recht üblich ist, gibt es allerdings für historische Sprachformen nur vergleichsweise wenig entsprechend aufbereitetes Material,[574] und aufgrund der Schwierigkeiten bei der automatisierten Analyse frühneuhochdeutscher Sätze ist ein weitgehend zuverlässiges POS-*Tagging* jedenfalls nach dem derzeitigen Stand für diese Sprachstufe mit einem erheblichen Aufwand verbunden und wird für Projekte, die nicht vorrangig an linguistischen Fragestellungen orientiert sind, wohl bis auf Weiteres eher die Ausnahme bleiben.

[573] Vgl. zur frühneuhochdeutschen Interpunktion Reichmann/Wegera 1993, S. 28–31. Inwieweit man bei modernen Texten davon ausgehen kann, dass sich eine Satzgliederung zweifelsfrei vornehmen lässt, soll ebenso wie die Frage, was überhaupt ein *Satz* ist, hier nicht weiter betrachtet werden. Jedenfalls dient auch heute die Setzung von Punkten zum Teil eher rhetorischen Zwecken als einer syntaktischen Gliederung – so etwa als im journalistischen Bereich verbreitetes Stilmittel. Zudem hat auch heute der Punkt eine Doppelfunktion als Satzgliederungs- und Abkürzungszeichen. Gleichwohl kann man bei modernen schriftsprachlichen Texten in aller Regel von einer Zeichensetzung ausgehen, die eine Untergliederung in syntaktisch sinnvolle Einheiten ermöglicht, was für das Frühneuhochdeutsche in sehr viel geringerem Maße der Fall ist.

[574] Für das Frühneuhochdeutsche sind insbesondere das zwischen 1972 und 1985 für die Erarbeitung der *Grammatik des Frühneuhochdeutschen* erstellte *Bonner Frühneuhochdeutschkorpus* zu nennen, das nur ca. 1200 exemplarisch ausgewählte Seiten aus verschiedenen Sprachräumen und Jahrhunderten umfasst (vgl. *http://www.korpora.org/Fnhd/*), sowie das im Aufbau befindliche und noch nicht allgemein zugängliche *Referenzkorpus Frühneuhochdeutsch* des Projekts *Deutsch Diachron Digital (DDD)* (vgl. *http://www.germanistik.uni-halle.de/forschung/altgermanistik/referenzkorpus_fruehneuhochdeutsc/*).

Alternativ lässt sich natürlich eine Filterung über eine Stoppwortliste vornehmen, deren Zusammenstellung allerdings gewisse Probleme bereitet. Einfach zu implementieren ist eine Auswahl, die sich allein an der Vorkommenshäufigkeit der Wortformen orientiert. Wenn dabei die codierten Formen zugrunde gelegt werden, entfällt das Problem unterschiedlicher Schreibungen – jedenfalls bei starker Varianzreduktion – weitgehend; allerdings lässt sich anhand der Formen nicht sicher erkennen, um welche originalen Wörter es sich handelt. Eine Rückführung auf in Frage kommende Originalformen ist immerhin möglich, wenn zusätzlich eine Liste aller originalen Schreibungen erstellt und zu jeder dieser Formen die entsprechende Codierung verzeichnet wird. Auf dieser Basis kann eine manuelle Überarbeitung der Stoppwortliste durchgeführt werden, wenn eine hohe *Precision* angestrebt wird. Allerdings kommen nicht wenige Funktionswörter seltener vor als besonders häufig gebrauchte inhaltstragende Wörter – von einer primär nach statistischen Kriterien erstellten Auswahl darf man sich also keine Vollständigkeit erhoffen und muss jedenfalls bei Verzicht auf eine Prüfung der Einträge auch damit rechnen, dass sie auch Wortformen enthält, die eigentlich von Interesse wären.

Soweit entsprechende Metainformationen vorliegen, lassen sich auch bestimmte Abschnitte der Texte mit gehäuftem Auftreten von nicht relevanten Textübereinstimmungen ausschließen. Insbesondere sind hier wohl Inhaltsverzeichnisse und Register zu nennen, die vielfach Formulierungen aus dem Haupttext übernehmen, so dass Matches dieser Textstücke mit anderen Texten jeweils zu einer Mehrfachzuordnung führen, von denen aber nur die Stelle im Haupttext tatsächlich von Interesse sein dürfte. Da die Daten des hier ausgewerteteten Korpus entsprechend ausgezeichnet sind, konnte dafür eine solche Filterung durchgeführt werden, die im weiteren Verlauf dieser Untersuchung zur Erläuterung bestimmter Einzelbefunde vergleichend mit herangezogen wird.

Schließlich soll noch auf eine Filterungsmöglichkeit hingewiesen werden, die sich aus den Matchpositionen ergibt. Bei Ansetzung einer niedrigen Mindestmatchlänge lässt sich für viele Matches feststellen, dass die betreffenden Textstücke nicht nur im Korpus, sondern auch in einem oder mehreren Einzeltexten jeweils mehrfach vorkommen. Da sich daraus nicht automatisch ergibt, dass solche Matches einfach irrelevant sind, soll im folgenden Unterkapitel betrachtet werden, wie sich Mehrfachzuordnungen für eine Bewertung und damit möglicherweise auch für eine feinere Filterung nutzen lassen. Es gibt aber auch einen Sonderfall, für den – je nach Weiterverarbeitung der Matchdaten – ein zusätzlicher Filterschritt sinnvoll sein kann: Es kann vorkommen, dass ein Match komplett in einem anderen Match enthalten ist, dass also die Positionen in beiden Texten innerhalb des Bereichs liegen, der durch ein umfangreicheres Match abgedeckt wird. Das beruht darauf, dass eine Zeichenfolge, die die geforderte Mindestlänge erreicht, innerhalb des größeren Matchs zweimal vorkommt. Zumindest dann, wenn die wiederholte Zei-

chenfolge nicht am Rand des längeren Matchs liegt,[575] ist wohl recht sicher davon auszugehen, dass das kürzere Match gestrichen werden kann. Entsprechende Fälle sind im Untersuchungskorpus auch bei kurzen Mindestmatchlängen nicht sehr zahlreich, so dass sich ein solcher Filterschritt kaum auf das Gesamtbild auswirkt. Wenn aber eine detailliertere Untersuchung über einen Feinvergleich erfolgt, können Überlappungen bei der Ermittlung einer möglichst optimalen Zuordnung eine Rolle spielen, deshalb kann es das Vergleichsergebnis verbessern, wenn Matches, die gänzlich in anderen Matches enthalten sind, vorab eliminiert werden.

Bis zu diesem Unterkapitel wurde die Untersuchung an Daten zu verschiedenen Codierungsvarianten durchgeführt. Das soll in dieser Form nicht weiter fortgeführt werden. Es ist wohl deutlich geworden, dass ein Vergleich auf der Basis einer Codierung entsprechend Unterkapitel 3.1.2 für die Zielsetzung dieser Arbeit erhebliche Vorteile bringt. Zwar führt die Abbildung unterschiedlicher Zeichen auf gleiche Codezeichen dazu, dass bei der Ermittlung sehr kurzer Matches auch Wortfolgen einander zugeordnet werden, die keinerlei Übereinstimmungen aufweisen. Je kürzer die ermittelten Matches sind, desto mehr muss aber auch bei tatsächlicher Wortgleichheit damit gerechnet werden, dass sie durch sprachliche Muster oder durch zufällig gleiche Wortwahl zu erklären sind, so dass es elaborierter Verfahren bedarf, um aus Häufungen sehr kurzer Matches Hinweise auf Textbeziehungen abzuleiten.

Dieser Weg soll hier nicht beschritten werden, auch wenn im folgenden Unterkapitel gewisse Möglichkeiten zur Bewertung von Matches vorgestellt werden. Das Anliegen dieser Untersuchung ist vielmehr insbesondere, das Potential eines Vergleichs auf der Basis etwas längerer Übereinstimmungen in einer codierten Textfassung auszuloten. Dafür wird im Folgenden beispielhaft die Codierung *ohne A/I/U/B/F* ohne Leerzeichen zugrunde gelegt, also eine Variante, die zwar in Bezug auf die Nivellierung gewisser schreibsprachlicher Unterschiede gegenüber anderen Varianten begrenzte Vorteile bietet, dabei aber den Konsonantenbestand stark reduziert und deshalb für den Vergleich kurzer Zeichenfolgen mit Sicherheit nicht geeignet ist. Die Auswahl dieser Codierung beruht also nicht auf der Annahme, dass es sich um die beste Variante handele, sondern soll zeigen, dass selbst mit einem so radikalen Ansatz sehr aussagekräftige Ergebnisse zu erzielen sind, wenn eine sinnvolle Mindestlänge für die MEM-Ermittlung angesetzt wird.

Diese Mindestlänge liegt im Folgenden – soweit nicht anders angegeben – bei 18. Dieser Wert führt nach den Abbildungen 3.5 und 3.7 zum einen zum höchsten Anteil textinterner MEMs, zum anderen zum niedrigsten Umfang von bei der Anpassung

[575] Andernfalls muss auch die Möglichkeit einkalkuliert werden, dass die kürzere Übereinstimmung im Zusammenhang mit dem darauf folgenden Text steht. Das ist wohl insbesondere dann in Erwägung zu ziehen, wenn auf das kürzere Match mit wenig Abstand in beiden Texten ein weiteres Match folgt.

der Matches an die Wortgrenzen entfallenden Zeichen für diese Codierungsvariante und scheint deshalb nach den oben angestellten Überlegungen eine gute Wahl zu sein, um einen hohen *Recall* zu erreichen, ohne zugleich die *Precision* stark zu beeinträchtigen.

In Einzelfällen erfolgt auch ein Vergleich mit den Daten für eine andere Mindestlänge oder für die Fassung in Kleinbuchstaben, um einen gewissen Eindruck davon zu verschaffen, welche Ergebnisse sich ohne Codierung beziehungsweise bei einer strengeren Auswahl der Matches anhand der Länge erreichen lassen.

3.3.2 Bewertung von Textübereinstimmungen

Durch die Überarbeitung der Matchdaten nach den bisher beschriebenen Kriterien kann zwar die Wahrscheinlichkeit erhöht werden, dass die gefundenen Übereinstimmungen durch Textbeziehungen zu erklären sind, vor allem bei einer relativ niedrig gewählten Mindestlänge ist aber damit zu rechnen, dass die Matches vielfach stattdessen auf sprachlich fest gefügten Mustern oder einfach auf zufällig gleicher Wortwahl beruhen, und auch längere Wortgleichheit in bestimmten Passagen muss nicht auf eine textuelle Verwandtschaft im engeren Sinne hindeuten, sondern kann auch auf die punktuelle Verwendung von Vorlagen, die oft weit verbreitet waren, zurückzuführen sein. Gerade bei der Formulierung von Geschäftsschriftgut ist es eine seit alten Zeiten etablierte Praxis, Textmuster als Basis zu verwenden. Während für eine Untersuchung zur Entwicklung derartiger Texte nun gerade diese Übereinstimmungen von besonderem Interesse sein dürften, kann es für andere Fragestellungen durchaus sinnvoll sein, sie auszuklammern.

Die Auswahl der relevanten Übereinstimmungen ist also sicherlich von der leitenden Fragestellung abhängig, und sie lässt sich in vielen Fällen nicht mit völliger Sicherheit vornehmen. Dementsprechend fehlt auch hier der Maßstab, der eine Bezifferung von *Precision* und *Recall* automatisierter Verfahren ermöglichen würde. Immerhin lassen sich einige Kriterien benennen, die bei der automatischen Einstufung von Matches genutzt werden können, wobei die konkrete Grenzziehung für die Fallunterscheidung vermutlich je nach Korpuszusammensetzung recht unterschiedlich ausfallen kann.

Ähnlich wie die Vorkommenshäufigkeit von Wort- und Zeichen-N-Grammen eine enorme Spannbreite aufweist und von wenigen hochfrequenten Formen aus schnell zu einer Vielzahl von selten oder sogar nur einmal belegten Formen hin abfällt,[576] lässt sich auch für die nach dem vorgestellten Verfahren ermittelten Übereinstimmungen feststellen, dass jedenfalls bei niedrig angesetzter Mindestlänge der Großteil von ihnen auf einer relativ kleinen Zahl von bestimmten Wortfolgen beruht. Hierbei handelt es sich also offenbar um feststehende Formulierungen,

[576] Vgl. oben Unterkapitel 2.3.1.

die – wie eben beschrieben – sprachlich etabliert waren oder auch auf das Kopieren von Mustertexten zurückzuführen sein können.

Die Bezifferung der jeweiligen Vorkommenshäufigkeit ist allerdings nicht ganz so klar wie bei N-Grammen, da die Grenzen eines übereinstimmenden Textstücks nicht notwendig bei jedem Textpaar die gleichen sind. Zwar kann durch die beschriebene Anpassung der Matches an Wortgrenzen ein Großteil der zufälligen Abweichungen vor einer Zählung der Vorkommenshäufigkeit eliminiert werden, aber auch bei formelhaften Wendungen muss mit gewissen Variierungsmöglichkeiten und dementsprechend unterschiedlichen Graden an Übereinstimmungen zwischen Matches gerechnet werden (bei der Einarbeitung von Textvorlagen in neue Zusammenhänge ohnehin), und schon aus der durch die Codierung nicht völlig aufzufangenden Schreibungsvarianz ergibt sich, dass die ermittelten Matches kürzer sein können als die Übereinstimmungen, die sich von einem sprachkundigen Leser ermitteln ließen. Wenn man also eine Zählung der Häufigkeit der ermittelten Match-Zeichenfolgen vornehmen will, muss man ein Verfahren festlegen, wie mit solchen unterschiedlichen Abgrenzungen zu verfahren ist.

Relativ leicht lässt sich ein Berechnungsschema implementieren, das für jedes Match eine durchschnittliche Häufigkeit der zugehörigen Zeichen oder auch den entsprechenden Maximalwert ermittelt. Dazu müssen zunächst aus den Positionsangaben aller Matches die Positionen der enthaltenen Einzelzeichen ermittelt und für diese dann berechnet werden, in wie vielen Matches sie vorkommen. Anschließend sind für den Durchschnittswert für jedes Match die Werte der Einzelzeichen zu addieren und durch die Länge des Matchs zu teilen. Die Berechnung erfolgt nach diesem Verfahren also nicht auf der Basis eines Vergleichs der Zeichenfolgen als *Types*, sondern vielmehr für jedes Einzelvorkommen separat, und die Frage der Abgrenzung relativ fester Formeln braucht gar nicht beantwortet zu werden, um für ein konkretes Match festzustellen, ob ein festzulegender Häufigkeits-Schwellenwert überschritten wird, der als Indikator für die Klassifikation als formelhaft dienen kann.

Ein Problem bei der Interpretation von Matches mit mittlerer Häufigkeit ergibt sich allerdings daraus, dass ein Mehrfachvorkommen nicht immer auf Formelhaftigkeit beruht, sondern auch ein Hinweis auf die stärkere Rezeption einer Textvorlage und damit gerade von besonderem Interesse sein kann, jedenfalls bei einer Fragestellung, wie sie im Hintergrund dieser Untersuchung steht. Da sich kaum theoretisch festlegen lässt, mit welcher maximalen Häufigkeit aufgrund solcher Traditionszusammenhänge zu rechnen ist, ist die Festlegung eines mittleren Schwellenwertes für die Klassifikation einer Übereinstimmung als formelhaft mit dem Risiko behaftet, größere Gruppen von verwandten Texten gar nicht erst in den Blick zu bekommen. Und umgekehrt ist bei einem relativ hoch angesetzten Schwellenwert zu erwarten, dass vielfach feste Formeln nicht als solche eingestuft werden, da selbst ein großes Textkorpus nur einen Ausschnitt der Ausdrucksmöglichkeiten einer Sprache erken-

nen lassen kann und erst recht nicht mit einer häufigen Belegung der Phänomene zu rechnen ist, sondern vielmehr auch Einzelvorkommen durchaus auf zugrunde liegenden Mustern beruhen können.[577]

Aus den Matchpositionen lassen sich neben der Häufigkeit im Korpus noch einige weitere Informationen ermitteln, die bei der Klassifikation der Matches herangezogen werden können:

- Grundsätzlich dürfte das gehäufte Vorkommen einer (in einem weiteren Sinne) signifikanten[578] Wortfolge in einem Text darauf hindeuten, dass es sich um eine feste Wendung handelt, die zu den sprachlichen Ausdrucksmitteln des jeweiligen Autors zählt, während ein jeweils einmaliges Vorkommen in einer Reihe von Texten ein Indiz dafür ist, dass diese Formulierung etwas Besonderes ist und die Textstellen etwas miteinander zu tun haben könnten. Natürlich ist auch hier mit Fällen zu rechnen, die keinem der beiden eben genannten Muster genau entsprechen – so können Übereinstimmungen innerhalb von Texten auch auf Querverweisen beruhen, und ohnehin gibt es viele Fälle von Texten, die eigentlich eine Sammlung enthaltener Einzeltexte sind, bei denen also das wiederholte Vorkommen bestimmter Formulierungen bei Verteilung auf diese Einzeltexte so zu interpretieren sein kann wie bei anderen Gruppen von Texten.
- Wenn sich zwischen zwei Texten eine Vielzahl von Übereinstimmungen feststellen lässt, die einander nicht oder nur selten überlappen und vielleicht auch noch nur durch relativ kurze Stücke dazwischen voneinander abgegrenzt sind, lässt das eine Textverwandtschaft vermuten. Bei isolierten Matches liegt es hingegen nahe, sie auf eine feste Formel oder auch eine zufällig gleiche Wortwahl zurückzuführen.
- Bei kurzen Übereinstimmungen ist prinzipiell damit zu rechnen, dass es sich um feststehende Formulierungen handelt, die mehr oder weniger allgemein verbreitet sein und ohne bewusstes Anknüpfen an eine bestimmte Texttradition verwendet werden können. Wenn Übereinstimmungen hingegen die Länge überschreiten, die solche Formeln in aller Regel haben, ist davon auszugehen, dass hier bewusst eine Vorlage zugrunde gelegt wird. Es kann sich dabei allerdings durchaus um ein Formulierungsmuster mit weiter Verbreitung handeln, etwa bei Texten, die auf Formularen basieren.

Eine Zusammenfassung der verschiedenen angeführten Kriterien in einer Formel, die für die automatische Klassifikation von Matches genutzt werden könnte, dürfte

[577] Dies betrifft schon die Wortebene: Typischerweise wird in einem relativ großen Korpus etwa die Hälfte der unterschiedlichen *Types* nur durch ein einziges *Token* vertreten, vgl. *http://en.wikipedia.org/wiki/Hapax_legomenon* (mit Literaturhinweis).

[578] Als nicht signifikant sind wohl – abgesehen von seltenen Ausnahmen – Wortfolgen zu betrachten, die außer Stoppwörtern (und anderen Funktionswörtern) kein oder nur ein einziges Wort umfassen.

sich kaum abstrakt begründen lassen; vielmehr ist, wie gesagt, mit starken Unterschieden je nach Zusammensetzung des zugrunde gelegten Textkorpus und der Art der interessierenden Übereinstimmungen zu rechnen. Die im Folgenden betrachteten Bewertungsverfahren beruhen auf Plausibilitätsüberlegungen und wurden anhand der Ergebnisse von Testläufen verfeinert, wobei das Ziel eine möglichst gute weitgehend automatische Abgrenzung zwischen solchen Matches war, die im Sinne der hier untersuchten Thematik relativ enger textueller Abhängigkeiten relevant sind, und solchen, die durch zufällige Übereinstimmungen oder sprachlich etablierte Muster erklärt werden können. Für andere Korpora und Fragestellungen ist mit Anpassungsbedarf zu rechnen, und es ist zu betonen, dass die Gewichtung sowie überhaupt die Berücksichtigung der einzelnen Faktoren in den vorgestellten Formeln in einem nicht unerheblichen Maße willkürlich ist.

Für die hier im Sinne einer präzisen Beschreibung vorgestellten Formeln werden folgende Termbezeichnungen zugrunde gelegt:

- Die einzelnen Texte werden mit T und einer laufenden Nummer als Subskript bezeichnet (zum Beispiel T_1), Buchstaben als Subskripte dienen als Variablen für die laufende Nummer.
- $freq(T_a, i)$ ist die (absolute) Häufigkeit des $i$ten Zeichens von T_a in allen Matches dieses Textes im Korpus (in den untersuchten Daten: einschließlich Matches mit anderen Stellen in T_a).
- $freq(T_a, T_b, i)$ ist die (absolute) Häufigkeit des $i$ten Zeichens von T_a in allen Matches dieses Textes mit Text T_b. Dementsprechend ist $freq(T_b, T_a, i)$ ist die (absolute) Häufigkeit des $i$ten Zeichens von T_b in allen Matches dieses Textes mit Text T_a.
- Die einzelnen Matches eines Textpaars werden mit M und einer laufenden Nummer als Subskript bezeichnet; auch hier dienen Buchstaben als Variablen für die Zählung. Um die Formeln nicht zu unübersichtlich werden zu lassen, wird das Textpaar dabei nicht angegeben, sondern es wird vorausgesetzt, dass es sich um die Texte T_a und T_b handelt.
- $start(M_x)$ ist die Position des ersten Zeichens des Matchs M_x im jeweils primär betrachteten Text T_a, $end(M_x)$ die entsprechende Endposition.
- $len(M_x)$ ist die Länge von M_x (also $end(M_x) - start(M_x) + 1$).
- $avw(M_x)$ ist $len(M_x)$ geteilt durch die durchschnittliche Wortlänge, also die fiktive Anzahl von Wörtern mit Durchschnittslänge im Match.
- $m(x, i)$ ist das Zeichen in T_b, das im Match M_x dem Zeichen i in T_a zugeordnet ist.
- $r(M_x)$ ist die vorgenommene Bewertung (*Rating*).

An erster Stelle soll ein relativ einfaches Bewertungsverfahren näher betrachtet werden, das zum einen die Überlappung von Matches im Korpus berücksichtigt, zum anderen die jeweilige Länge.

Oben auf S. 173 wurde schon ein Schema für die Bezifferung der Vorkommens-

häufigkeit von Matches beschrieben, das auf der Auszählung der Vorkommenshäufigkeit jedes einzelnen beteiligten Zeichens basiert. Entsprechend lassen sich auch Kehrwerte bilden und summieren:

$$r(M_x) = \sum_{i=start(M_x)}^{end(M_x)} \frac{1}{freq(T_a, i)} \tag{3.1}$$

Bei Verwendung dieser Formel ergibt sich eine Zahl, die um so höher ist, je länger das Match und je geringer die Überlappung mit einer Vielzahl anderer Matches ist; maximal – bei einem singulären Vorkommen – entspricht sie der Länge des Matchs, bei einem sehr hohen Übereinstimmungsgrad mit einer großen Zahl anderer Matches liegt sie nahe null. Längere Übereinstimmungen können also eine wesentlich höhere Bewertung erhalten als kürzere, aber das ist keineswegs immer der Fall; vielmehr gibt es im hier untersuchten Korpus auch vergleichsweise lange Matches mit einer großen Zahl von Entsprechungen, denen nach dieser Formel nur eine sehr niedrige Bewertung zugeordnet wird, und dies ist jedenfalls für die niedrigsten Werte sachlich durchaus korrekt.[579]

Die einfache Formel bietet den Vorteil, dass sich die ermittelten Bewertungen bei Kenntnis der Länge einer Übereinstimmung leicht interpretieren lassen. Wenn der Wert zum Beispiel in etwa die Hälfte der Matchlänge beträgt, liegt die Vermutung nahe, dass sich für die zu diesem Match gehörenden Textstücke noch eine weitere Parallelstelle finden lässt und dass die Grenzen der Übereinstimmungen einander weitgehend entsprechen.[580]

Abbildung 3.9 auf S. 183 veranschaulicht in einer Reihe von Säulendiagrammen, wie häufig bestimmte Bewertungen bei bestimmten Matchlängen von 20 bis 90 (in diesem Fall als exakte Werte, nicht als Mindestlängen) in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (nach den in Unterkapitel 3.3.1 beschriebenen Filterschritten 1–3) im untersuchten Korpus vorkommen. Für die Berechnung der Überlappung mit anderen Matches werden dabei hier wie auch in den Abbildungen 3.10 und 3.12–3.14[581] sämtliche Matches mit einer Mindestlänge von 18 Zeichen berücksichtigt, also nicht nur die mit der dem jeweiligen Diagramm zugrunde liegenden Länge. Die absoluten Werte sind zwar insbesondere auf der y-Achse recht unterschiedlich, aber es gibt in der Grundstruktur der Werteentwicklung einige auffällige Ähnlichkeiten.

Bemerkenswert ist vor allem, dass die Zahl der Fälle mit der jeweils möglichen

[579] Eine Überprüfung der Bewertungen unter 2 für Matches mit der Länge 90 in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen ergab, dass es sich dabei jeweils um einen Teil der kaiserlichen Titulatur handelt.

[580] Natürlich sind auch andere Konstellationen vorstellbar, wie etwa die, dass der Anfang eines Matchs dem Ende eines anderen entspricht und der Rest dem Anfang eines dritten. Dies wäre aber selbst bei zufälligen Übereinstimmungen recht unwahrscheinlich.

[581] Vgl. zu Abbildung 3.11 unten Anm. 584

Maximalbewertung – also singuläre Matches – um ein Vielfaches höher ist als die von Bewertungen, die etwas niedriger sind, und dass sich auch bei der Hälfte und zum Teil bei weiteren Teilern dieser Bewertung lokale Maxima feststellen lassen. Daneben ist aber auch deutlich zu erkennen, dass bei kürzeren Matches ein recht oder sogar sehr hoher Anteil nur eine sehr niedrige Bewertung erhält – bei der kürzesten hier betrachteten Länge liegt bei 35.656 Matches von insgesamt 41.554 die Bewertung unter 10, bei 30.540 unter 5 und bei 15.546 sogar unter 1, aber nur bei 3.104 bei 20, also beim Wert für ein singuläres Match dieser Länge.

Wenn man kurze Übereinstimmungen also zum Beispiel dann berücksichtigen möchte, wenn es maximal eine weitere Entsprechung mit derselben Abgrenzung gibt, reduziert sich die Zahl der noch zu betrachtenden Fälle ganz erheblich, es werden zugleich aber immer noch wesentlich mehr Matches für die weitere Untersuchung berücksichtigt als bei einer Auswahl allein auf der Basis einer deutlich größeren Mindestlänge. Von längeren Matches entfällt bei einer entsprechenden Grenzziehung hingegen nur ein geringer Anteil – so erhalten von den 748 Matches mit 70 Zeichen 73 und von den 430 Matches mit 90 Zeichen 24 eine Bewertung unter 10.

Formel 3.1 liefert zwar Werte, die im Hinblick auf eine bestimmte Codierung gut zu interpretieren sind, allerdings eignet sie sich nicht für den Vergleich verschiedener Codierungsvarianten, wenn diese zu einer unterschiedlich starken Textverdichtung führen. Um hierfür einen einheitlichen Maßstab zu haben, lässt sich wie schon oben[582] eine Umrechnung in Wörter mit durchschnittlicher Länge vornehmen. Dazu ist der nach Formel 3.1 ermittelte Wert durch die Länge des Matchs zu teilen und dann mit der der Matchlänge entsprechenden Zahl an Wörtern durchschnittlicher Länge zu multiplizieren.[583] Daraus ergibt sich Formel 3.2:

$$r(M_x) = \sum_{i=start(M_x)}^{end(M_x)} \frac{1}{freq(T_a, i)} \; : \; len(M_x) \cdot avw(M_x) \tag{3.2}$$

Bei einer Bewertung nach dieser Formel ist – analog zu Formel 3.1 – der erreichbare Maximalwert die Länge, umgerechnet in die Zahl von Wörtern mit durchschnittlicher Länge. Die Abbildungen 3.10 und 3.11 auf S. 184 und 185 veranschaulichen die Wertverteilung für zwei Codierungsvarianten, wobei die jeweils untersuchten Längen einander und denen in Abbildung 3.9 weitgehend entsprechen.[584] Offenbar

[582] Vgl. S. 147.

[583] Das ist äquivalent dazu, den Wert von 3.1 durch die durchschnittliche Zeichenzahl je Wort zu teilen. Die im Haupttext dargestellte Form wurde gewählt, weil sich auf dieser Basis die weiteren Formeln übersichtlicher gestalten lassen.

[584] Die für die Abbildungen 3.10 und 3.11 zugrunde gelegten Längen differieren zum Teil geringfügig, da aufgrund des unterschiedlichen Umrechnungsfaktors die den Längen von Abbildung

ergibt sich auch nach diesem Berechnungsverfahren eine Werteverteilung, die von der Grundstruktur her der von Formel 3.1 entspricht, wobei das Spektrum auf der x-Achse natürlich kleiner ist.

Die Werteunterschiede zwischen Abbildung 184 und 185 lassen sich auf mehrere Ursachen zurückführen: Bei einer Codierungsvariante mit Leerzeichen ist schon deshalb mit einer geringeren Zahl von Matches zu rechnen, weil manche Entsprechungen aufgrund einer Abweichung in der Leerzeichensetzung nicht gefunden werden können. Außerdem verteilen sich die Matches in der Variante ohne Leerzeichen auf eine geringere Zahl unterschiedlicher Längen; deshalb sind die Fallzahlen für die einzelnen Längen im Schnitt deutlich größer, als es dem Verhältnis der Matchzahlen insgesamt entspricht. Und drittens gibt es zwischen den beiden Abbildungen zwar (ungefähre) Entsprechungen hinsichtlich der Umrechnung in Wörter durchschnittlicher Länge, inwieweit die zugrunde liegenden originalen Textstücke jeweils übereinstimmen, ist damit aber nicht gesagt. Auch wenn eine bestimmte Entsprechung über beide Codierungsvarianten gefunden wird und die Abgrenzung im Original exakt übereinstimmt, können die Längen der entsprechenden codierten Textstücke einander hinsichtlich der Zahl von Durchschnittswörtern unterschiedlich groß sein. Da hier in jedem Einzeldigramm nur die Matches der angegebenen Länge berücksichtigt werden, macht sich bei den relativ niedrigen Fallzahlen, die den Diagrammen für größere Längen zugrunde liegen, deutlich bemerkbar, ob eine der zugrunde liegenden Zeichenfolgen häufig vorkommt.

Eine Prüfung von Stellen mit einer größeren Länge, aber eher niedriger Bewertung nach einer der beiden bisher vorgestellten Formeln zeigt allerdings, dass es sich vielfach doch um recht signifikante Formulierungen handelt. Und da exakte Textübereinstimmung in einer Passage, deren Länge die von sprachlich fest etablierten Mustern überschreitet, grundsätzlich ein recht deutliches Zeichen für die Verwendung von Vorlagen ist, scheint es sachlich durchaus angemessen, solche längeren Übereinstimmungen höher zu gewichten, zumal die beiden bisher beschriebenen Bewertungsschemata die Erkennung von Textübernahmen dann unterminieren können, wenn es sich dabei nicht um Einzelbeziehungen, sondern um Gruppen von Texten mit entsprechend häufiger Überlappung von Matches handelt.

Um die Länge der Übereinstimmungen stärker einfließen zu lassen, kann man sie mit einem Exponenten versehen, also zum Beispiel ihren Quadratwert zugrunde legen:

3.9 entsprechende durchschnittliche Wortzahl auch bei Rundung auf eine Nachkommastelle nicht unbedingt eine exakte Entsprechung in den Daten für die Codierung *ohne A/I/U/B/F* mit Leerzeichen hat. Abbildung 3.11 liegen für die Ermittlung der Überlappungen alle Matches zugrunde, die bei einer Ermittlung aller MEMs mit der Mindestlänge 25 und der anschließenden Überarbeitung der Matchdaten nach den Regeln der Filterschritte 1–3 (vgl. oben Unterkapitel 3.3.1) verbleiben. Die Mindestlänge 25 wird hier deshalb angesetzt, weil sie etwa 6,98 Wörtern durchschnittlicher Länge entspricht und damit fast exakt der gleichen Zahl wie die Mindestlänge 18 für die Codierung *ohne A/I/U/B/F* ohne Leerzeichen.

$$r(M_x) = \sum_{i=start(M_x)}^{end(M_x)} \frac{1}{freq(T_a, i)} \; : len(M_x) \cdot avw(M_x)^2 \tag{3.3}$$

Abbildung 3.12 auf S. 186 ist parallel zu Abbildung 3.10 gestaltet, basiert aber auf Formel 3.3. Die relative Werteverteilung in jedem Einzeldiagramm ist im Prinzip die gleiche wie in Abbildung 3.9 und 3.10, da der hinzugekommene Faktor für eine bestimmte Länge stets derselbe ist, es ergibt sich allerdings eine größere Spreizung des Wertespektrums.[585] Und insbesondere vergrößert sich der Unterschied zwischen den Wertebereichen der Einzeldiagramme, so dass ein Schwellenwert, der zum Beispiel wie oben so gewählt ist, dass die kürzesten Matches maximal zwei Entsprechungen haben können, einen merklich geringeren Anteil längerer Matches ausschließt als nach dem ersten Verfahren.[586]

Eine weitere Variante soll zusätzlich auch das Ausmaß an Überlappungen mit anderen Matches desselben Textpaars berücksichtigen. Sie werden zwar schon bei der Zählung der Gesamtvorkommen berücksichtigt, aber wenn man davon ausgeht, dass die zu untersuchenden Texte jeweils für sich keine oder kaum gezielte Textwiederholungen enthalten, können, wie schon erwähnt, solche Matchüberlappungen auch dann ein Indiz für die Formelhaftigkeit der betreffenden Passagen sein, wenn sich das von der Gesamthäufigkeit her noch nicht unbedingt nahelegt. Bei der Auswertung, die den Diagrammen in Abbildung 3.13 auf S. 187 zugrunde liegt, wurde im Wesentlichen dieselbe Formel verwendet wie für die Daten in Abbildung 3.12, allerdings anschließend die Summe der Vorkommenshäufigkeiten der betreffenden Einzelzeichen beider beteiligten Texte in den verschiedenen Matches eines Textpaars, geteilt durch die Matchlänge, abgezogen und zuletzt wieder 2 addiert:

$$\begin{aligned} r(M_x) = & \sum_{i=start(M_x)}^{end(M_x)} \frac{1}{freq(T_a, i)} \; : len(M_x) \cdot avw(M_x)^2 \\ & - \left(\sum_{i=start(M_x)}^{end(M_x)} freq(T_a, T_b, i) \; + \sum_{i=start(M_x)}^{end(M_x)} freq(T_b, T_a, m(x, i)) \right) : len(M_x) \\ & + \; 2 \end{aligned} \tag{3.4}$$

[585] Aufgrund dieser Spreizung werden hier in den einzelnen Diagrammen jeweils unterschiedlich große Wertebereiche in einer Säule zusammengefasst, daraus ergeben sich gewisse Abweichungen im optischen Gesamteindruck.

[586] Der Schwellenwert des Beispiels liegt etwa bei $7{,}7^2/2$, also etwa bei 30. Darunter liegen bei einer Matchlänge von 7,7 Durchschnittswörtern 35.656 von 41.554, bei einer Länge von 27,1 Durchschnittswörtern 25 von 748 und bei einer Länge von 34,9 Durchschnittswörtern 16 von 430 Matches.

Damit bleibt bei Matches ohne Überlappungen innerhalb eines Textpaars der Wert unverändert (da der Subtrahend genau 2 beträgt und dies durch die Addition wieder ausgeglichen wird), der zu subtrahierende Wert nimmt aber – bei gleicher Abgrenzung der Matches – mit jedem weiteren entsprechenden Textstück um 2 zu.

Aus den Diagrammen in Abbildung 3.13 auf S. 187 kann der Effekt dieser Subtraktion allerdings nur begrenzt abgelesen werden, da hier aus darstellerischen Gründen alle Werte unter null in einer einzigen Säule zusammengefasst sind – dabei liegt auch die Annahme zugrunde, dass ein Schwellenwert für die Abgrenzung zwischen näher zu betrachtenden und mit hoher Wahrscheinlichkeit irrelevanten Matches höher liegen dürfte und deshalb die genaue Wertverteilung in diesem Bereich für diese Untersuchung keine Rolle spielt.

Das gerade beschriebene Bewertungsverfahren führt natürlich zu maximal gleichen, in vielen Fällen aber etwas und einigen Fällen viel niedrigeren Werten als das zuvor vorgestellte. Dementsprechend steigt der Anteil der Matches, denen eine Bewertung unterhalb eines bestimmten Schwellenwertes zugeordnet wird, allerdings sind die Unterschiede bei höheren Bewertungen und/oder längeren Matches gering, während kurze Matches mit ohnehin sehr niedriger Bewertung in vielen Fällen noch einmal deutlich herabgestuft werden.

Eine nicht unerhebliche Schwäche der bisher vorgestellten Bewertungsformeln – wie auch der Untersuchung von Textbeziehungen auf der Basis längerer exakter Übereinstimmungen – besteht darin, dass die tatsächlichen Textentsprechungen zum Teil wesentlich umfangreicher sind, als über die Matchpositionen beschrieben wird, weil die Matches öfters nur durch kleine Textvarianten unterbrochen werden. Selbst wenn die für die Erkennung vorausgesetzte Mindestlänge von den Teilstücken einer Textentsprechung jeweils erreicht wird, werden solche Teilstücke bei einer Bewertung, die insbesondere auch die Matchlänge zugrunde legt, wesentlich schlechter eingestuft, als wenn sie ein zusammenhängendes Match bilden würden. Eine Verbesserung kann deshalb darin bestehen, in die Bewertung eines Matchs auch die von nahe benachbarten Matches mit einzubeziehen.

Hierbei ergeben sich allerdings verschiedene Schwierigkeiten. Zunächst einmal lässt sich nicht allgemein festlegen, bis zu welcher Grenze eine solche Berücksichtigung des Textumfeldes sinnvoll ist. Auch wenn zwei Matches nicht nur durch unterschiedliche Einzelwörter getrennt sind, sondern zum Beispiel durch einen eingefügten Teilsatz, ist die Wahrscheinlichkeit eines Zusammenhangs natürlich sehr groß, und auch längere Einschübe sind jedenfalls für das hier untersuchte Korpus nicht ungewöhnlich.

Daneben steht ein praktisches Problem: Es ist nicht unbedingt offensichtlich, welche Paare von Matches überhaupt als beste Kandidaten für eine Bewertung als benachbart in Frage kommen. Ein Match, das im Text T_a einen geringen Abstand zum Match M_x hat, kann im Text T_b an einer weit entfernten Stelle liegen, und

es kann andere Matches geben, die in T_b näher bei M_x stehen. Es liegt wohl nahe, eine Auswahl danach zu treffen, bei welchem Match die Summe der Abstände in beiden Texten am geringsten ist.[587] Um für alle Matches die in diesem Sinne nächstgelegenen darauf folgenden Matches zu ermitteln, ohne dabei jedes Match in Beziehung zu jedem anderen setzen zu müssen, bietet es sich an, zunächst einmal eine Sortierung nach den Positionen zum Beispiel innerhalb von T_a vorzunehmen. Dann kann für die in dieser sortierten Liste auf das jeweils untersuchte Match folgenden Einträge ermittelt werden, wie groß die Abstandssumme ist, und zwar bis zu dem Punkt, an dem der Abstand in T_a genau so groß ist wie die geringste bisher ermittelte Abstandssumme. Entsprechend lässt sich auch für die Ermittlung der nächstgelegenen vorangehenden Matches verfahren, wobei natürlich die Endpunkte der Matches für die Abstandssumme ausschlaggebend sind und dementsprechend eine andere Sortierung zugrunde zu legen ist.

Auch bei diesem Vorgehen kann sich für ein Korpus wie das hier untersuchte vor allem bei relativ kurzen Mindestmatchlängen ein nicht unerheblicher Verarbeitungsaufwand ergeben, wenn für jedes Match die Nachbarmatches davor und danach ermittelt werden sollen. Für kurze Matches ist davon auszugehen, dass die Matchpositionen zu einem großen Teil in keiner Beziehung zueinander stehen und deshalb ausgehend von M_x zwar möglicherweise viele andere Matches im näheren Umfeld in T_a zu finden sind und auch viele Matches im näheren Umfeld in T_b, aber nur in eher seltenen Fällen Matches, die in beiden Texten nahe benachbart sind. Dementsprechend ist die Zahl der für ein einzelnes Match im Durchschnitt erforderlichen Überprüfungen deutlich höher als bei längeren Matches,[588] und sie ist zu multiplizieren mit der hohen Gesamtzahl solcher kurzen Matches, wobei freilich die einzelnen Überprüfungen sehr schnell durchgeführt werden können.

Soweit es aber nicht darum geht, für jeden Fall die optimale Nachbarkombination zu ermitteln, sondern primär darum, kurze Matches aufzuwerten, die sich problemlos in eine längere Folge von Matches einordnen lassen, dürfte es vielfach ausreichen, eine einfache Prüfung anhand der Sortierung nach den Positionen in T_a vorzunehmen und eventuell bei ausgeprägten Überlappungen auch die nächstfolgenden beziehungsweise -vorangehenden Matches ohne eine solche Überlappung in die Prüfung mit einzubeziehen.[589] Allein schon die Tatsache, dass die Reihen-

[587] Für Matches, die in beiden Texten in unterschiedlicher Reihenfolge vorkommen, ist außerdem zu entscheiden, ob sie überhaupt entsprechend diesem Kriterium einander als Nachbarn zugeordnet werden können. Wenn das bejaht wird, sollte natürlich jeweils der absolute Wert des Abstands in die Rechnung einfließen und kein negativer Wert. Außerdem ist bei der Auswahl darauf zu achten, dass keine größeren Überlappungen vorliegen (vgl. dazu im Folgenden Anm. 589).

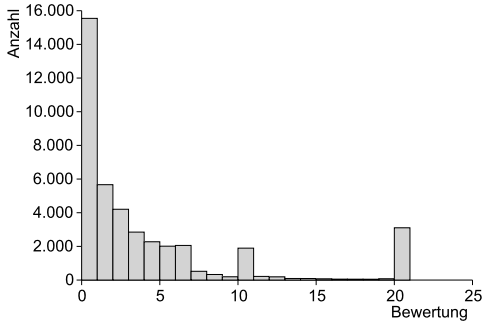
[588] Beispielsweise erforderte die Ermittlung der nächstgelegenen nachfolgenden Matches in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen durchschnittlich 24,4 Vergleiche für die Mindestlänge 16, 10,8 für die Mindestlänge 24 und 3,9 für die Mindestlänge 66.

[589] Dass Matches mit ausgeprägten Überlappungen nicht als benachbart einzustufen sind, dürfte sich von selbst verstehen. Matches mit kurzen Überlappungen hingegen kommen deshalb als

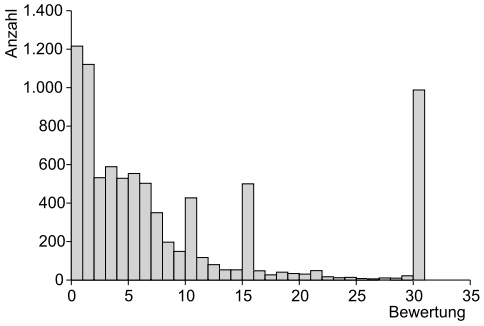
folge einer nicht geringen Zahl von Matches in zwei Texten gleich ist, kann ein Indiz für einen tatsächlichen Traditionszusammenhang sein. Eine Lösung, die auch bei Matchfolgen mit Sortierproblemen eine recht gute Ermittlung von benachbarten Matches ermöglicht, kann insbesondere für den Feinvergleich von Textpaaren wichtig sein und wird deshalb in Unterkapitel 3.4.5 vorgestellt.

Abbildung 3.14 auf S. 188 stellt auch hierzu eine Reihe von Diagrammen zusammen, wobei die Berechnung auf Formel 3.4 basiert und dann, wenn nach dem in Unterkapitel 3.4.5 entwickelten Verfahren ein vorangehendes oder folgendes Match gefunden wird, dessen Länge als Summand hinzugefügt wird. Aus darstellerischen Gründen sind hier nicht nur am unteren, sondern auch am oberen Rand der x-Achse Säulen zusammengefasst, so dass das Wertespektrum in etwa dem von Abbildung 3.13 entspricht.

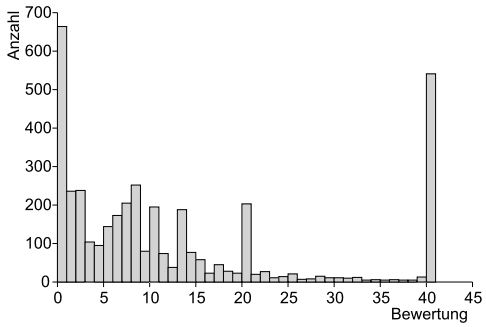
Kandidaten für eine Nachbarschaftszuordnung (unter Kürzung eines der Matches) in Frage, weil die Überlappung auf einer sachlich irreführenden Matchausdehnung zum Beispiel aufgrund von Stoppwörtern oder feststehenden Wortfolgen im Randbereich beruhen kann.



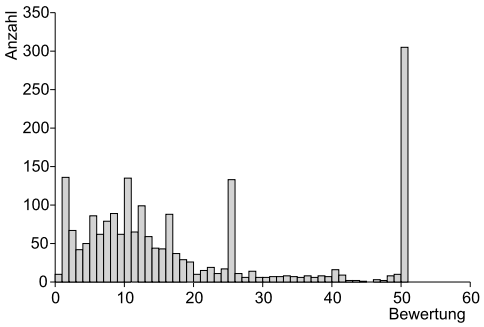
(a) Matches mit genau 20 Zeichen



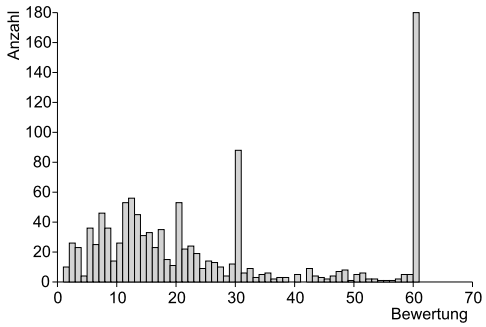
(b) Matches mit genau 30 Zeichen



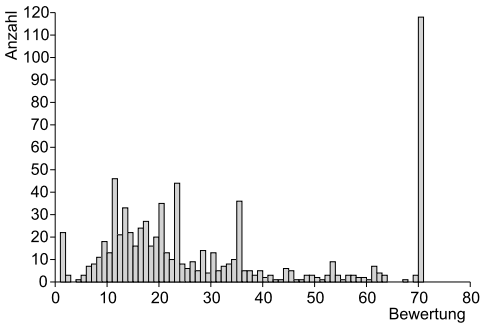
(c) Matches mit genau 40 Zeichen



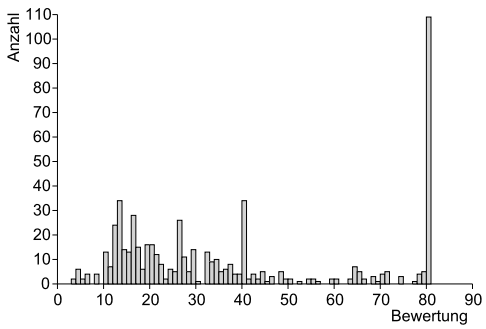
(d) Matches mit genau 50 Zeichen



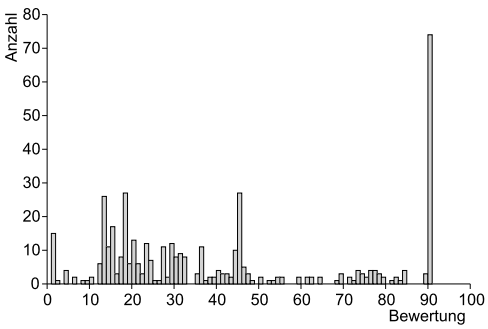
(e) Matches mit genau 60 Zeichen



(f) Matches mit genau 70 Zeichen

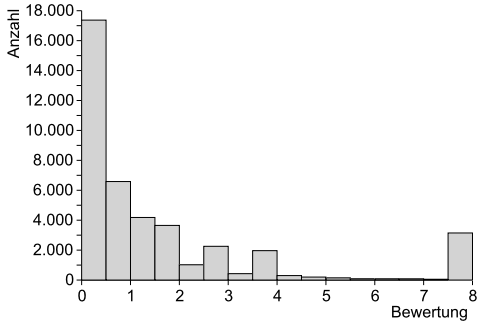


(g) Matches mit genau 80 Zeichen

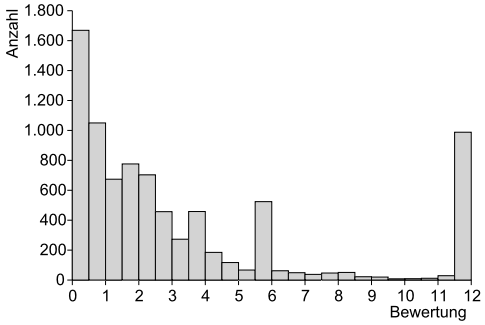


(h) Matches mit genau 90 Zeichen

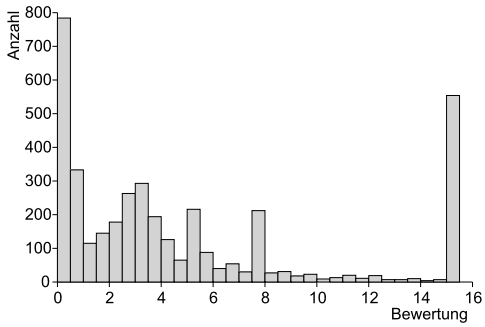
Abb. 3.9: Bewertung von Matches der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (nach Filterschritt 1–3) entsprechend Formel 3.1 (S. 176)



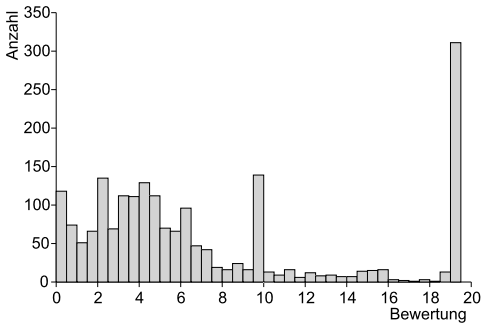
(a) Matchlänge 7,7 Durchschnittswörter



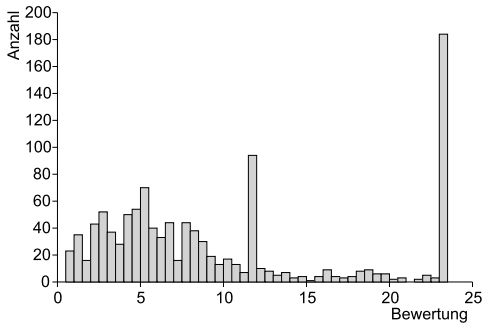
(b) Matchlänge 11,6 Durchschnittswörter



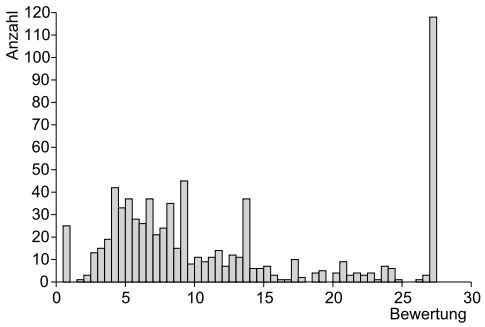
(c) Matchlänge 15,5 Durchschnittswörter



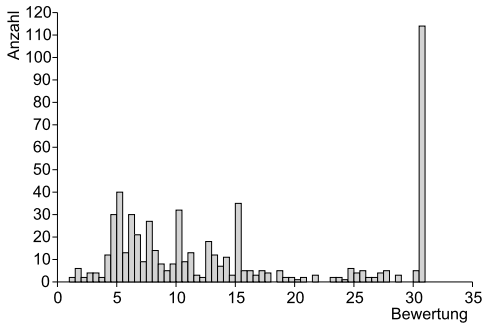
(d) Matchlänge 19,4 Durchschnittswörter



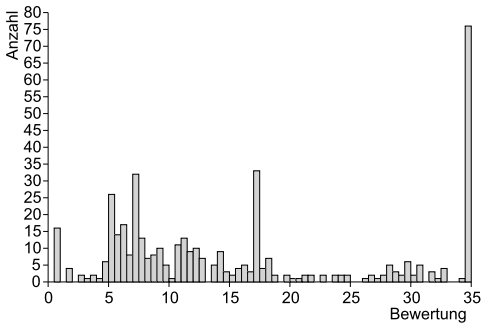
(e) Matchlänge 23,2 Durchschnittswörter



(f) Matchlänge 27,1 Durchschnittswörter

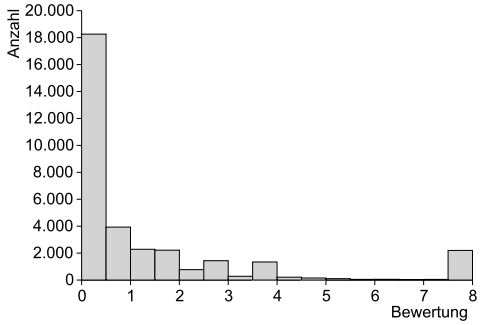


(g) Matchlänge 31 Durchschnittswörter

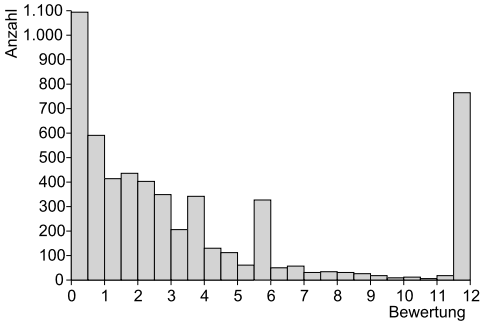


(h) Matchlänge 34,9 Durchschnittswörter

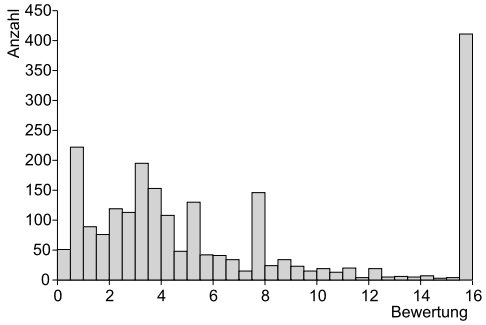
Abb. 3.10: Bewertung von Matches der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (nach Filterschritt 1–3) entsprechend Formel 3.2 (S. 177)



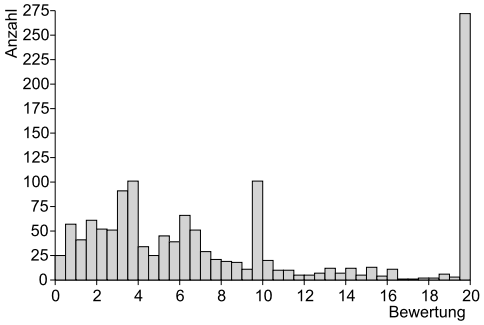
(a) Matchlänge 7,8 Durchschnittswörter



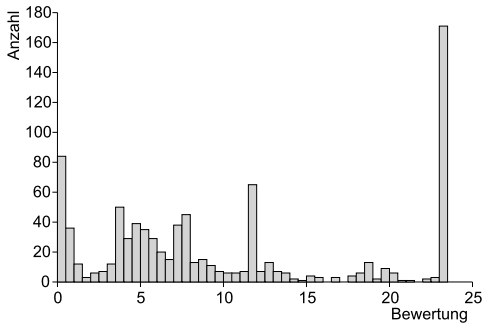
(b) Matchlänge 11,7 Durchschnittswörter



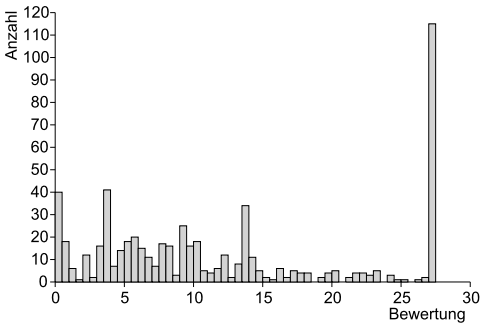
(c) Matchlänge 15,6 Durchschnittswörter



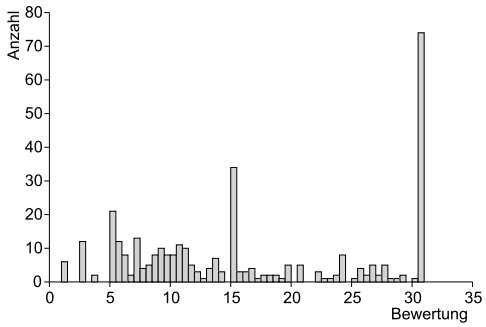
(d) Matchlänge 19,5 Durchschnittswörter



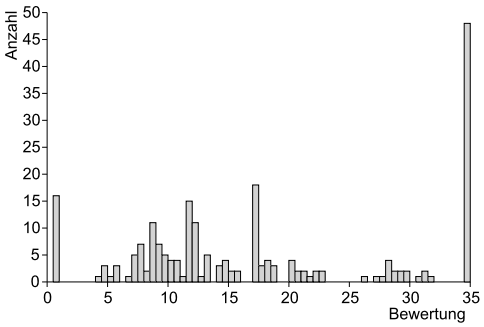
(e) Matchlänge 23,2 Durchschnittswörter



(f) Matchlänge 27,1 Durchschnittswörter

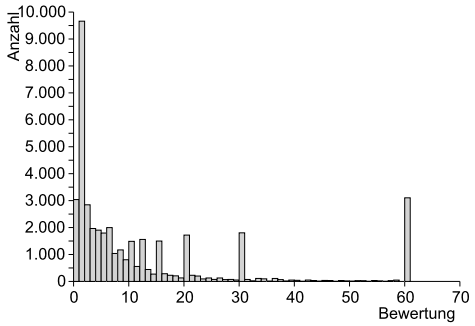


(g) Matchlänge 31 Durchschnittswörter

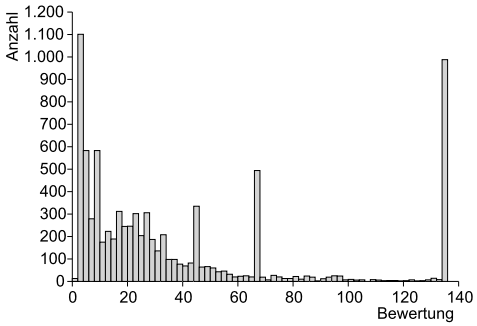


(h) Matchlänge 34,9 Durchschnittswörter

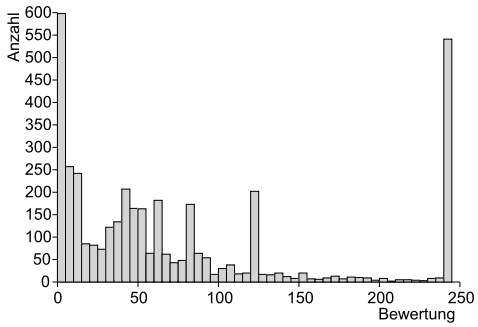
Abb. 3.11: Bewertung von Matches der Codierung *ohne A/I/U/B/F* mit Leerzeichen (nach Filterschritt 1–3) entsprechend Formel 3.2 (S. 177)



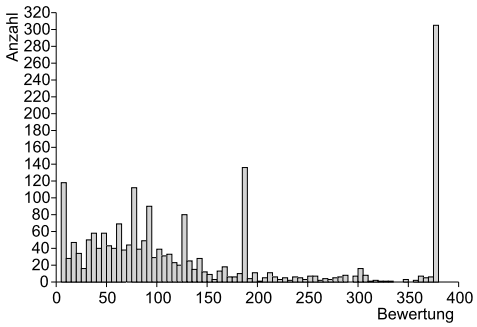
(a) Matchlänge 7,7 Durchschnittswörter



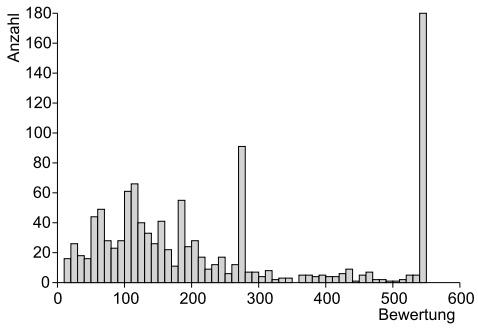
(b) Matchlänge 11,6 Durchschnittswörter



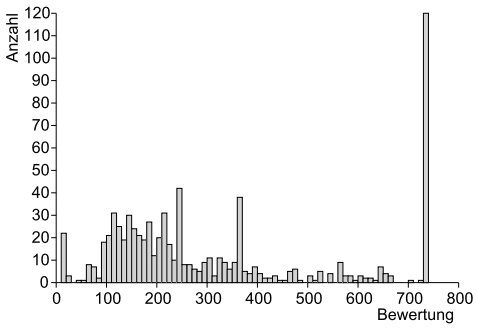
(c) Matchlänge 15,5 Durchschnittswörter



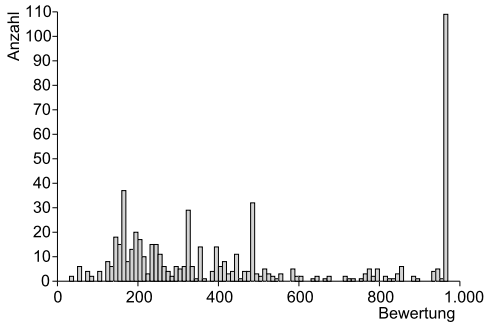
(d) Matchlänge 19,4 Durchschnittswörter



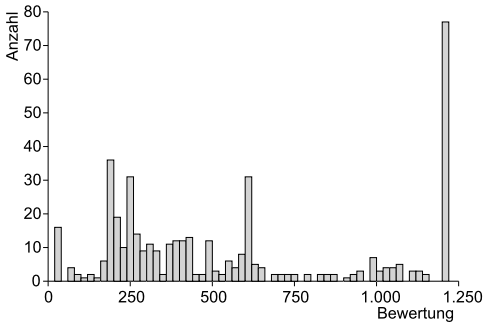
(e) Matchlänge 23,2 Durchschnittswörter



(f) Matchlänge 27,1 Durchschnittswörter

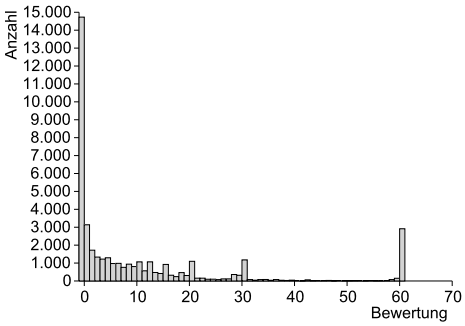


(g) Matchlänge 31 Durchschnittswörter

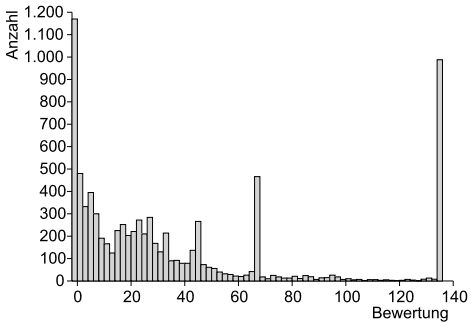


(h) Matchlänge 34,9 Durchschnittswörter

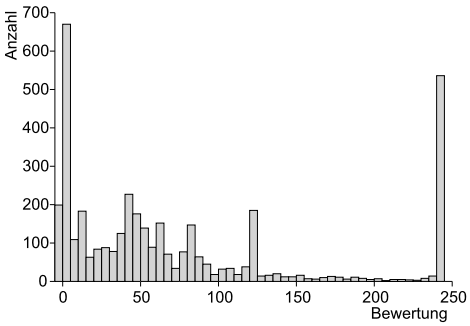
Abb. 3.12: Bewertung von Matches der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (nach Filterschritt 1–3) entsprechend Formel 3.3 (S. 179)



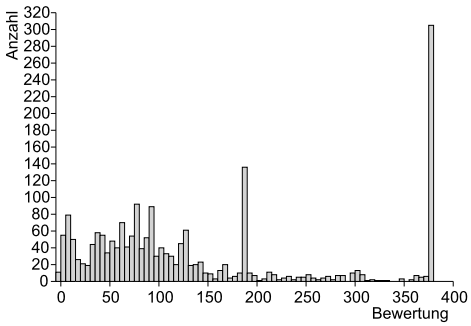
(a) Matchlänge 7,7 Durchschnittswörter



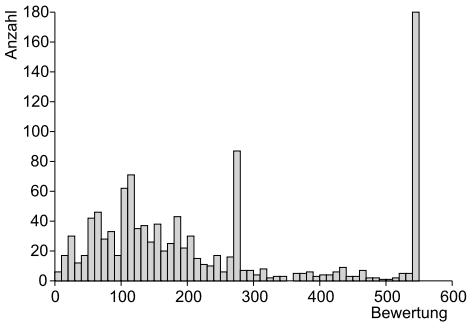
(b) Matchlänge 11,6 Durchschnittswörter



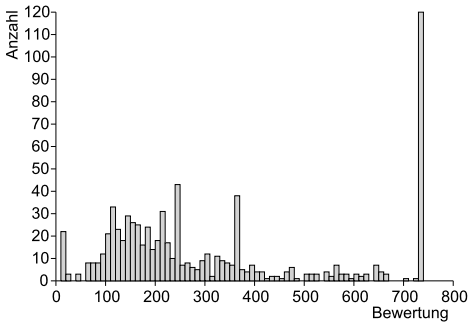
(c) Matchlänge 15,5 Durchschnittswörter



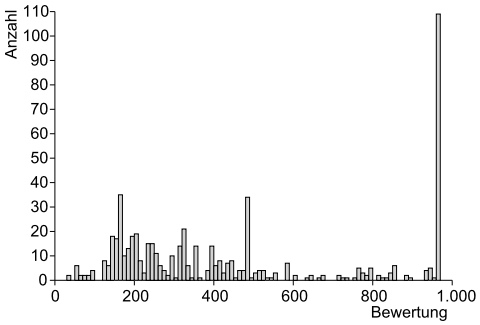
(d) Matchlänge 19,4 Durchschnittswörter



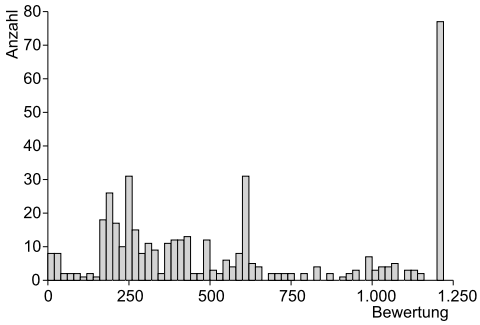
(e) Matchlänge 23,2 Durchschnittswörter



(f) Matchlänge 27,1 Durchschnittswörter

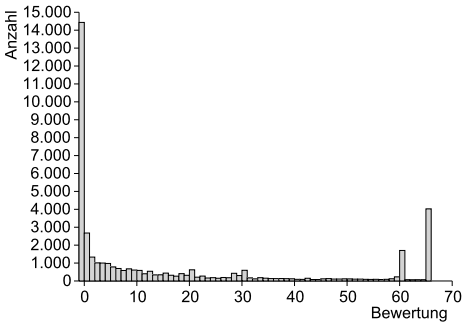


(g) Matchlänge 31 Durchschnittswörter

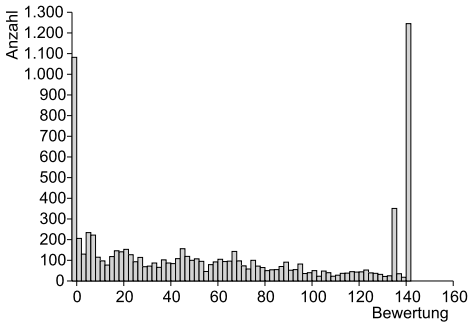


(h) Matchlänge 34,9 Durchschnittswörter

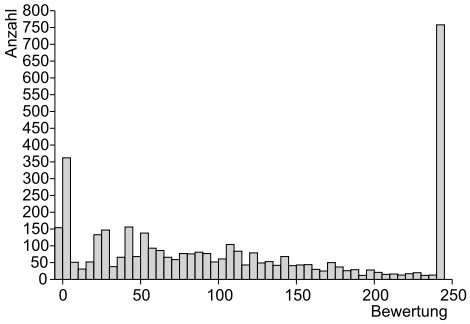
Abb. 3.13: Bewertung von Matches der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (nach Filterschritt 1–3) entsprechend Formel 3.4 (S. 179); alle negativen Werte sind in einer Säule zusammengefasst



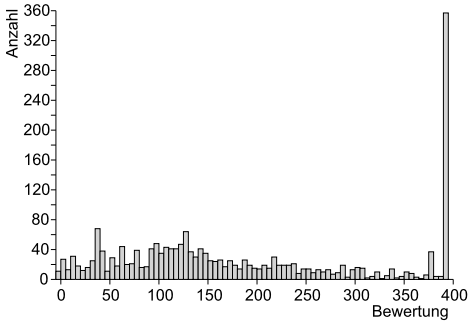
(a) Matchlänge 7,7 Durchschnittswörter



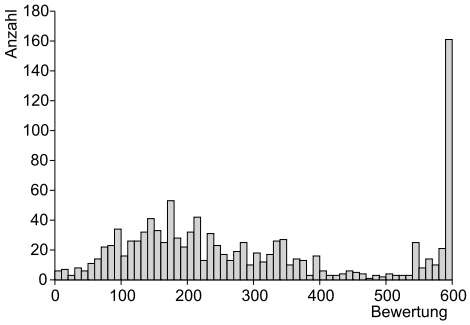
(b) Matchlänge 11,6 Durchschnittswörter



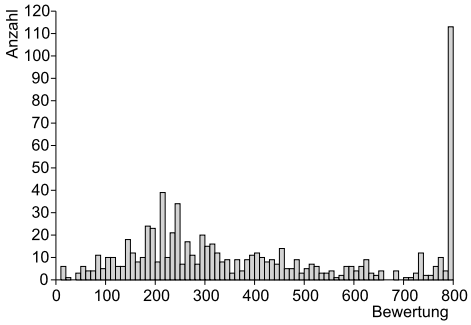
(c) Matchlänge 15,5 Durchschnittswörter



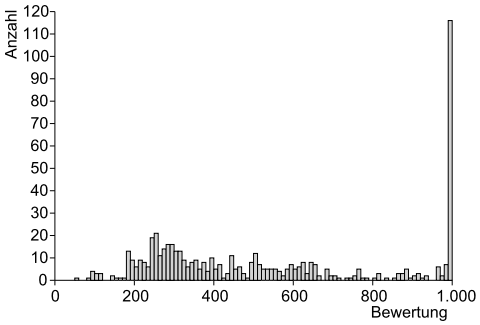
(d) Matchlänge 19,4 Durchschnittswörter



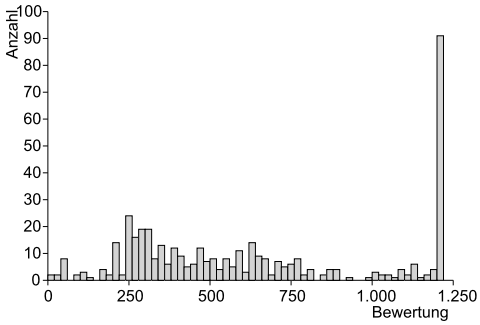
(e) Matchlänge 23,2 Durchschnittswörter



(f) Matchlänge 27,1 Durchschnittswörter



(g) Matchlänge 31 Durchschnittswörter



(h) Matchlänge 34,9 Durchschnittswörter

Abb. 3.14: Bewertung von Matches der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (nach Filterschritt 1–3) entsprechend Formel 3.4 (S. 179) mit Einbeziehung von Nachbarmatches; alle negativen und alle sehr hohen Werte sind in einer Säule zusammengefasst

3.4 Auswertung und Anzeige von Matches

In diesem Kapitel soll es darum gehen, wie die in der bereits dargestellten Weise ermittelten und gegebenenfalls nach Signifikanzkriterien gefilterten Matches ausgewertet beziehungsweise aufbereitet werden können, um sie für die Untersuchung textueller Beziehungen zu nutzen. Ähnlich wie die Forschungsgruppe um G. Heyer eine Präsentation von Vergleichsergebnissen im *macro view* und im *micro view* vorsieht,[590] soll dabei insbesondere zwischen einer eher mathematisch orientierten Datenauswertung im Überblick und der Rückbindung an die originalen Textstellen für die vergleichende Textlektüre unterschieden werden.

3.4.1 Quantifizierung der Matches zwischen Textpaaren

Ein einfacher Ansatz, um anhand der ermittelten Matchpositionen einen Überblick über das Ausmaß an Übereinstimmung zwischen Texten oder auch Textabschnitten zu gewinnen, besteht in der Zählung der Zeichen, die sich den Matches in dem jeweils untersuchten Text- oder Textabschnittspaar zuordnen lassen. Eng miteinander verwandte Texte, die zu großen Teilen wörtliche Entsprechungen aufweisen, sind dann an hohen Werten leicht zu erkennen.

Die ermittelten absoluten Zahlen können freilich bei kurzen Texten nur entsprechend niedrig ausfallen. Es bietet sich deshalb an, die Zahl der Zeichen in Matches in Relation zur Textlänge zu setzen. Dabei ist allerdings mit unterschiedlichen Fällen zu rechnen. Neben Textpaaren, die in allen oder fast allen Abschnitten Entsprechungen aufweisen, gibt es solche, bei denen einer der Texte als ganzer einem Teil des anderen zugeordnet werden kann, und solche, bei denen es zwar umfangreiche Entsprechungen gibt, diese sich aber jeweils vor allem oder ausschließlich in einem gewissen Bereich finden und deshalb nur einen geringen Anteil des Gesamttextes ausmachen. Die Erkennung aller umfangreicheren Textübernahmen ist also bei Ansetzung eines hohen prozentualen Schwellenwerts nicht gewährleistet.

Der *Recall* lässt sich immerhin verbessern, wenn sowohl der absolute Umfang der Matches als auch ihr Anteil an den betreffenden Gesamttexten[591] berücksichtigt werden, etwa in der Form, dass für den absoluten Umfang eine Mindestzahl als erster Schwellenwert festgelegt wird und eine Klassifikation eines Textpaars als irgendwie verwandt dann erfolgt, wenn außerdem ein höherer zweiter absoluter Schwellenwert oder in einem der Texte ein bestimmter Mindestanteil erreicht wird. Was für Grenzwerte dabei sinnvollerweise anzusetzen sind, hängt natürlich – wie so vieles bei der Untersuchung von Textbeziehungen – von den jeweiligen Rahmen-

[590] Vgl. zum Beispiel Gessner 2010, vor allem S. 31–35 oder Büchler u. a. 2010, vor allem S. 11–13.

[591] Sofern entsprechend S. 170 Matches in bestimmten Textstücken wie zum Beispiel Inhaltsverzeichnissen und Registern ausgeklammert werden, legt es sich nahe, hier jeweils die Länge der Texte nach Eliminierung dieser Stücke zugrundezulegen.

bedingungen ab, hier insbesondere davon, inwieweit es nur um sehr eng oder auch um entfernter verwandte Texte geht, in welchem Ausmaß mit Schreibvarianten zu rechnen ist, die nicht gleich codiert werden, und wie *Precision* und *Recall* gewichtet werden.

In Tabelle 3.10 ist in mehreren Spalten zusammengestellt, mit welcher Häufigkeit im untersuchten Korpus die summierte Länge der entsprechend den drei Filterschritten der Tabelle 3.9 ausgewählten und gekürzten Matches der MEM-Mindestlänge 18 in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen bestimmte Anteile des Gesamttextes erreicht, wenn die summierte Matchlänge in einem Textpaar nicht als Auswahlkriterium herangezogen wird oder wenn nur Textpaare berücksichtigt werden, bei denen diese Länge mindestens 100, 500, 1.000 beziehungsweise 10.000 Zeichen beträgt.[592] Zu Vergleichszwecken finden sich in Tabelle 3.11 die entsprechenden Werte für entsprechend gefilterte Matches ab einer MEM-Länge von mindestens 48 Zeichen – also bei einem Wert, bei dem nach Tabelle 3.7 auf S. 157 im Hinblick auf die Zahl der Matches nur etwa ein Zehntel, im Hinblick auf die gesamte Matchlänge nur etwa ein Drittel der Entsprechungen gefunden wird, die bei Ansetzung einer Mindestlänge von 18 Zeichen zu verzeichnen sind –, in Tabelle 3.12 die Werte für gefilterte Matches ab einer MEM-Länge von 18 Zeichen bei einem Vergleich auf der Basis einer Textumwandlung in Kleinbuchstaben und in Tabelle 3.13 die Daten für die Codierung *ohne A/I/U/B/F* ohne Leerzeichen und die Mindestlänge 18 bei Einbeziehung aller MEMs ohne irgendeine Überarbeitung.

Es ist wohl unmittelbar zu entnehmen, dass bei den Matchkriterien der Tabelle 3.10 jedenfalls im hier untersuchten Korpus der allergrößte Teil der ermittelten Textpaare hinsichtlich der absoluten oder der relativen Länge der Übereinstimmungen in keiner Weise auffällig ist, und in Anbetracht der niedrigen Mindestmatchlänge erklärt sich das leicht durch eine zufällige beziehungsweise durch feststehende Wendungen bedingte Gleichheit kürzerer Wortfolgen. Bei immerhin 921 Textpaaren macht der durch die Matches abgedeckte Teil des jeweils ersten Texts mindestens 5 % aus, und von diesen 921 beträgt bei 844 die summierte Länge der Übereinstimmungen mindestens 100 Zeichen. Die Differenz kann nur auf Matches in Texten zurückzuführen sein, die selbst weniger als 2000 Zeichen lang sind, da andernfalls eine Übereinstimmung von weniger als 100 Zeichen nicht zu einer Übereinstimmung von mindestens 5 % führt. Dass die Differenz so groß ist, erklärt sich daraus, dass im Projekt DRQEdit Sammelwerke als eine Reihe von Einzeltexten behandelt werden und die Textdatei für ein Sammelwerk selbst nur das enthält, was sich

[592] Textpaare sind jeweils doppelt berücksichtigt (beim zweiten Mal mit umgekehrter Textreihenfolge), allerdings führt das nicht einfach zu verdoppelten Werten, sondern es können sich auch ungerade Zahlen ergeben. Das beruht darauf, dass zum einen auch die Ermittlung von Übereinstimmungen innerhalb eines Textes einbezogen ist (dies allerdings natürlich nicht doppelt) und zum anderen – wie gerade schon beschrieben – die prozentualen Anteile im ersten und zweiten Text stark voneinander abweichen können. Ein Textpaar wird in einer Zeile nur dann doppelt gezählt, wenn der in dieser Zeile angesetzte Grenzwert in beiden Texten überschritten wird.

Matchanteil (1. Text)	alle Textpaare	mindestens 100 Matchzeichen	mindestens 500 Matchzeichen	mindestens 1.000 Matchzeichen	mindestens 10.000 Matchzeichen
mindestens 0 %	21.739	7.843	2.258	1.365	218
mindestens 5 %	921	844	686	610	193
mindestens 10 %	460	454	349	322	158
mindestens 15 %	310	307	244	219	130
mindestens 20 %	228	228	191	167	109
mindestens 25 %	178	178	159	137	92
mindestens 30 %	140	140	129	107	75
mindestens 35 %	114	114	109	87	64
mindestens 40 %	93	93	91	70	55
mindestens 45 %	82	82	82	63	50
mindestens 50 %	66	66	66	52	42
mindestens 55 %	55	55	55	44	37
mindestens 60 %	43	43	43	35	30
mindestens 65 %	38	38	38	31	27
mindestens 70 %	33	33	33	28	24
mindestens 75 %	32	32	32	27	24
mindestens 80 %	28	28	28	24	21
mindestens 85 %	22	22	22	19	16
mindestens 90 %	17	17	17	15	12
mindestens 95 %	8	8	8	8	6

Tab. 3.10: Textpaare mit bestimmten Matchanteilen in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (Basis: alle MEMs ab 18 Zeichen, überarbeitet entsprechend Tab. 3.9)

Matchanteil (1. Text)	alle Textpaare	mindestens 100 Matchzeichen	mindestens 500 Matchzeichen	mindestens 1.000 Matchzeichen	mindestens 10.000 Matchzeichen
mindestens 0 %	3.778	2.317	832	528	116
mindestens 5 %	317	294	251	229	102
mindestens 10 %	198	198	170	161	82
mindestens 15 %	142	142	126	118	73
mindestens 20 %	109	109	98	90	59
mindestens 25 %	78	78	72	64	46
mindestens 30 %	64	64	60	52	39
mindestens 35 %	56	56	54	46	36
mindestens 40 %	49	49	48	41	32
mindestens 45 %	39	39	39	33	29
mindestens 50 %	38	38	38	32	28
mindestens 55 %	34	34	34	29	26
mindestens 60 %	31	31	31	27	24
mindestens 65 %	30	30	30	26	23
mindestens 70 %	28	28	28	24	21
mindestens 75 %	25	25	25	21	18
mindestens 80 %	18	18	18	15	12
mindestens 85 %	13	13	13	11	8
mindestens 90 %	10	10	10	8	5
mindestens 95 %	3	3	3	3	1

Tab. 3.11: Textpaare mit bestimmten Matchanteilen in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (Basis: alle MEMs ab 48 Zeichen, überarbeitet entsprechend Tab. 3.9)

nicht einem der Einzeltexte zuordnen lässt, sondern für das Gesamtwerk gilt. Zum Beispiel ist darin öfters ein Druckprivileg enthalten, das häufig formelhaft ist und entsprechend viele Übereinstimmungen mit anderen Texten aufweist.

Dass sich bei höheren Prozentwerten für die Übereinstimmung die Zahlen der ermittelten Textpaare in allen Spalten stark annähern, lässt sich analog zum letzten

Matchanteil (1. Text)	alle Textpaare	mindestens 100 Matchzeichen	mindestens 500 Matchzeichen	mindestens 1.000 Matchzeichen	mindestens 10.000 Matchzeichen
mindestens 0 %	48.470	40.741	26.029	17.800	1.856
mindestens 5 %	2.860	2.843	2.590	2.443	900
mindestens 10 %	900	897	814	760	477
mindestens 15 %	482	479	454	411	313
mindestens 20 %	313	311	307	284	230
mindestens 25 %	196	195	195	187	147
mindestens 30 %	128	128	128	124	93
mindestens 35 %	83	83	83	82	57
mindestens 40 %	60	60	60	60	44
mindestens 45 %	45	45	45	45	32
mindestens 50 %	35	35	35	35	27
mindestens 55 %	28	28	28	28	22
mindestens 60 %	24	24	24	24	18
mindestens 70 %	20	20	20	20	14
mindestens 75 %	13	13	13	13	9
mindestens 80 %	10	10	10	10	6
mindestens 85 %	5	5	5	5	2
mindestens 90 %	2	2	2	2	0

Tab. 3.12: Textpaare mit bestimmten Matchanteilen bei Umwandlung in Kleinschreibung (Basis: alle MEMs ab 18 Zeichen, überarbeitet entsprechend Tab. 3.8)

Matchanteil (1. Text)	alle Textpaare	mindestens 100 Matchzeichen	mindestens 500 Matchzeichen	mindestens 1.000 Matchzeichen	mindestens 10.000 Matchzeichen
mindestens 0 %	21.777	8.148	2.367	1.415	224
mindestens 5 %	962	887	722	642	198
mindestens 10 %	488	483	367	339	162
mindestens 15 %	329	325	258	232	136
mindestens 20 %	245	245	202	177	115
mindestens 25 %	181	181	161	139	95
mindestens 30 %	145	145	133	111	77
mindestens 35 %	121	121	116	94	68
mindestens 40 %	96	96	94	73	57
mindestens 45 %	86	86	86	66	53
mindestens 50 %	72	72	72	58	47
mindestens 55 %	61	61	61	47	40
mindestens 60 %	48	48	48	39	33
mindestens 65 %	39	39	39	32	28
mindestens 70 %	36	36	36	30	26
mindestens 75 %	33	33	33	28	24
mindestens 80 %	31	31	31	27	24
mindestens 85 %	23	23	23	19	16
mindestens 90 %	21	21	21	18	15
mindestens 95 %	10	10	10	10	8

Tab. 3.13: Textpaare mit bestimmten Matchanteilen in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (Basis: alle MEMs ab 18 Zeichen ohne Überarbeitung)

Absatz erklären: Eine kleine Zahl von Matchzeichen kann nur dann einen größeren Teil eines Textes ausmachen, wenn dieser Text entsprechend kurz ist. Da in den Spalten jeweils die Textpaare mit einer Mindestübereinstimmungslänge verzeichnet werden, sind die Paare, die eine höhere Länge erreichen, mit enthalten; es ergibt sich also nur dann eine Differenz, wenn es Textpaare mit Matches gibt, die nicht das in der anderen Spalte angesetzte Längenkriterium erfüllen, aber den in der jeweiligen Zeile verzeichneten Anteil – und je höher der Prozentwert und je niedri-

ger die Mindestübereinstimmungslänge ist, desto weniger Texte können überhaupt aufgrund ihrer Länge einem solchen Paar zuzuordnen sein.

Schließlich sollen im Hinblick auf Tabelle 3.10 noch die Fälle kurz betrachtet werden, in denen sich insgesamt Übereinstimmungen von erheblicher Länge feststellen lassen, die aber nur einen sehr kleinen Prozentwert erreichen. In immerhin 25 Textpaaren im Korpus machen Matches von insgesamt mindestens 10.000 Zeichen weniger als 5 % der Länge des primär betrachteten Textes aus. In der Mehrzahl dieser Fälle liegt der Anteil im jeweils anderen Text des Paars allerdings über 10, bei elf Paaren über 20 und in einem Fall sogar über 50 %, was den Schluss nahelegt, dass die Ähnlichkeit doch größer sein, aber nur einen Teil des Textes betreffen dürfte, und die Vermutung bestärkt, dass eine Gesamtübereinstimmung in diesem Umfang auch bei einem niedrigen Prozentwert ein einigermaßen brauchbares Indiz für eine Textbeziehung ist. Übereinstimmungen dieser Länge werden allerdings in immerhin 32 Fällen auch bei der Ermittlung von Übereinstimmungen innerhalb eines Textes verzeichnet, und bei 30 davon liegt der Anteil der Matches bei über 5 %. Wenn Inhaltsverzeichnisse und Register sowie gänzlich in anderen Matches enthaltene Matches ausgeklammert werden, reduzieren sich diese Zahlen erheblich, es bleiben aber immer noch dreizehn Fälle mit einer Gesamtmatchlänge über 10.000 Zeichen und davon zehn mit einem Anteil von über 5 % übrig. Ein so hoher Anteil von wörtlichen Entsprechungen innerhalb eines Textes ist teilweise durch Struktur und Inhalt der betreffenden Texte zu erklären;[593] dass textinterne Matches relativ häufig einen größeren Umfang erreichen, legt aber doch die Einschätzung nahe, dass aus einer höheren absoluten Matchlänge nicht einfach unbesehen auf wörtliche Übernahmen geschlossen werden sollte. Vielmehr ist gerade bei der Auswertung von ziemlich kurzen Matches damit zu rechnen, dass sie auch in größerer Zahl durchaus auch durch zufällig gleiche Wortwahl und vor allem durch feste Wortverbindungen und damit sprachliche Ähnlichkeit zustande gekommen sein können.

[593] Bei den zuletzt genannten zehn Texten handelt es sich um Bemel,TraktTestam. 1587, Hugen, Rhetor. 1528, Klagspiegel(Brant) 1516, KölnErzstiftRef. 1538, Lettscher,Notariat 1576, Perneder, Proz. 1544, RKGO. 1548, RKGO. 1555, TirolLO. 1573 (1574) und Zwengel,Form. 1568. Dabei erreicht Hugen,Rhetor. 1528 mit 17 % Zeichen, die sich einer Übereinstimmung mit mindestens einer anderen Stelle im selben Text (ohne Inhaltsverzeichnis und Register) zuordnen lassen, einen Spitzenwert. Wenn nur Übereinstimmungen mit einer Länge von mindestens 48 Zeichen berücksichtigt werden, gibt es vier Texte, bei denen mindestens 10.000 Zeichen einem Match mit einer anderen Stelle im selben Text (ohne Inhaltsverzeichnis und Register) zugeordnet werden können, darunter einen (Bemel,TraktTestam. 1587) mit einem Anteil über 5 %, was tatsächlich auf gezielte Textwiederholung zurückzuführen ist (der Text eines Formulars wird zunächst abschnittsweise vorgestellt und am Ende noch einmal in Gänze wiedergegeben). Bei den übrigen eben genannten Texten lässt sich der hohe Grad an Wiederholungen relativ kurzer Wortfolgen im Text zum Teil darauf zurückführen, dass sie Formularsammlungen enthalten, daneben teilweise auch auf standardisierte Verweise auf andere Textstellen und überhaupt auf feste Wortverbindungen.

Der Vergleich mit den Tabellen 3.11 und 3.12 verdeutlicht noch einmal die unterschiedliche Werteentwicklung in Abhängigkeit von gewählter Codierung und Mindestlänge. Während die Zahl der Textpaare ohne weitere Auswahl nach absoluter oder relativer Länge in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen bei Matches mit einer Mindestlänge von 18 Zeichen mehr als 5,7mal so hoch ist wie die Zahl der Textpaare bei Matches von mindestens 48 Zeichen, nähern sich die beiden Zahlen schon bei einem sehr niedrigen absoluten oder relativen Schwellenwert deutlich an; bei einem Matchumfang von mindestens 50 % liegt das Verhältnis fast durchgängig unter 2 und steigt nur bei den höchsten Schwellenwerten teilweise wieder über diesen Wert.

Bei einem Vergleich auf der Basis von Texten, die nur durch Umwandlung in Kleinschreibung vereinheitlicht wurden, und bei Zugrundelegung einer Mindestmatchlänge von 18 Zeichen ist hingegen die Zahl der Textpaare ohne Zugrundelegung von Schwellenwerten viel höher, sie nimmt aber bei einer Auswahl anhand eines absoluten und/oder relativen Schwellenwerts deutlich stärker ab und unterschreitet die für die codierte Fassung bei einer Mindestmatchlänge von 48 Zeichen ermittelte Zahl ab Ansetzung einer Übereinstimmungsquote von 50 % beziehungsweise – bei Ansetzung einer Gesamtlänge von mindestens 1.000 Matchzeichen – 55 %.[594]

Bei Ansetzung eines relativen Schwellenwerts von unter 30 % ergibt der Vergleich der Tabellen 3.10 und 3.12 für die Mindestmatchlänge 18, dass die Zahlen für die Fassung in Kleinbuchstaben höher – und teilweise viel höher – liegen als für die Codierung *ohne A/I/U/B/F* ohne Leerzeichen. Insbesondere gilt das bei Verzicht auf einen relativen Schwellenwert. Dass zum Beispiel nach der obersten Datenzeile von Tabelle 3.12 etwa die Hälfte der etwa 52.000 Paare, die sich überhaupt aus den Texten des Korpus bilden lassen, Übereinstimmungen von insgesamt mindestens 500 Matchzeichen aufweist, verdeutlicht wohl noch einmal, dass die hier zugrunde gelegten Matches in einem Großteil der Fälle im Sinne der hieru untersuchten Fragestellung nicht aussagekräftig sind.

Der Vergleich von Tabelle 3.13 mit Tabelle 3.10 schließlich soll noch einen Eindruck davon verschaffen, wie sich die oben in Unterkapitel 3.3.1 beschriebene Überarbeitung der Matchdaten auf der Basis der Wortgrenzen auf die Ermittlung von Textpaaren nach quantifizierenden Kriterien auswirkt. Insgesamt kann dabei festgestellt werden, dass die Zahlen beider Tabellen recht nahe beieinander liegen.[595]

[594] Dass sich in den verschiedenen Spalten der Tabellen eine etwas unterschiedliche Werteentwicklung zeigt, lässt sich dadurch erklären, dass die Codierung *ohne A/I/U/B/F* ohne Leerzeichen zu einer starken Komprimierung führt. Die in den verschiedenen Spalten angesetzte Mindestzahl an Matchzeichen entspricht also in den Tabellen 3.10 und 3.11 einer größeren Textmenge im Original als in Tabelle 3.12 und stellt damit ein strengeres Auswahlkriterium dar.

[595] Bei den höchsten relativen Schwellenwerten geht das Verhältnis der einander entsprechenden Werte in den beiden Tabellen zwar etwas stärker auseinander, aber es handelt sich dabei um sehr kleine Fallzahlen, und da es hier nicht primär um die Ermittlung von Quasi-Dubletten geht, sondern insbesondere auch um das Aufspüren von weniger umfassenden Entsprechungen,

Wenn es darum geht, über einen Schwellenwert eine Auswahl von Textpaaren zu treffen, die in ihren Beziehungen näher untersucht oder deren Übereinstimmungen visualisiert werden sollen, beeinträchtigt es das Gesamtbild also wohl nur wenig, wenn die MEMs ohne irgendeine Filterung zugrunde gelegt werden. Für einen präziseren Vergleich, zum Beispiel wie er unten in Unterkapitel 3.4.5 beschrieben wird, ist eine Überarbeitung der Matchdaten natürlich trotzdem sinnvoll, kann dann aber auf die ausgewählten Paare eingeschränkt werden.

3.4.2 Graphen auf der Basis von Textpaarähnlichkeiten

Soweit es um die Ermittlung engerer Verwandtschaftsbeziehungen eines bestimmten Textes geht, lässt sich schon durch Berechnung der summierten Matchlängen und der entsprechenden Textanteile sowie gegebenenfalls eine Sortierung nach den absoluten oder relativen Übereinstimmungswerten eine gute Vorauswahl erreichen, insbesondere wenn durch entsprechende Schwellenwerte isoliert stehende kurze oder zum Beispiel nach einem der in Unterkapitel 3.3.2 beschriebenen Verfahren sehr niedrig bewertete Matches von vorneherein ausgeklammert werden. Um jedoch einen Überblick über solche Beziehungen in einem ganzen Korpus zu gewinnen, ist eine Auflistung, also eine sequentielle Darstellung, wenig geeignet. Hier bietet sich eine Visualisierung an, wie sie durch Graphen erfolgen kann. Auch hier zeigt sich aber, dass sich jedenfalls bei einem Korpus wie dem hier untersuchten nur bei einer restriktiven Auswahl der zu präsentierenden Ähnlichkeiten ein einigermaßen übersichtliches Bild gewinnen lässt. Deshalb ist Abbildung 3.15 auf S. 201 auf die Auswertung der (entsprechend Tabelle 3.9 gefilterten) Matches zwischen Gerichtsordnungen[596] beschränkt und berücksichtigt zudem nur Textpaare, bei denen die Zahl der einem Match zugeordneten Zeichen mindestens 1000 beträgt[597] und die Matches in einem der Texte mindestens 10 % des Gesamtumfangs ausmachen.

Die ermittelten Übereinstimmungsanteile sind jeweils neben den Verbindungslinien vermerkt und können sich zwischen den beiden Texten eines Paars stark unterscheiden. Der jeweils höhere Prozentwert dient als Grundlage für die Berechnung der Dicke der Verbindungslinie, um stärkere Ähnlichkeiten zwischen Texten optisch hervorzuheben. Dass die angegebenen Werte bei Texten, die einander fast völlig entsprechen, im untersuchten Korpus auch bei der hier zugrunde gelegten, im Hinblick auf Schreibungsvarianten besonders stark nivellierenden Codierung

sind die höchsten Schwellenwerte hier nur der Vollständigkeit halber verzeichnet und dürften praktisch kaum von Interesse sein.

[596] Die Auswahl wurde dabei rein formal nach dem Siglenbestandteil „GO." getroffen – Texte, die ebenfalls Gerichtsordnungen sind oder enthalten, aber aufgrund des Titels oder der in der Wissenschaft eingeführten Bezeichnung eine andere Sigle haben, sind nicht berücksichtigt.

[597] Da sich aufgrund von Mehrfachzuordnungen einer Textstelle für beide Texte unterschiedliche Zahlen ergeben können, wird hier die kleinere zugrunde gelegt.

ohne A/I/U/B/F ohne Leerzeichen und mit einer niedrigen Mindestmatchlänge von 18 Zeichen maximal 96 beziehungsweise 97 % erreichen (bei der *Bambergensis* und der *Brandenburgensis*[598]), zeigt allerdings, dass diese Zahlen nicht einfach als Bezifferung des tatsächlichen Ausmaßes an wörtlicher Übereinstimmung verstanden werden sollten. Es gibt eine sicher nicht ganz geringe Zahl an Schreibungsvarianten, die durch die hier zugrunde gelegten Ersetzungsregeln nicht vereinheitlicht werden.[599]

Auch hier soll zum Vergleich betrachtet werden, zu welchem Bild eine entsprechende Auswertung auf der Basis der Fassung in Kleinbuchstaben führt. Dabei wird ebenfalls die Mindestlänge 18 zugrunde gelegt und damit die geringste Mindestlänge, die in dieser Untersuchung für diese Fassung überhaupt in Betracht gezogen wird. Da dadurch auch viel kürzere Übereinstimmungen in den Originaltexten berücksichtigt werden als bei derselben Länge auf der Basis der Codierung *ohne A/I/U/B/F* ohne Leerzeichen, gibt es – wie oben schon in den Tabellen 3.10 und 3.12 gezeigt – im gesamten Korpus erheblich mehr Textpaare, die den hier angesetzten quantitativen Auswahlkriterien entsprechen, und dementsprechend zeigt Abbildung 3.16 mehr Linien zwischen Textpaaren.

Oben wurde aber bereits festgestellt, dass sich für Textpaare mit sehr umfangreichen Entsprechungen insgesamt ein ganz anderes Bild ergibt, und hier lässt sich im Vergleich mit Abbildung 3.15 für einzelne Textpaare entnehmen, wie groß die Unterschiede im Hinblick auf den relativen Umfang sind. Diese fallen recht unterschiedlich aus. Zum einen erklärt sich das durch das Ausmaß der Schreibungsabweichungen, das vom jeweiligen Textpaar abhängt. Zum anderen gibt es natürlich auch in Textpaaren, für die auf der Basis der Codierung *ohne A/I/U/B/F* ohne Leerzeichen ein hoher Übereinstimmungsgrad ermittelt wird, kürzere einander entsprechende Stücke, die die für diese Codierung angesetzte Mindestlänge nicht erreichen, aber aufgrund exakter Übereinstimmung bei einem Vergleich auf der Basis einer Umwandlung in Kleinbuchstaben gefunden werden können, wodurch sich der Abstand zwischen den angegebenen Prozentwerten reduziert. Je umfassender aber die längeren Matches sind, desto kleiner sind die Bereiche, in denen durch die Einbeziehung kurzer Matches etwas hinzukommen kann.

Dementsprechend ist der Unterschied zwischen den angegebenen Prozentwerten zum Beispiel für die Beziehung zwischen der Bamberger Halsgerichtsordnung von 1507 und der revidierten Fassung von 1580 besonders groß, da sich die Schreibpräferenzen im Lauf der Zeit entwickelten und hier der zeitliche Abstand erheblich ist und da die längeren Matches den Großteil beider Texte abdecken.

[598] Vgl. oben S. 37.

[599] Eine gewisse Verbesserung lässt sich immerhin mit dem in Unterkapitel 3.4.5 vorgestellten Feinvergleichsverfahren erreichen, das unter anderem auch eine Klassifikation von einander nicht genau entsprechenden, aber überwiegend übereinstimmenden Wortformen im Umfeld von Matches als *ähnlich* ermöglicht. Tabelle 3.16a auf S. 245zeigt die dabei ermittelten Daten für den Vergleich von *Bambergensis* und *Brandenburgensis*.

Sogar zwischen den beiden Einzeltexten eines Textpaars kann es Unterschiede geben. Die Prozentangaben berücksichtigen alle Zeichen, die sich – nach den drei in den Tabellen 3.8 und 3.9 dokumentierten Fiterschritten – mindestens einem Match mit dem jeweils anderen zuordnen lassen. Die Häufigkeit insbesondere einer kurzen Zeichenfolge kann aber zwischen den beiden jeweils betrachteten Texten differieren, und dementsprechend ist die Zahl der den Matches zuzuordnenden Zeichen in vielen Fällen in den beiden Texten nicht gleich. Das erklärt, warum die beiden jeweils für ein Textpaar verzeichneten Prozentwerte in den beiden Abbildungen nicht unbedingt im selben Verhältnis zueinander stehen.

Es zeigt sich also, dass bei Verzicht auf eine Schreibungsvereinheitlichung auch bei der Ansetzung einer sehr kurzen Mindestlänge der *Recall* bei Textpaaren, die recht umfassende Übereinstimmungen aufweisen, erheblich hinter dem zurückbleibt, der sich über einen Vergleich auf der Basis einer die Varianz stark reduzierenden Codierung erzielen lässt. Dass die zusätzlichen Funde, die sich durch die Einbeziehung sehr kurzer Übereinstimmungen ergeben, gegenüber den Matches in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen (oder auch einer anderen Codierung) teilweise Hinweise auf textuelle Beziehungen geben könnten, ist nicht auszuschließen, die bisher betrachteten Daten lassen aber für das hier untersuchte Korpus deutlich erkennen, dass es sich dabei jedenfalls zum größten Teil um häufig vorkommende und in dieser Hinsicht kaum aussagekräftige Wortfolgen handelt.

Daneben soll – wie schon oben in Tabelle 3.11 – auch noch betrachtet werden, welche Auswirkungen es hat, wenn die für die Codierung *ohne A/I/U/B/F* ohne Leerzeichen angesetzte Mindestlänge auf 48 erhöht wird. Gegenüber der Mindestlänge 18 reduziert sich damit der Umfang der entsprechend Tabelle 3.9 gefilterten Matchdaten insgesamt auf etwa 35 %. In Textpaaren mit starker Übereinstimmung weisen die Matches aber vielfach eine recht große Länge auf, und dementsprechend erreichen die in Abbildung 3.17 verzeichneten Prozentwerte teilweise eine nur wenig niedrigere Höhe als in Abbildung 3.15 und eine deutlich höhere als in Abbildung 3.12.[600]

Die Anordnung der Knoten in diesen drei Graphen – wie auch weitestgehend in den folgenden[601] – ist nach bestimmten Kriterien automatisch generiert und stimmt deshalb in den verschiedenen Abbildungen nur teilweise überein. Dabei wird zwar das Erscheinungsjahr der Texte, wie es aus den Siglen zu entnehmen ist,

[600] Das gilt nicht in jedem Fall. Wenn ein Textpaar nur eine relativ geringe Zahl an Schreibungsunterschieden aufweist, kann es stärker ins Gewicht fallen, wie groß die angesetzte Mindestmatchlänge ist. Ein Beispiel, bei dem der Vergleich auf Basis einer Umwandlung in Kleinbuchstaben zu einer etwas höheren Zuordnungsquote führt als der Vergleich auf der Basis der Codierung *ohne A/I/U/B/F* ohne Leerzeichen mit einer Mindestmatchlänge von 48 Zeichen, ist das Textpaar WittenbergHofGO. 1550 – JenaHofGO. 1566.

[601] An wenigen Stellen wurden spezielle Festlegungen vorgenommen, um die Verteilung der Knoten auf der Seite und damit die Lesbarkeit zu verbessern.

als Basis für die Anordnung der einzelnen Textpaare verwendet, so dass die älteren Texte (bei Drehung der Seite entsprechend dem Schriftbild) jeweils oberhalb der zugeordneten jüngeren stehen, damit soll aber keineswegs der Eindruck erweckt werden, dass es zwischen den durch Linien verbundenen Texten tatsächlich jeweils ein Abhängigkeitsverhältnis gebe oder dass gar ein Stemma dargestellt werden solle. Vielmehr wird einfach dokumentiert, wo sich im jeweils angesetzten Ausmaß Entsprechungen ermitteln lassen, so dass zum Beispiel auch direkte Verbindungslinien vom ersten zum letzten Glied einer Kette von Texten verzeichnet sein können. Ganz abgesehen von dem Problem, automatisiert eine sachlich angemessene Auswahl zu treffen – es ist ja durchaus damit zu rechnen, dass ein jüngerer Text auf einen älteren zurückgreift, auch wenn andere das schon vor ihm getan haben –, hat dies den Vorteil, dass Textgruppen sowohl durch die Verbindungslinien als auch durch die davon entsprechend dem zugrunde liegenden Algorithmus beeinflusste Anordnung leichter erkennbar werden.[602]

Während sich bei den in diesen Graphen zugrunde gelegten Schwellenwerten unter den Gerichtsordnungen verschiedene Textgruppen gut voneinander abgrenzen lassen, ist der Befund wesentlich weniger übersichtlich, wenn auch niedrigere Übereinstimmungsgrade verzeichnet werden. Aber selbst wenn als Schwellenwert nur eine Übereinstimmung von 500 Zeichen oder von 3 % und zugleich mindestens 100 Zeichen angesetzt und keine Gewichtung nach dem Übereinstimmungsgrad vorgenommen wird, zeichnen sich diese Gruppen teilweise schon recht gut ab, wie aus den Abbildungen 3.18 und 3.19 auf S. 204 und 205 zu entnehmen ist: Hier fällt neben isoliert stehenden Textpaaren insbesondere die Gruppe der peinlichen Gerichtsordnungen ins Auge, die Verwandtschaft der sächsischen Gerichtsordnungen lässt sich erkennen, und in Abbildung 3.19 fällt eine Ballung von Linien auf, die von der Braunschweig-Wolfenbütteler Hofgerichtsordnung von 1556 ausgehen und Texte, die mit ihr direkt oder indirekt durch Übernahmen verbunden sind, als Gruppe zusammenfassen.

Eine einfache Ermittlung von Textgruppen kann auch auf der Basis erfolgen, dass nicht alle Textpaare mit einer ein bestimmtes Mindestmaß überschreitenden Übereinstimmung verzeichnet werden, sondern vielmehr für die diesem Kriterium genügenden Textpaare untersucht wird, welche weiteren Texte es gibt, mit denen die beiden Texte des Paars ebenfalls durch Matches in einem Umfang oberhalb des Schwellenwerts verbunden sind. Wenn dabei als Auswahlkriterium zugrunde gelegt wird, dass die Zahl der Texte, die diesem Kriterium entsprechen, die Zahl der

[602] Verwendet wurde das Programm *dot* aus der *Graphviz*-Programmfamilie (vgl. *http://www.graphviz.org/*). Die Lesbarkeit und Übersichtlichkeit wurden durch verschiedene, zum größten Teil auf der Basis einer Graphanalyse automatisch generierte Zusätze in den Graphbeschreibungsdateien verbessert, die sich auf die Anordnung auswirken. Letztlich beruht diese Anordnung aber auf dem in *dot* eingebauten Verfahren.

Texte, mit denen nur einer der Texte Ähnlichkeiten in diesem Ausmaß aufweist, in einem bestimmten Maß übersteigt, entfallen zwar manche Textpaare, bei denen man wohl mit guter Berechtigung von einer – vielleicht entfernten oder partiellen – Verwandtschaft sprechen kann, bei denen es aber zugleich in einem Text oder auch in beiden ein erhebliches Maß von Ähnlichkeiten mit anderen Texten gibt, die nicht derselben Gruppe zuzuordnen sind.[603] Da dem Kriterium aber gerade solche Textpaare genügen, die aufgrund der ermittelten sonstigen Ähnlichkeiten die Existenz einer Textgruppe vermuten lassen, zeichnen sich solche Textgruppen hierbei schon bei niedrigeren Schwellenwerten ab als bei der Berücksichtigung aller Textpaare, deren Übereinstimmungen den jeweils angesetzten Wert überschreiten.

Die Abbildungen 3.20 und 3.21 auf S. 206 und 207 sollen das verdeutlichen. Darin sind alle Textpaare berücksichtigt, bei denen die Zahl anderer Texte, mit denen beide Texte Ähnlichkeiten im angesetzten Ausmaß aufweisen, mehr als doppelt so groß ist wie die Zahl der Texte, mit denen einer der beiden entsprechende Ähnlichkeiten aufweist, der andere aber nicht. Das Paar wird dabei auch dann ausgewählt, wenn der zweite Text mit einer größeren Zahl von weiteren Texten Übereinstimmungen in diesem Umfang hat, da auf diese Weise auch die Gruppenzusammenhänge der in Textsammlungen enthaltenen Werke gegebenenfalls mit einbezogen werden können. Die angesetzten Schwellenwerte entsprechen denen in den Abbildungen 3.18 und 3.19 auf S. 204 und 205, offenkundig sind aber Textgruppen wesentlich besser zu erkennen und zeichnen sich weitgehend entsprechend zu denen ab, wie sie aus Abbildung 3.15 zu entnehmen sind. Zugleich gibt es aber auch eine Reihe von Abweichungen, die wiederum hoffentlich verdeutlichen, dass die beschriebenen Auswahlkriterien natürlich die aufgrund von Einzelmatches in Betracht zu ziehenden Textpaare nicht in einer Weise filtern können, die der Komplexität textueller Beziehungen wirklich angemessen wäre, dass sich aber mit einer Kombination verschiedener Kriterien eine signifikante Verbesserung des Auswahlergebnisses erzielen lässt.

Die bisher betrachteten Graphen beschränken sich, wie gesagt, auf eine Darstellung der ermittelten Ähnlichkeiten zwischen Gerichtsordnungen, da die Darstellung sonst zu unübersichtlich würde. Um aber auch vom gesamten Korpus einen Eindruck zu vermitteln, zeigt Abbildung 3.22 auf S. 208 den Graphen, der sich aus der Analyse der Beziehungen im gesamten Textbestand ergibt, wenn der höchste hier angesetzte relative Schwellenwert von 10 % zugrunde gelegt wird. Während für die Gerichtsordnungen mit diesen Werten eine recht klare Gruppierung erreicht werden kann, ist das Bild für das ganze Korpus deutlich weniger übersichtlich. Auch hier lässt sich allerdings über das zusätzliche Kriterium der überwiegend

[603] Das kann – neben einer eklektischen Textformulierung auf der Basis mehrerer Vorlagen – insbesondere auch dadurch zu erklären sein, dass Texte auch Zusammenstellungen von Einzeltexten sein können, die jeweils eigene Verwandtschaftsverhältnisse haben.

gemeinsamen Entsprechungstexte (entsprechend den Abbildungen 3.20 und 3.21 auf S. 206 und 207) eine erhebliche Komplexitätsreduktion erreichen, die auch bei einer Absenkung des Schwellenwerts auf 5 % noch eine recht gut strukturierte Darstellung ermöglicht, wie aus Abbildung 3.23 auf S. 209 entnommen werden kann.

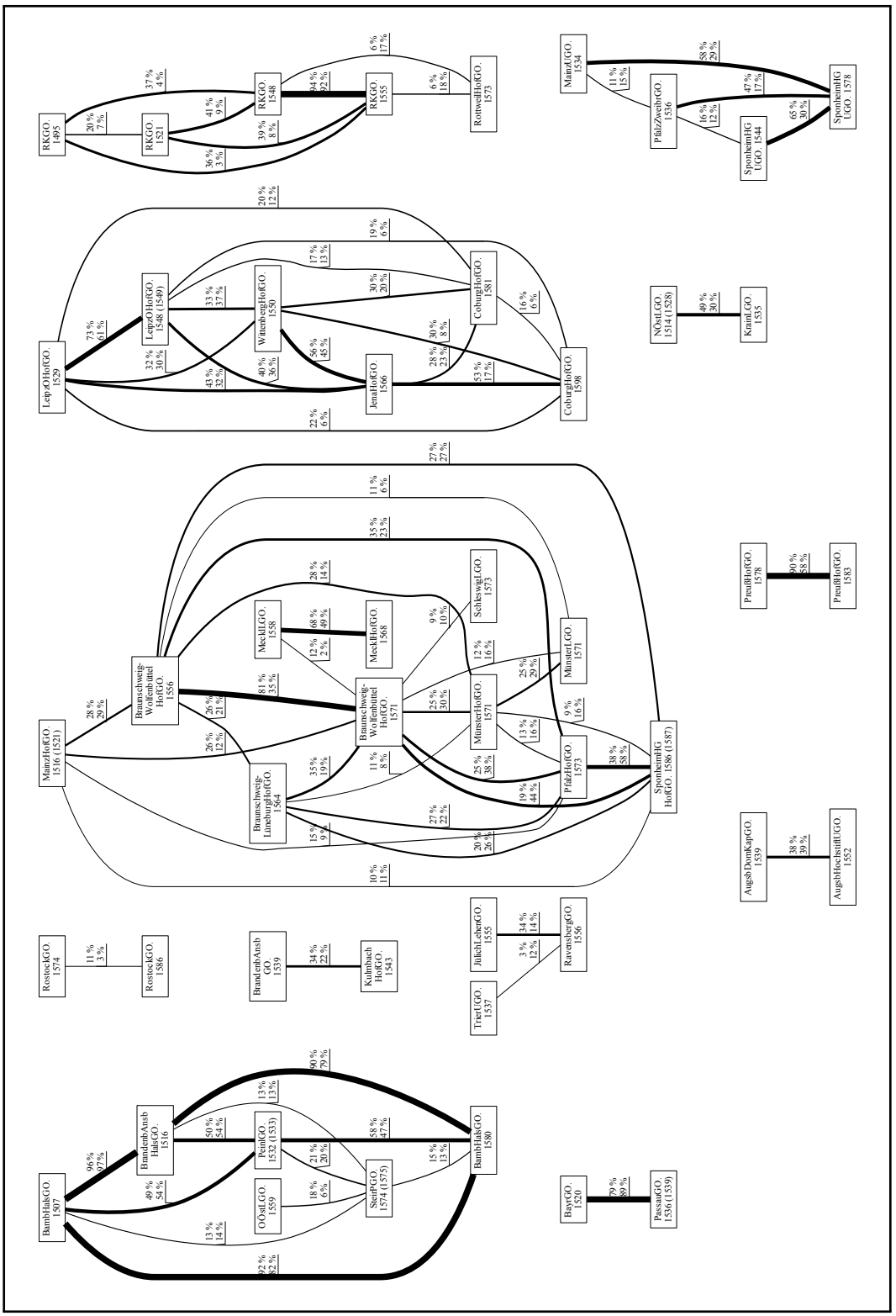


Abb. 3.15: Textpaare (Gerichtsordnungen) mit gefilterten Matches ab 1000 Zeichen und 10 % in mindestens einem der Texte (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, MEMs ab 18 Zeichen, vgl. S. 195 und zur Gestaltung S. 197)

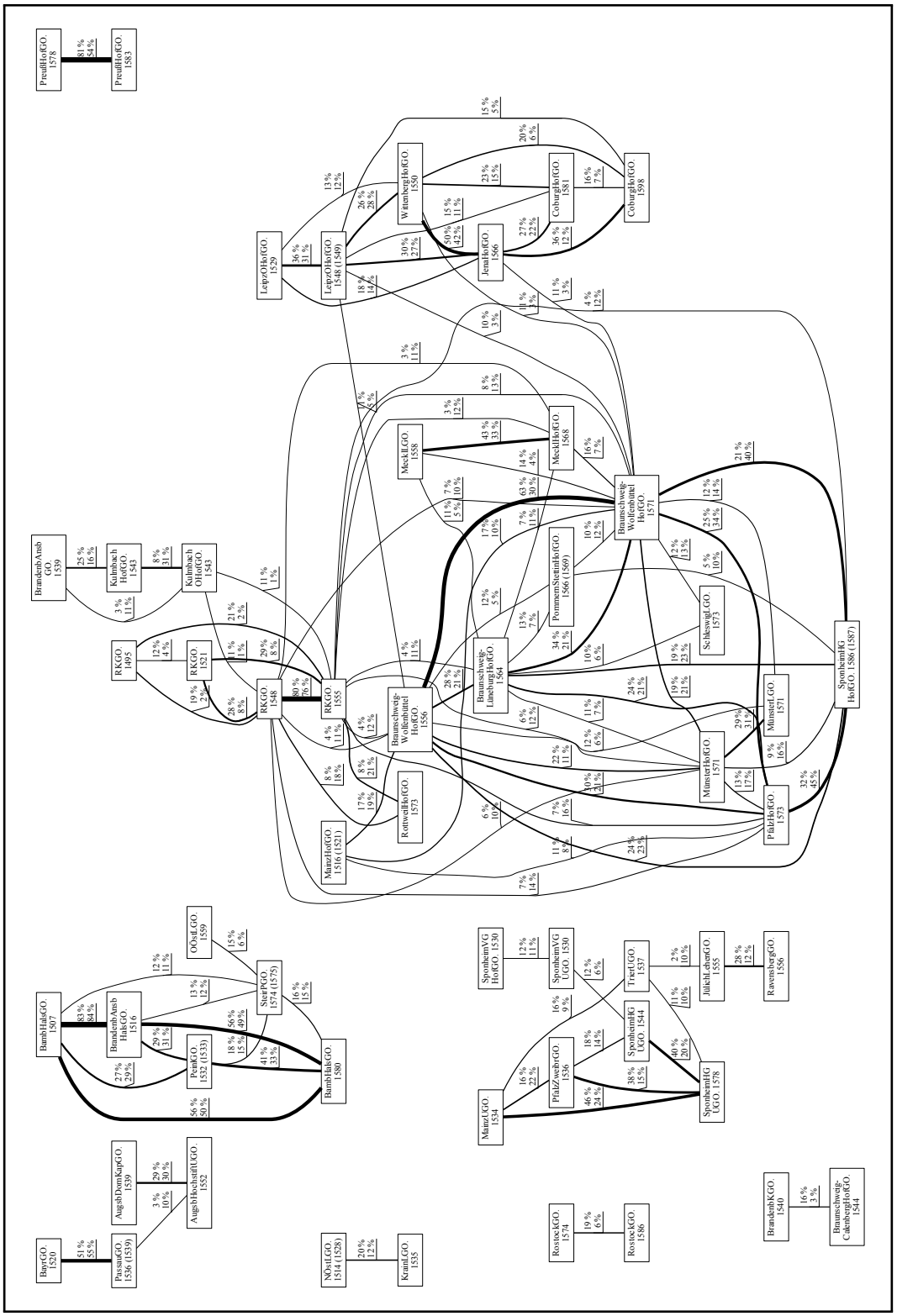


Abb. 3.16: Textpaare (Gerichtsordnungen) mit gefilterten Matches ab 1000 Zeichen und 10 % in mindestens einem der Texte (Textfassung in Kleinbuchstaben mit Leerzeichen, MEMs ab 18 Zeichen, vgl. S. 196 und zur Gestaltung S. 197)

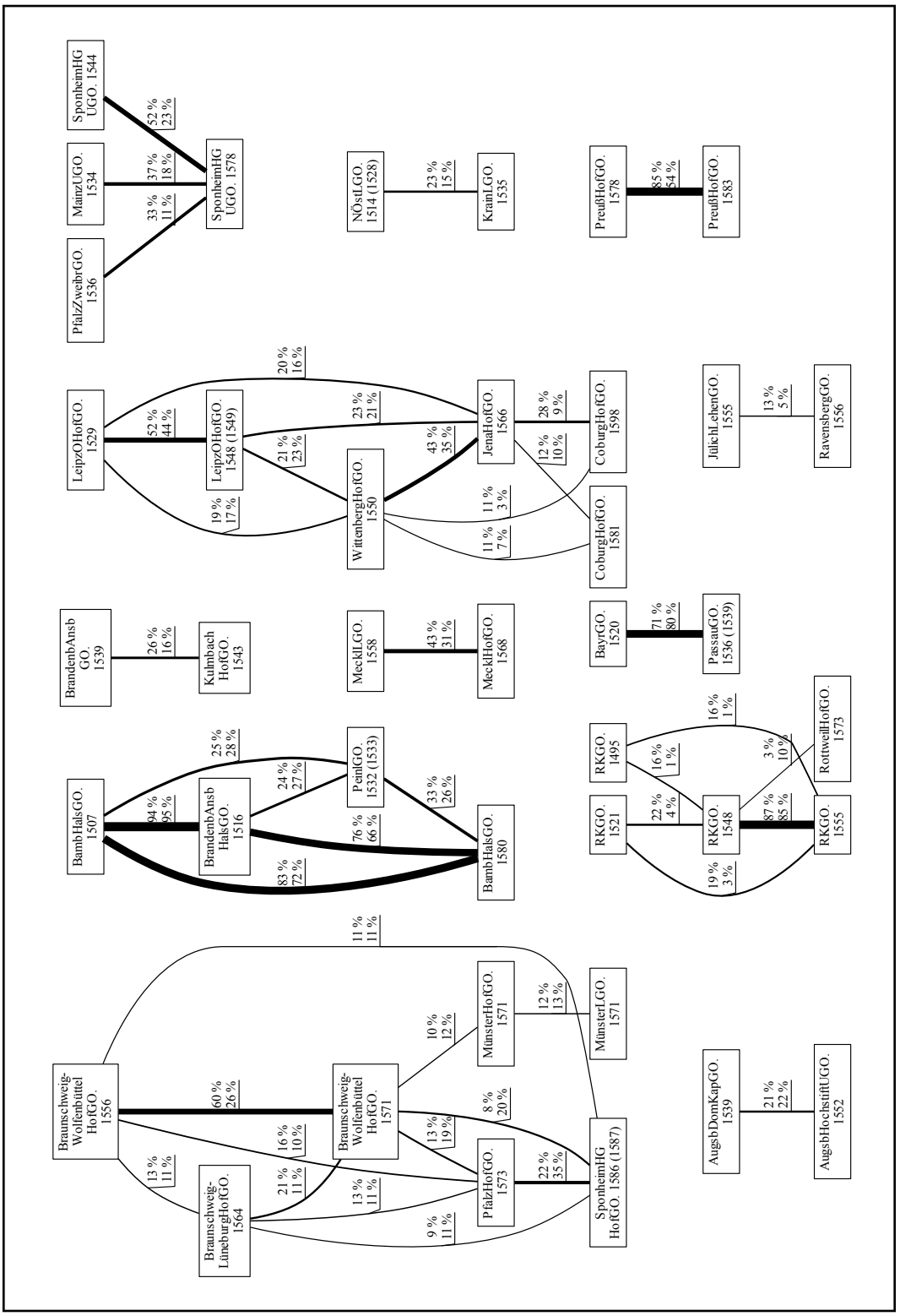


Abb. 3.17: Textpaare (Gerichtsordnungen) mit gefilterten Matches ab 1000 Zeichen und 10 % in mindestens einem der Texte (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, MEMs ab 48 Zeichen, vgl. S. 197 und zur Gestaltung S. 197)

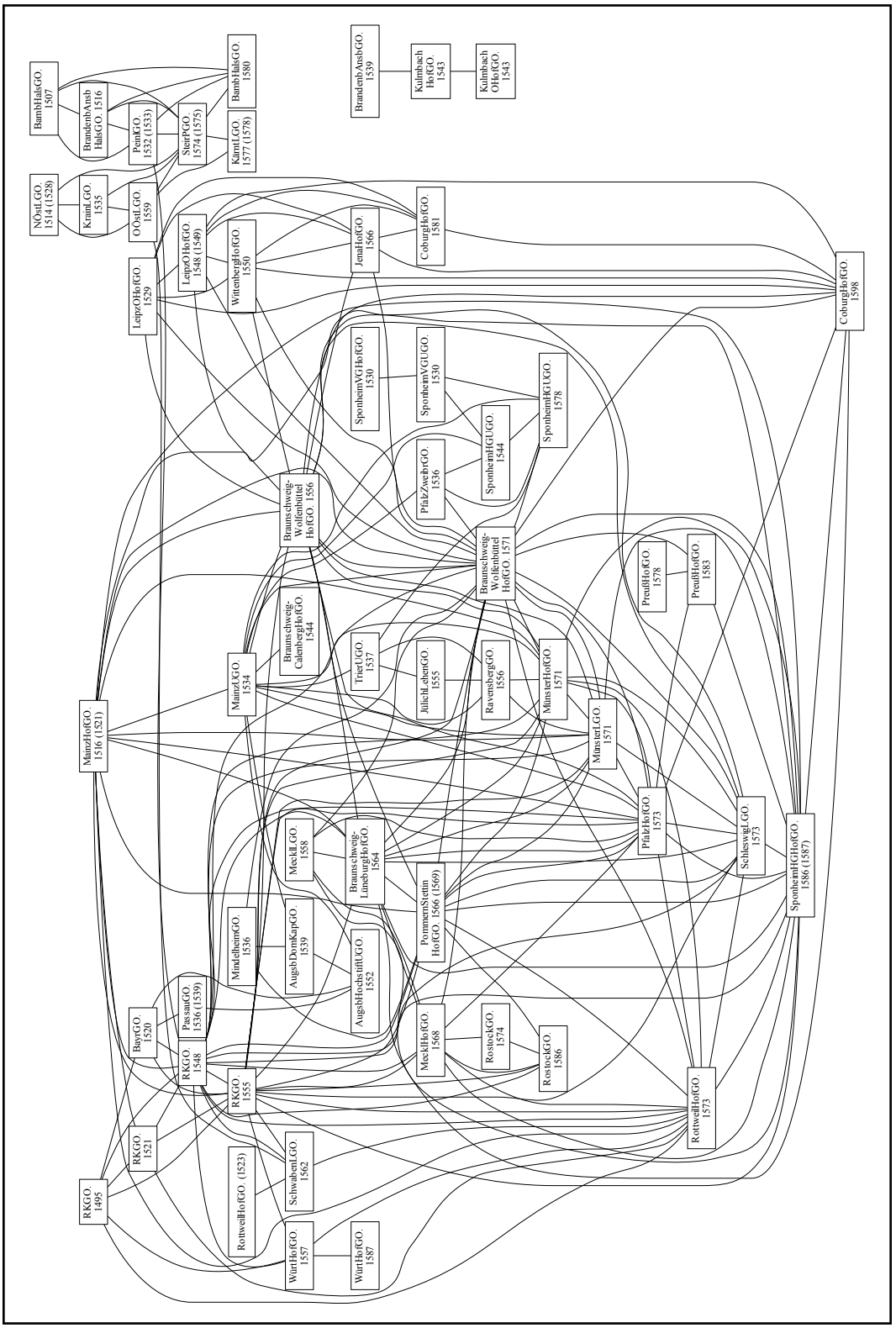


Abb. 3.18: Textpaare (Gerichtsordnungen) mit gefilterten Matches ab 500 Zeichen (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, Matches ab 18 Zeichen, vgl. S. 198 und zur Gestaltung S. 197)

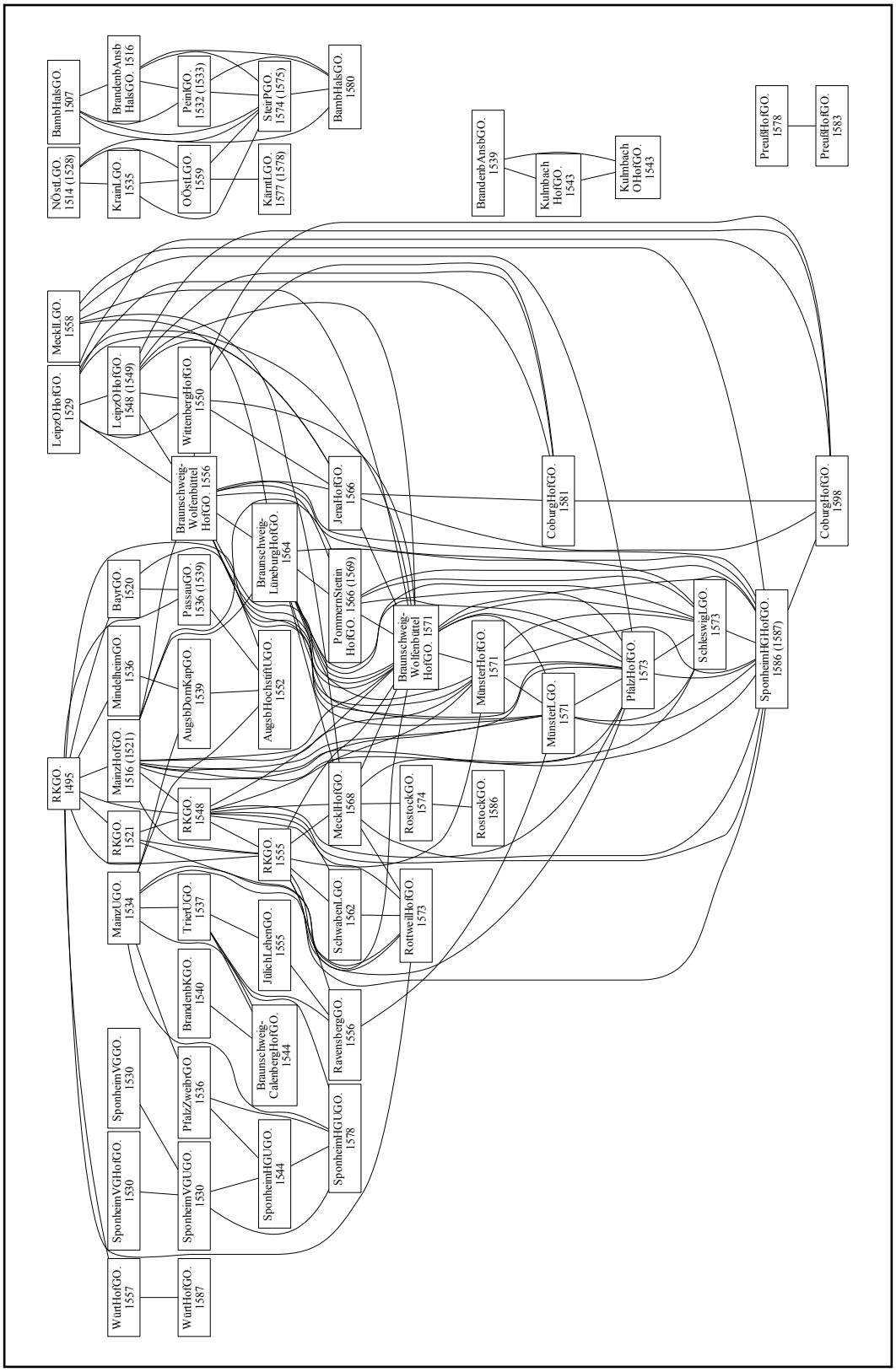


Abb. 3.19: Textpaare (Gerichtsordnungen) mit gefilterten Matches ab 3 % (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, Matches ab 18 Zeichen, vgl. S. 198 und zur Gestaltung S. 197)

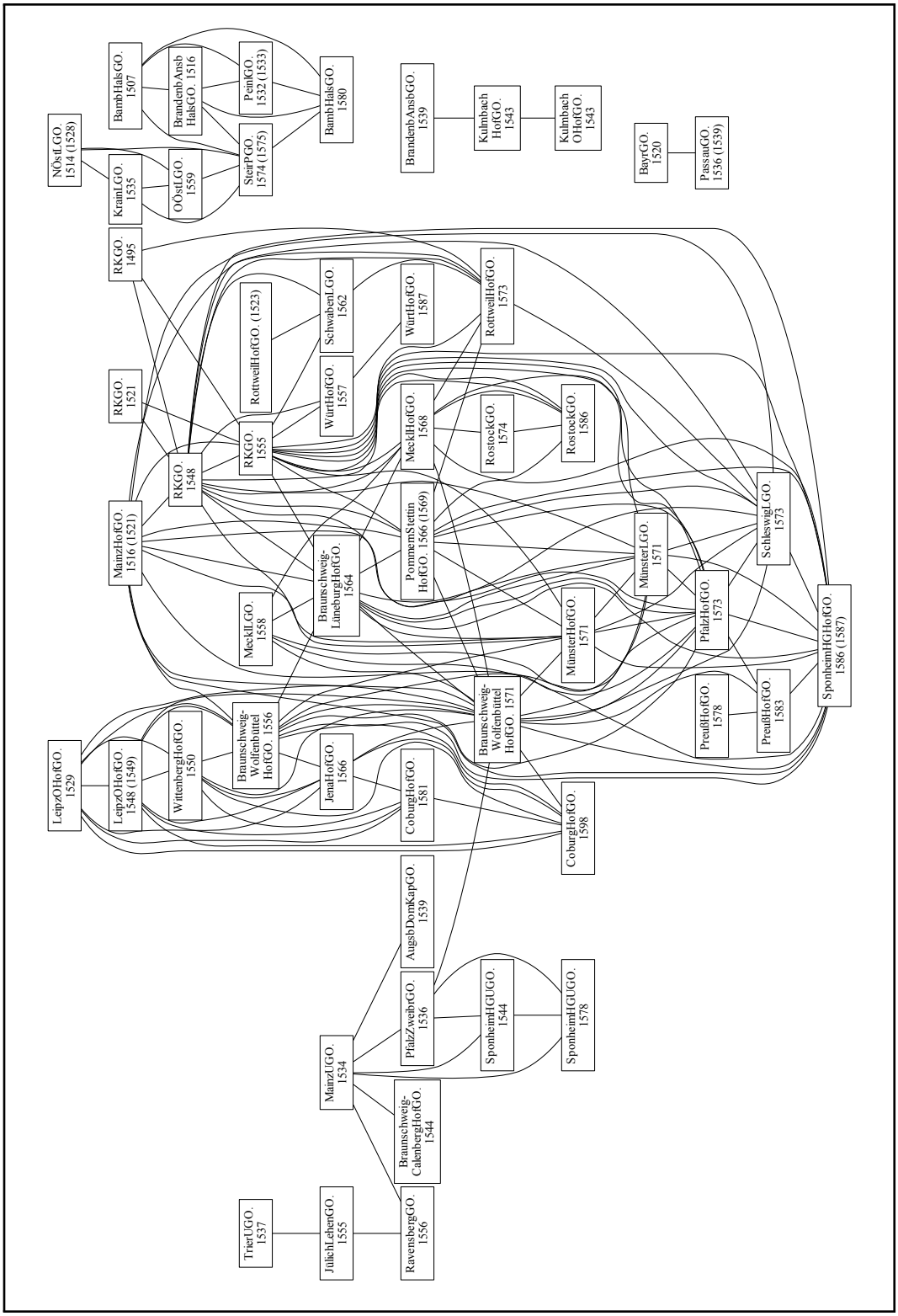


Abb. 3.20: Textpaare (Gerichtsordnungen) mit gefilterten Matches ab 500 Zeichen, die dem auf S. 199 beschriebenen Textgruppenkriterium entsprechen (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, Matches ab 18 Zeichen, vgl. zur Gestaltung S. 197)

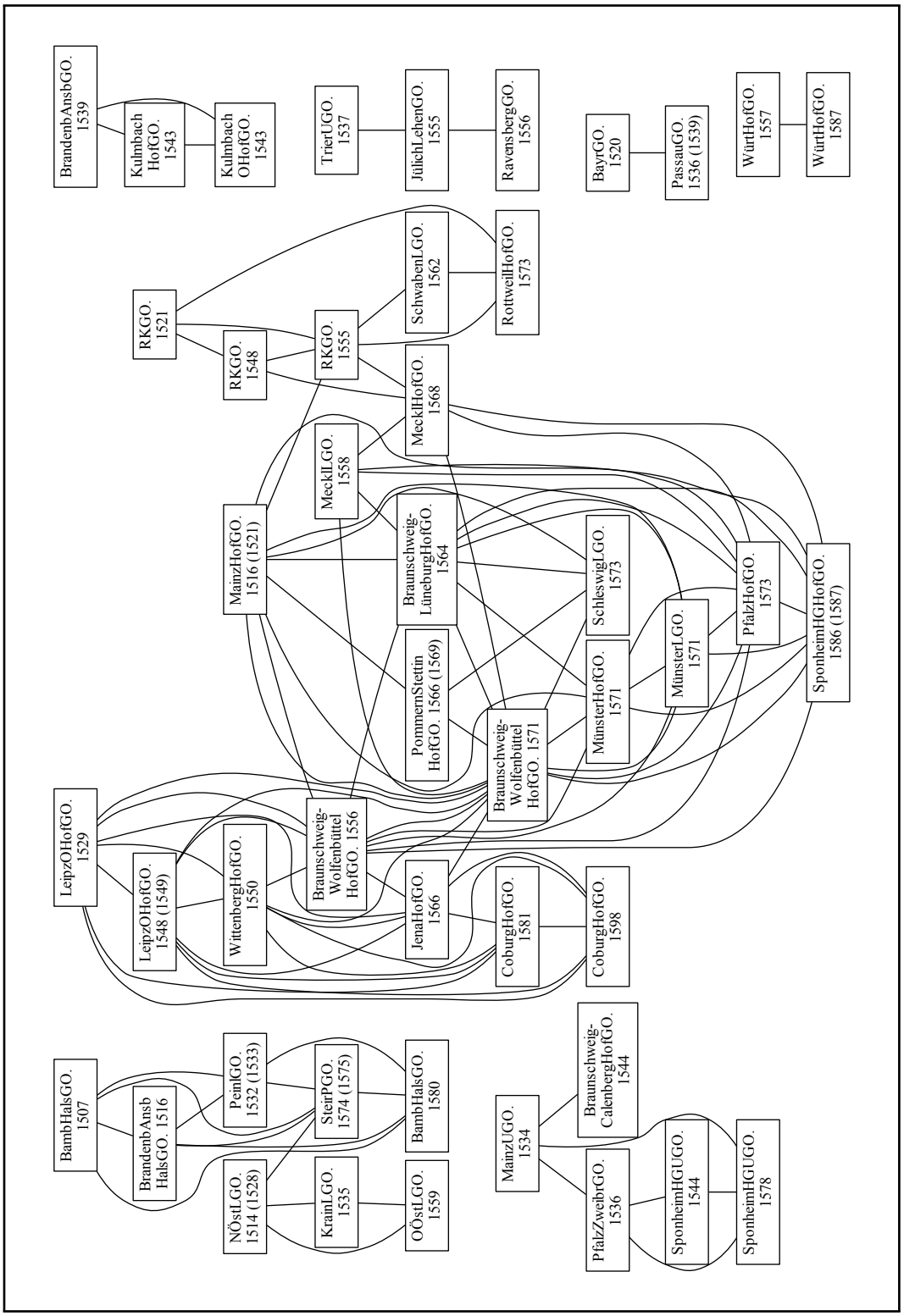


Abb. 3.21: Textpaare (Gerichtsordnungen) mit gefilterten Matches ab 3 %, die dem auf S. 199 beschriebenen Textgruppenkriterium entsprechen (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, Matches ab 18 Zeichen, vgl. zur Gestaltung S. 197)

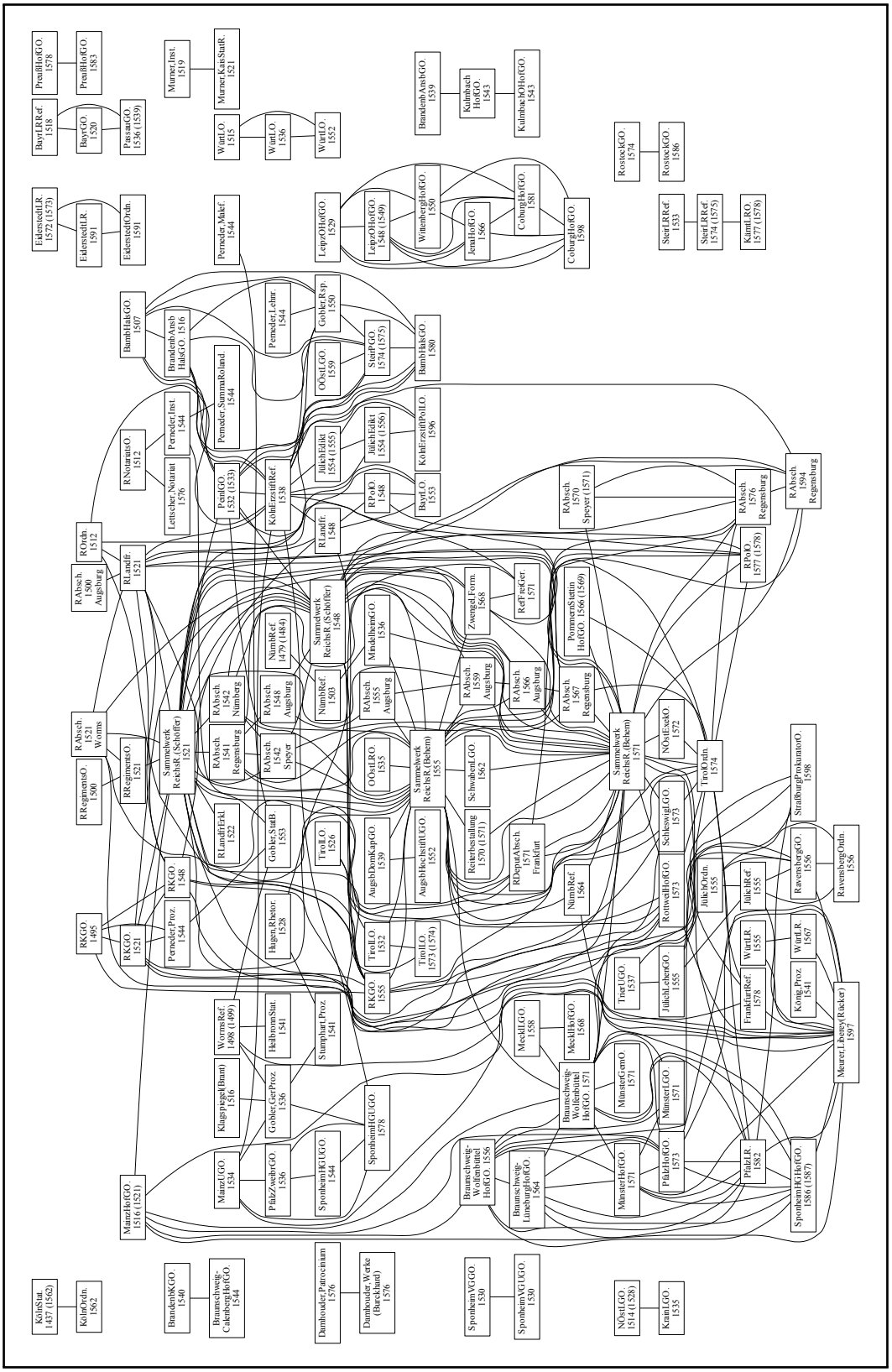


Abb. 3.22: Textpaare (alle Texte) mit gefilterten Matches ab 10 % (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, Matches ab 18 Zeichen, vgl. S. 199 und zur Gestaltung S. 197)

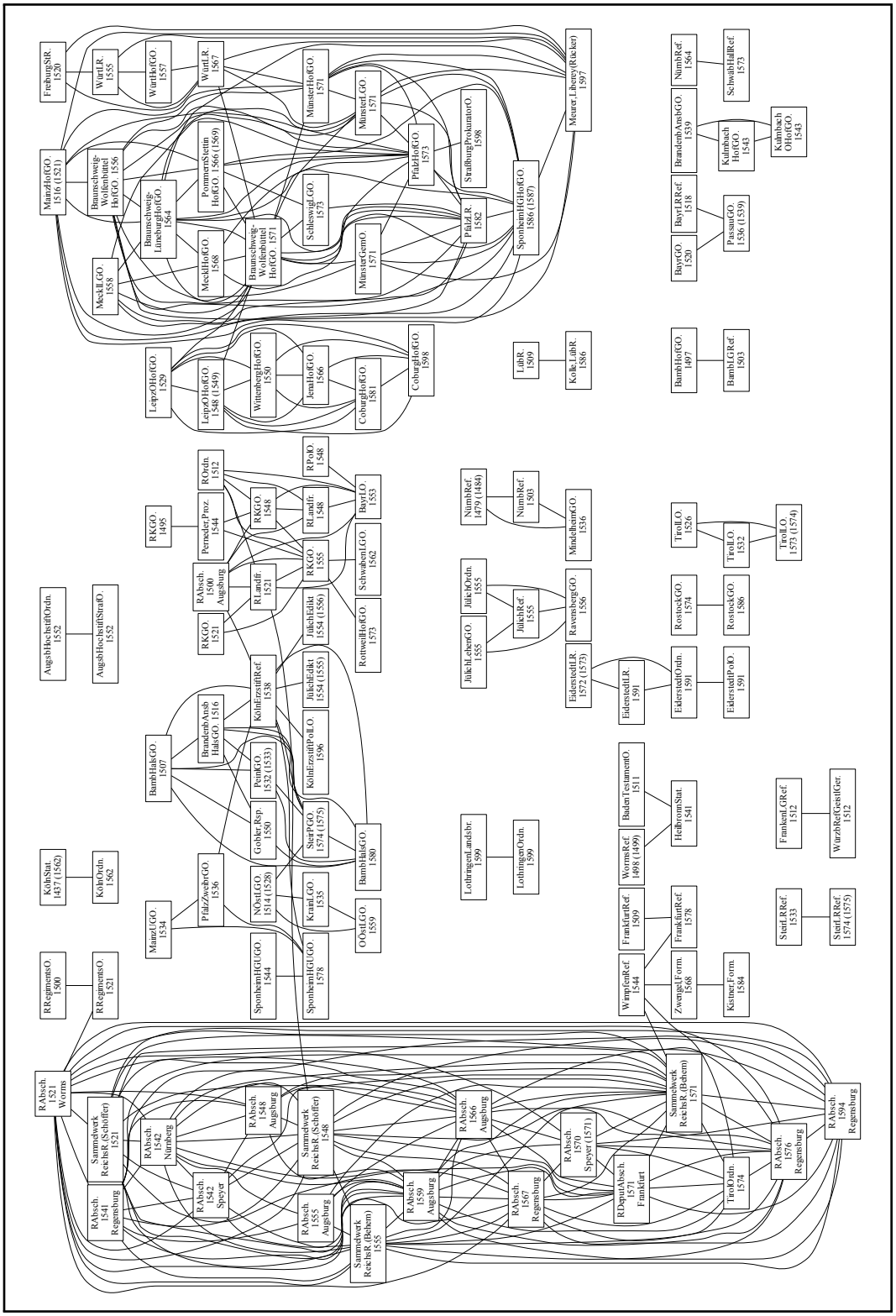


Abb. 3.23: Textpaare (alle Texte) mit gefilterten Matches ab 5 %, die dem auf S. 199 beschriebenen Textgruppenkriterium entsprechen (Codierung *ohne A/I/U/B/F* ohne Leerzeichen, Matches ab 18 Zeichen, vgl. zur Gestaltung S. 197)

3.4.3 Dotplotdarstellung und -auswertung von Matches

Eine rein quantitative Analyse der ermittelten Matchbereiche, wie sie bisher beschrieben wurde, kann zwar – wie gezeigt – jedenfalls für das hier untersuchte Korpus schon einen guten Überblick über ausgeprägtere Ähnlichkeiten zwischen Texten verschaffen und dabei auch eine inhaltlich recht plausible Textgruppierung ermöglichen, da dabei jedoch nur die Existenz und Länge der Matches (gegebenenfalls nach einer Überarbeitung oder Auswahl entsprechend Kapitel 3.3), nicht aber ihre Positionen ausgewertet werden, ist es damit nicht möglich, Textentsprechungen als signifikant zu erkennen, wenn diese zwar in einem bestimmten Textbereich einen recht großen Teil ausmachen, insgesamt aber zu kurz sind, um den Auswahlkriterien zu genügen. Und auch für Textpaare, deren Matches den angesetzten Mindestumfang erreichen, lässt sich so nicht feststellen, welche Abschnitte überhaupt und in welchem Ausmaß Matches enthalten und ob zum Beispiel die Textgruppen, die sich in Graphen auf der Basis solcher Kriterien abzeichnen, tatsächlich durch gemeinsame Textpassagen verbunden sind oder ob die Entsprechungen zwischen den einzelnen Textpaaren vielleicht gar nichts miteinander zu tun haben.

Auch ohne die betreffenden Textstücke selbst zu vergleichen, lässt sich in dieser Hinsicht jeweils für einzelne Textpaare ein wesentlich präziseres Bild gewinnen, wenn nach der oben in Unterkapitel 2.2.2 vorgestellten Dotplot-Methode in Diagrammform verzeichnet wird, welche Matchpositionen einander zugeordnet werden. Wie dort schon erwähnt, kann eine 1:1-Abbildung der Positionen in Punkte der Dotplot-Matrix allerdings insbesondere für umfangreiche Texte problematisch sein, und zwar sowohl hinsichtlich des Aufwands als auch hinsichtlich der auf dieser Grundlage erzeugten graphischen Darstellung: Da die Zahl der möglichen Positionspaare dem Produkt der beiden Textlängen entspricht, muss bei Verarbeitungsschritten, die jede Matrixzelle berücksichtigen, mit einem quadratisch steigenden Laufzeitverhalten gerechnet werden, und die Größe einer Rasterabbildung der Matrix nimmt (wenn die Auflösung nicht für die Darstellung reduziert wird) ebenfalls in diesem Maße zu, so dass sich bei größeren Textmengen nur schwer der Gesamtüberblick gewinnen lässt, der eigentlich durch dieses Verfahren ermöglicht werden soll.

Wenn man allerdings davon ausgeht, dass Matches insgesamt nur einen sehr kleinen Teil der möglichen Positionspaare abdecken, dürfte es in aller Regel unproblematisch sein, mit einer wesentlich weniger präzisen Repräsentation zu arbeiten. Für die hier vorgestellten Beispiele, die wiederum auf den entsprechend Tabelle 3.9 gefilterten Matches mit einer MEM-Mindestlänge von 18 Zeichen in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen basieren, wurde jeweils eine Abbildung der Matches auf die Matrixzellen vorgenommen, die sich aus den durch 100 geteilten Matchpositionen (vom ersten bis zum letzten Zeichen jedes Matchs) ergeben. Somit

sind in diesen Dotplotdarstellungen dann zwei benachbarte Punkte auf einer Diagonale belegt, wenn ein einzelnes Match in die entsprechenden Gruppen von jeweils 100 Zeichen hineinreicht oder wenn in jedem der beiden Blöcke ein Match zu finden ist, also in beiden Texten, je nachdem wo innerhalb des ersten Blocks das erste Match endet, maximal 99 bis 198 Zeichen zwischen den Matches liegen. Wenn sich in einer solchen Darstellung größere zusammenhängende Linien auf Diagonalen ergeben, bedeutet das also nicht zwangsläufig, dass eine weitgehend vollständige Zuordnung der betreffenden Textpassagen zueinander möglich ist, der Vergleich der erzeugten Diagramme zeigt aber, dass schon relativ kurze Linien nur selten auftreten und kaum rein zufällig sein dürften. Dass die Blöcke gerade 100 Zeichen groß sind, ist letztlich willkürlich gewählt, der Wert hat sich aber insbesondere auch für die Darstellung als recht gut verwendbar erwiesen. Der Effekt unterschiedlicher Blockgrößen wird am Ende dieses Unterkapitels noch untersucht.

Zu beachten ist, dass sich der Abstand eines Matchs vom Anfang des jeweils ersten zugehörigen Zeichenblocks in den beiden Texten in den meisten Fällen unterscheidet. Deshalb liegen die einem Match zuzuordnenden Punkte, wenn es die Zeichenblockgrenzen überschreitet, häufig nicht auf einer einzigen, sondern auf zwei benachbarten Diagonalen. Entsprechend können auch Matches, die einander in beiden Texten sehr nahe benachbart sind, nebeneinander liegenden Diagonalen zuzuordnen sein. Dies kann insbesondere relevant sein, wenn bei der Prüfung auf auffällige Strukturen Schwellenwerte für die Anzahl der Matchzeichen in einem Bereich festgelegt werden. Hier ist es wohl angemessen, jeweils auch die Nachbardiagonalen mit einzubeziehen, um eine Abwertung von Matches, denen zwei Diagonalen zuzuordnen sind, zu verhindern.

Zur Illustration bietet Abbildung 3.24 eine schematische Darstellung, in der die waagerechten und senkrechten Linien den Grenzen der Blockeinteilung entsprechen und diagonale Linien den tatsächlichen Matches. Die quadratischen Blöcke, die Matches enthalten, sind durch Graufärbung hervorgehoben. Diese Darstellung entspricht also nicht den im Folgenden präsentierten Dotplots auf der Basis von Blöcken, die Matches enthalten, vielmehr soll sie verdeutlichen, welche Unschärfe sich dabei gegenüber einer präzisen Verzeichnung der Matchpositionen ergibt. Während sich bei der rechten diagonalen Linie der Sonderfall zeigt, dass der Anfang eines Matchs in beiden Texten den gleichen Abstand von den Blockgrenzen hat und somit auch in beiden Texten an derselben Stelle die Blockgrenze überschritten wird, sind links einige Matches zu finden, bei denen der Abstand unterschiedlich ist, so dass die Diagonalen nicht durch die Ecken der quadratischen Blöcke laufen, sondern die waagerechten und senkrechten Grenzlinien an unterschiedlichen Punkten kreuzen und sich damit auf Quadrate verteilen, die zu zwei benachbarten Diagonalen (auf der Ebene der Zeichenblöcke) gehören, wobei den markierten Quadraten oft nur eine kleine Zahl von Matchzeichen zugeordnet werden kann.

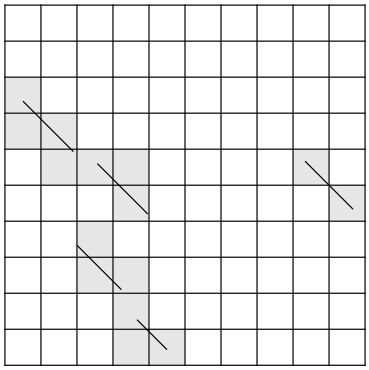


Abb. 3.24: Schema: Dotplot auf der Basis von Zeichenblöcken

Die Abbildungen 3.25–3.28 auf S. 219 und 222 enthalten jeweils ein bis drei Dotplotgraphiken und rechts neben jeder von ihnen zwei Säulendiagramme. Aus dem ersten kann in etwa entnommen werden, wie sich die Punkte auf die verschiedenen Diagonalen verteilen; dabei sind jeweils fünf Diagonalen in einer Säule zusammengefasst. Das zweite Säulendiagramm verzeichnet, wie groß die Zahl von Punkten (Matchblöcken) ist, die sich auf diagonalen Linien mit der auf der x-Achse angegebenen Länge befinden beziehungsweise (bei der Länge 1) isoliert sind.[604] Beide Säulendiagramme sind selbstverständlich mit einem erheblichen Informationsverlust verbunden, sie sollen aber auch nur verdeutlichen, wie sich auf der Basis der Punktverteilung in einer Dotplotmatrix einfache Berechnungen anstellen und damit Hinweise auf möglicherweise signifikante Strukturen entnehmen lassen.

Daneben sind auch andere Datenerhebungen zur Punktverteilung plausibel. Insbesondere könnte es sinnvoll sein, bei der Berechnung der jeweiligen Längen auch Linienunterbrechungen und horizontale oder vertikale Verschiebungen (also Auslassungen beziehungsweise Einfügungen in einem der Texte) in einem gewissen Maße zuzulassen, womit sich wohl eine bessere Annäherung an den optischen Eindruck zusammenhängender Linien auch bei kleinen Abweichungen von einer durchgehenden Diagonale erreichen ließe und insgesamt mit einer besseren Abgrenzung zusammengehöriger Textbereiche zu rechnen wäre. Allerdings wäre dies auch mit einer entsprechend dem Ausmaß geduldeter Unterbrechungen und Verschiebungen steigenden Berechnungskomplexität verbunden und insbesondere bei Dotplots mit einem relativ dichten Punktmuster mit zunehmenden Entscheidungsschwierigkeiten, welcher Punkt welcher Linie zuzuordnen ist. Eine weitere einfache Abbildungsvorschrift wäre die Zählung der Punkte in den einzelnen Reihen und Spalten – die Bedeutung von Häufungen auf diesen Achsen wird im Folgenden noch thematisiert.

[604] Es werden die verbundenen Matchblöcke gezählt. Die Zahl der dadurch gebildeten diagonalen Linien der angegebenen Länge ergibt sich, wenn man den Wert auf der y>-Achse durch die Länge, also den Wert auf der x-Achse, teilt.

Hintergrund für eine solche mathematische Aufbereitung der Daten ist, dass die Erstellung der Dotplots (und auch ihre Anzeige und Durchsicht) selbst bei der hier vorgenommenen Informationsreduktion durch Abbildung der Matchpositionen auf 100er-Zeichenblöcke einen zumindest für den Gesamtbestand an zu untersuchenden Textpaaren ganz erheblichen Aufwand bereitet und deshalb sinnvollerweise auf solche Fälle beschränkt werden sollte, bei denen signifikante Ergebnisse einigermaßen wahrscheinlich sind. Überlegungen zu einer Auswahl auf der Basis von lokalen Matchkonzentrationen finden sich am Ende dieses Unterkapitels.

In den Abbildungen sind jeweils solche Dotplots zusammengefasst, die in einer gewissen Hinsicht Ähnlichkeiten aufweisen. Zu betonen ist, dass die gewählte Auflösung aus darstellerischen Gründen sehr unterschiedlich ist. In den Säulendiagrammen lässt sich das an den Werten der x- und y-Achse erkennen, bei den Dotplots ist das allerdings nicht unmittelbar abzulesen. Die Auflösung lässt sich immerhin in etwa aus der Größe und dem Schwärzungsgrad isolierter Punkte erschließen.

Abbildung 3.25 auf S. 219 zeigt, welches Bild entsteht, wenn Textpaare sehr umfangreiche Entsprechungen haben. Dabei beruht Abbildung 3.25a auf einem Textpaar, bei dem der zweite Text (dessen 100er-Zeichenblöcke hier wie auch in den übrigen Dotplots den Zeilen der Matrix entsprechen) eine Überarbeitung und Erweiterung des ersten Textes darstellt. Dementsprechend ist die insgesamt gut erkennbare Diagonale immer wieder durch Lücken mit einer Verschiebung der Fortsetzung nach unten unterbrochen. Abbildung 3.25b beruht auf dem Vergleich eines Normtextes mit einem von diesem offenbar stark abhängigen Autorenwerk. Auch hier ist deutlich zu erkennen, dass der zweite Text mit dem ersten nicht nur durch eine Vielzahl von Entsprechungen verbunden ist, sondern dass diese Textpassagen auch zum allergrößten Teil in derselben Reihenfolge und mit einigermaßen ähnlichen Abständen zueinander wie im Ausgangstext zu finden sind, wobei es immer wieder kleinere Verschiebungen und an einigen Stellen auch größere Textstücke ohne Pendant gibt.

In Abbildung 3.25a lässt sich außerdem links unten eine relativ kurze Diagonale erkennen, und oberhalb sowie neben dieser Diagonale finden sich weitere Punkte, die sich trotz größerer Lücken zwischen ihnen als in etwa auf einer sehr steilen beziehungsweise einer sehr flachen Linie liegend wahrnehmen lassen. Diese Punkte lassen sich auf Matches zwischen jeweils einem Inhaltsverzeichnis und einem Haupttext zurückführen; die Diagonale links unten beruht auf Matches zwischen den Inhaltsverzeichnissen beider Texte, von denen eines vor und das andere nach dem jeweiligen Haupttext steht. Matches zwischen einem Inhaltsverzeichnis und einem Haupttext zeigen sich auch, aber wegen der Skalierung etwas weniger auffällig, in Abbildung 3.25b in einem schmalen Bereich am linken Rand. In beiden Dotplots finden sich daneben vergleichsweise wenige Punkte, die ganz oder fast isoliert sind

und vermutlich mehr oder weniger zufällige Entsprechungen widerspiegeln.

Die Auswertung der Punktverteilung auf den Diagonalen zeigt eine deutliche Konzentration in einem bestimmten Bereich, wobei die auf Einschüben beruhende mehrfache Verschiebung der Diagonale im Dotplot von Abbildung 3.25a als Entsprechung im zugehörigen oberen Säulendiagramm mehrere relativ nahe beieinander liegende, aber voneinander deutlich abgegrenzte lokale Maximalwerte hat und außerdem mit größerem Abstand dazu ein auf den Matches zwischen den Inhaltsverzeichnissen beruhendes weiteres lokales Maximum, während das entsprechende Diagramm von Abbildung 3.25b neben einer deutlich überwiegenden Hauptgruppe einen etwas breiteren Bereich mit mehreren im Vergleich wesentlich niedrigeren Spitzenwerten aufweist. Er lässt sich den Matches im letzten Viertel beziehungsweise Fünftel der Texte zuordnen, die im zweiten Text auf ein längeres Stück ohne Entsprechungen folgen und die im zweiten Text dichter beieinander stehen als im ersten; die entsprechende Linie, die man im Dotplot erkennen kann, ist deutlich abgeflacht.

Im jeweils zugehörigen unteren Säulendiagramm ist erkennbar, dass es trotz der starken Punktkonzentration, die den Gesamteindruck einer bei allen Unregelmäßigkeiten doch in etwa diagonal verlaufenden Linie hervorruft, in den hier betrachteten Fällen viele Zeichenblöcke mit Matches gibt, die mit keinem anderen Matchblock auf einer Diagonale unmittelbar verbunden sind – die dafür angegebene Häufigkeit ist in beiden Diagrammen mit Abstand die größte. Sie ist allerdings viel kleiner als die summierte Zahl von Matchblöcken, die sich einer zusammenhängenden Gruppe von zwei oder mehr Matchblöcken zuordnen lassen.[605]

Abbildung 3.26 auf S. 220 zeigt Dotplots von Textpaaren, die insgesamt recht wenige Übereinstimmungen aufweisen, darunter aber auch etwas längere oder solche, die so dicht und im mehr oder weniger gleichen Abstand aufeinander folgen, dass sie hier als kürzere diagonale Linien erscheinen. Auch dies wird in den Diagrammen der rechten Spalte gut widergespiegelt.

Allerdings lässt sich leicht feststellen, dass die Zahl der in den Abbildungen 3.26a und 3.26b deutlich abgegrenzten Matchbereiche größer ist als die Zahl der Säulen in den Diagrammen, die die Verteilung auf verschiedene Diagonalen dokumentieren. Ganz abgesehen davon, dass durch die Zusammenfassung mehrerer Diagonalen in einer Säule eine zusätzliche Unschärfe entsteht, ist die Abbildung der Punktpositionen auf die Differenz der zugehörigen Positionen natürlich mit einem erheblichen Informationsverlust verbunden, und es liegt in der Natur der

[605] Beim Vergleich von BraunschweigWolfenbüttelHofGO. 1556 mit BraunschweigWolfenbüttelHof GO. 1571 werden insgesamt 854 Blöcke mit Matches gefunden, davon 172 ohne direkte Nachbarn und 682, die sich auf 145 Gruppen von benachbarten Matchblöcken verteilen. Beim Vergleich von WormsRef. 1498 (1499) mit Gobler,StatB. 1553 gibt es 286 isolierte Matchblöcke und weitere 1665, die sich in derselben Weise zu 320 Gruppen zusammenfassen lassen.

Sache, dass isolierte Übereinstimmungen, die überhaupt nichts miteinander zu tun haben und weit voneinander entfernt zu finden sind, auf derselben Achse liegen können.

Wenn Übereinstimmungen in einem Textpaar aber insgesamt so selten sind wie in den in dieser Abbildung untersuchten Paaren, können auch schon etwas höhere Werte, wie sie hier verzeichnet sind, einen guten Anhaltspunkt dafür darstellen, zumindest eine Überprüfung der Matchverteilung über einen Dotplot vorzunehmen.

Die Relevanz der in den Dotplots erkennbaren Bereiche mit auffälliger Häufung von Entsprechungen gerade in Fällen, in denen sie in Relation zu den Gesamttexten nur einen sehr kleinen Teil ausmachen, ist natürlich anhand der konkreten Textstellen zu prüfen. Allerdings kann zumindest manchmal schon die Positionierung der Entsprechungen gewisse Hinweise geben. So ist in Abbildung 3.26b auffällig, dass die beiden kürzeren diagonalen Linien fast in der oberen linken beziehungsweise unteren rechten Ecke zu finden sind, also jeweils fast am Anfang beziehungsweise Ende der beiden Texte. Das legt die Vermutung nahe, dass es sich um formelhafte Passagen handelt, wie sie in Normtexten der Zeit vielfach in der Einleitung und am Schluss zu finden sind.

Abbildung 3.27 auf S. 221 soll verdeutlichen, dass übereinstimmende Stellen, die mehrfach in einem der Texte oder auch in beiden vorkommen, in Dotplots anhand von Punkten zu erkennen sind, die in waagerechten beziehungsweise senkrechten Linien angeordnet sind, wenn auch oft mit größeren Abständen. Gezeigt wird der Vergleich zwischen zwei Formularbüchern, also Texten, bei denen viele der übereinstimmenden Formulierungen sehr häufig vorkommen. Um die Anzeige nicht zu stark verkleinern zu müssen, wird vom Dotplot allerdings nur in etwa das oberste Viertel gezeigt; die Säulendiagramme beziehen sich aber auf den vollständigen Dotplot.

Im hier gezeigten Dotplotausschnitt sind wohl insbesondere zum einen drei parallele senkrechte Linien in etwa am Ende des oberen Drittels auffällig, zum anderen Punkthäufungen in einigen Zeilen im unteren Drittel, die trotz größerer Lücken den Gesamteindruck von waagerechten Linien erwecken, die zudem eine zu einem erheblichen Teil gleiche Punktverteilung aufweisen – man könnte sie also auch als eine Vielzahl von senkrechten Linien mit sehr großen Abständen zwischen den Punkten betrachten. Die Säulendiagramme zeigen deutlich, dass in diesem Dotplot kaum mit signifikanten Matches zu rechnen ist. Schon zwei zusammenhängende Matchblöcke sind selten und zum größten Teil darauf zurückzuführen, dass sich ein kürzeres Match zufällig auf zwei Blöcke verteilt, und Matchbereiche mit mehr als zwei Blöcken sind gar nicht vorhanden. Und auch die Zählung der Punkte auf den Diagonalen weist keine irgendwie signifikanten Spitzenwerte auf – dass sich vereinzelt in einem Bereich von fünf Diagonalen bis zu acht Blöcke mit Matches finden, kann (insbesondere in Relation zur Gesamtzahl der ermittelten Matchbe-

reiche) gut durch zufällig ähnliche Abstände der betreffenden Matches von den Textanfängen erklärt werden.

Als letzte in dieser Reihe soll Abbildung 3.28 auf S. 222 veranschaulichen, welche Muster ein Dotplot zeigt, wenn die in einem einzigen Text enthaltenen Teilstrings miteinander verglichen werden. Dabei ist allerdings zu beachten, dass hier (entsprechend der zugrunde liegenden Matchliste) nur Teilübereinstimmungen verzeichnet werden, nicht aber, dass der Text insgesamt natürlich mit sich selbst übereinstimmt. Dementsprechend fehlt hier die durchgehende Linie auf der zentralen Diagonale von der linken oberen zur rechten unteren Ecke. Manchmal werden aber doch Punkte auf dieser Diagonale verzeichnet, nämlich wenn die Positionen in einem Match so nahe beieinander liegen, dass sie in denselben 100er-Block fallen. Die Dotplots dieser Abbildung sind jeweils symmetrisch zur Diagonale von der linken oberen zur rechten unteren Ecke, da die beiden Positionen eines Matchs natürlich vertauscht werden können, wenn es sich um denselben Text handelt.

Abbildung 3.28a zeigt die Wiederholungen in einem der Texte, die Abbildung 3.27 zugrunde liegen, und lässt noch deutlicher erkennen, wie stark dieser Text durch wiederholt vorkommende Formulierungen geprägt ist, aber auch, dass sie sich nicht einigermaßen zufällig verteilen, sondern dass es neben Abschnitten mit starker Formelhaftigkeit offenbar auch größere Bereiche gibt, die jedenfalls im selben Text keine Entsprechung haben.

Insgesamt typisch für Dotplots, die Übereinstimmungen innerhalb eines einzigen Textes visualisieren, ist jedenfalls im hier untersuchten Korpus die Häufung dieser Übereinstimmungen in der Nähe der Hauptdiagonale von links oben nach rechts unten. Dass dies (jedenfalls in einem gewissen Maße) auch für Abbildung 3.28a gilt, ist zwar wohl im Dotplot nicht unmittelbar augenfällig, aber jedenfalls aus dem oberen Säulendiagramm zu entnehmen. In diesem Fall kann es aber wohl als sekundärer Effekt betrachtet werden, da die Punkthäufung in diesem Bereich zu einem nicht geringen Teil einem quadratischen Bereich im unteren rechten Viertel des Dotplots zugeordnet werden kann und sich darin Spalten und Zeilen kreuzen, die aufgrund vielfacher Übereinstimmungen als solche optisch hervortreten; die der Hauptdiagonale nahe stehenden Punkte in diesem Bereich dürften also mit recht hoher Wahrscheinlichkeit ebenfalls auf hier besonders stark gehäuft auftretende feststehende Formulierungen zurückzuführen sein.

In Texten, in denen sich keine solche Häufung von Mehrfachübereinstimmungen findet, ist es aber, wie gesagt, auffällig, dass die Matches relativ oft in einem recht engen Textumfeld zu finden sind. Das mag verschiedene Gründe haben – zu denken wäre wohl insbesondere an den jeweiligen Sachzusammenhang und an sprachliche Charakteristika des hier untersuchten Korpus, aber möglicherweise auch daran, dass beim Verfassen eines Textes die kurz zuvor gebrauchten Wörter und Wendungen noch irgendwie präsent sein können und deshalb leichter erneut gebraucht werden als an ganz anderer Stelle.

Matchzeichenzahl mindestens	Blockgröße 1	Blockgröße 100	Blockgröße 200	Blockgröße 300	Blockgröße 400
50	3394	6340	6831	7107	7283
100	1111	3709	3943	4154	4284
150	541	2341	2599	2771	2897
200	324	1723	1959	2072	2153
250	212	1288	1488	1579	1642
300	164	1092	1280	1356	1433
350	129	951	1113	1207	1250
400	104	833	964	1059	1111
450	83	703	844	923	970
500	57	635	773	849	894

Tab. 3.14: Dotplots mit Mindestanzahl von Matchzeichen in einer zusammenhängenden Gruppe von Blöcken, die Matches enthalten

Wie in Abbildung 3.27 lässt sich auch in Abbildung 3.28 feststellen, dass die Punkte in den Dotplots zum größten Teil keine unmittelbaren Nachbarn auf der jeweiligen Diagonale haben, wobei es auch hier in Abbildung 3.28a und Abbildung 3.28c einzelne etwas größere zusammenhängende Bereiche mit Matches gibt. Der deutliche Unterschied zur Verteilung in Abbildung 3.25 und 3.26, wie er sich aus den jeweils unteren Säulendiagrammen entnehmen lässt, legt die Vermutung nahe, dass diese Verteilung als Basis für ein Auswahlkriterium dienen kann, nach dem sich vorab einschätzen lässt, inwieweit eine Dotplotvisualisierung auffällige Strukturen enthält, die Hinweise auf Bereiche mit Textübernahmen geben könnten.

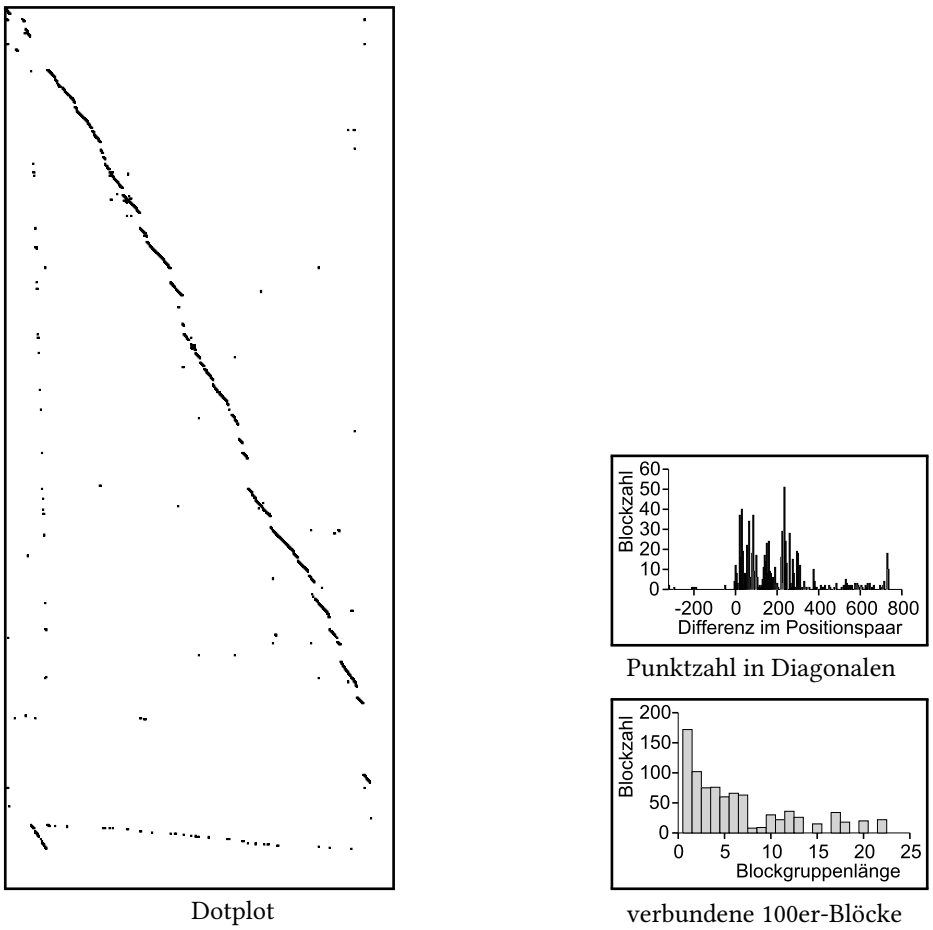
Dabei kann auch eine wesentlich feinere Auswertung erfolgen, als sie diesen Säulendiagrammen und auch den Dotplots zugrunde liegt. Da darin nur berücksichtigt wird, ob sich einem Zeichenblock überhaupt ein Match zugeordnet werden kann, ist nicht erkennbar, welches Ausmaß die Matches in diesem Bereich oder auch in einer zusammenhängenden Blockgruppe erreichen. Es lässt sich aber auch leicht zählen, wie viele Zeichen eines Matchs in einem Block liegen. Da sich – wie oben beschrieben – die Zeichen eines Matchs in vielen Fällen zwei benachbarten Diagonalen zuordnen lassen, ist es sinnvoll, dabei jeweils auch die entsprechende Zahl der direkt daneben liegenden Diagonalenabschnitte mit einzubeziehen. Damit ergeben sich Werte, die zum Beispiel jeweils für die festgestellten zusammenhängenden oder nur wenig unterbrochenen Diagonalen-Blockgruppen, die Matches enthalten, addiert werden können. Das bei diesem Verfahren erreichte Wertespektrum ist viel größer als in den bisher präsentierten Säulendiagrammen und führt insbesondere auch zu einer Aufwertung von eher kurzen Bereichen, die einen hohen Anteil übereinstimmender Zeichen enthalten.

Tabelle 3.14 zeigt, in wie vielen der Textpaare sich mindestens eine Blockgruppe mit einem oder mehreren Matches findet, deren (für die Blockgruppe) summierte Zeichenzahl einen bestimmten Schwellenwert überschreitet. Dabei werden unterschiedliche Größen für die Blockeinteilung von 1 bis 400 verglichen, um den

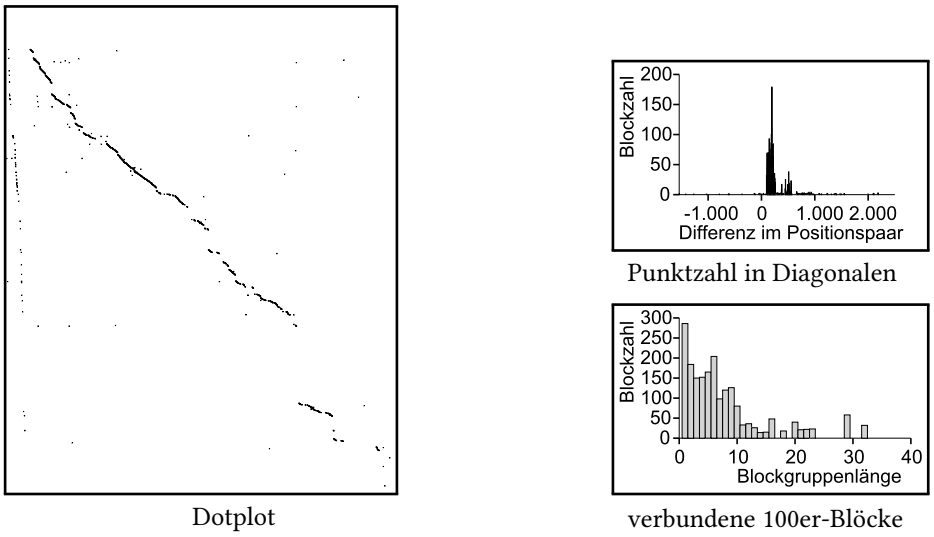
Effekt dieser Einteilung erkennbar zu machen. Aus der Spalte zur Blockgröße 1 kann entnommen werden, wie viele Textpaare es gibt, bei denen die angegebene Zeichenzahl in einem zusammenhängenden Match vorkommt; je größer die zugrunde gelegten Blöcke sind, desto größere Abstände und Verschiebungen zwischen Matches werden für die Erkennung von zusammenhängenden Bereichen toleriert und desto mehr Textpaare gibt es, die den jeweiligen Schwellenwert überschreiten. Man kann deutlich erkennen, dass die Zahl der Textpaare, die dem Auswahlkriterium genügen, zwar in jedem Fall bei einer Erhöhung des Schwellenwerts stark abnimmt, dass diese Abnahme aber bei Verzicht auf eine Blockbildung viel stärker ausfällt: Während von den 3394 Textpaaren, die Matches mit einer Länge von mindestens 50 Zeichen enthalten, nur 57 und damit weniger als 1,7 % auch Matches mit einer Länge über 500 Zeichen aufweisen, sind es bei einer Zusammenfassung von Matchbereichen in 100er-Blöcken 635 von 6340 und damit gut 10 % und bei 400er-Blöcken 894 von 7283, also über 12 %.

Aus der Tabelle lässt sich nicht entnehmen, welcher Schwellenwert und welche Blockgröße besonders gut geeignet sind, um eine Auswahl unter dem Gesichtspunkt zu erreichen, dass ein Textbereich durch stärker ausgeprägte Entsprechungen auffällt – auch hier hängt eine entsprechende Entscheidung davon ab, was das genaue Erkenntnisinteresse ist und wie *Precision* und *Recall* gewichtet werden. In vielen Fällen dürfte es sinnvoll sein, noch weitere Kriterien mit einzubeziehen. So scheint es zum Beispiel recht plausibel, Textpaare mit etwas längeren Matches auch dann zu berücksichtigen, wenn ein höherer Schwellenwert bei einer Blockeinteilung nicht erreicht wird, dabei aber Matches nicht zu berücksichtigen, die nach bestimmten Bewertungskriterien als weniger aussagekräftig eingestuft werden.[606] Soweit aber überhaupt eine Auswahl anhand von Länge und/oder Häufung der Matches in Teilbereichen sinnvoll erscheint, lässt sich offenbar schon mit einem niedrigen Schwellenwert und einer relativ hohen Blockgröße ohne die Berücksichtigung weiterer Merkmale der Matches eine deutliche Reduktion der näher zu betrachtenden Textpaare erreichen, jedenfalls wenn – wie bei den hier zugrunde gelegten Matchkriterien – die ermittelten Matches zu einem erheblichen Teil so kurz sind, dass sie mit nicht geringer Wahrscheinlichkeit auf sprachliche Muster beziehungsweise zufällige Übereinstimmungen zurückzuführen sind.

[606] Vgl. oben Unterkapitel 3.3.2.

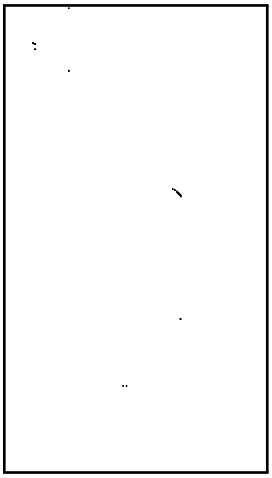


(a) Vergleich von BraunschweigWolfenbüttelHofGO. 1556 mit BraunschweigWolfenbüttelHofGO. 1571

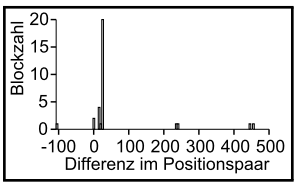


(b) Vergleich von WormsRef. 1498 (1499) mit Gobler,StatB. 1553

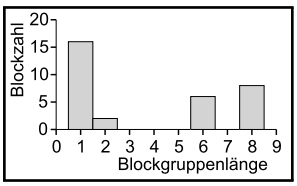
Abb. 3.25: Dotplots und Diagramme zur Punktzahl in Diagonalen in Textpaaren mit weitgehend unveränderten umfangreichen Übernahmen (vgl. S. 212 und 213)



Dotplot

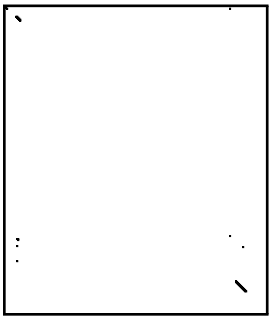


Punktzahl in Diagonalen

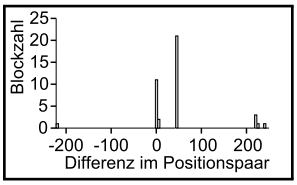


verbundene 100er-Blöcke

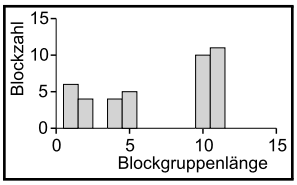
(a) Vergleich von MainzUGO. 1534 mit HeilbronnStat. 1541



Dotplot

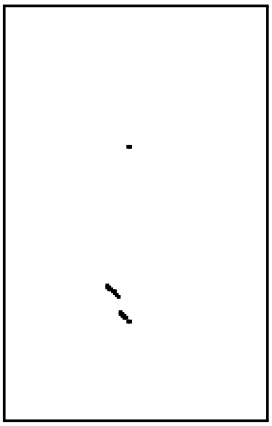


Punktzahl in Diagonalen

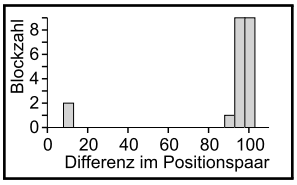


verbundene 100er-Blöcke

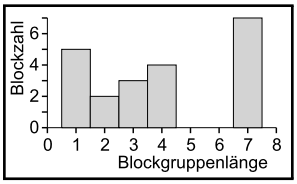
(b) Vergleich von KärntLGO. 1577 (1578) mit KärntLRO. 1577 (1578)



Dotplot



Punktzahl in Diagonalen

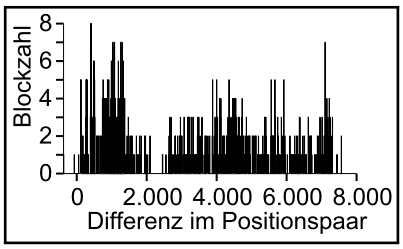


verbundene 100er-Blöcke

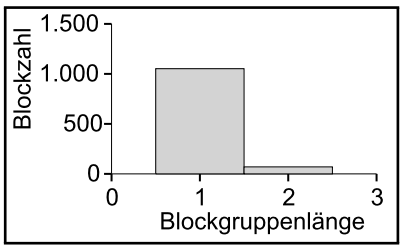
(c) Vergleich von SchlesLO. 1577 mit BreslauStat. 1588

Abb. 3.26: Dotplots und Diagramme zur Punktzahl in Diagonalen in Textpaaren mit etwas längeren Übereinstimmungen in kleinen Bereichen (vgl. S. 212 und 214)

Dotplotausschnitt

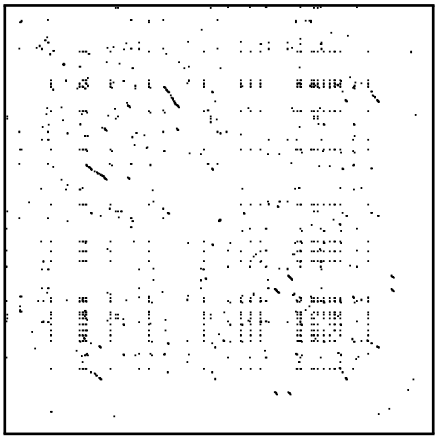


Punktzahl in Diagonalen

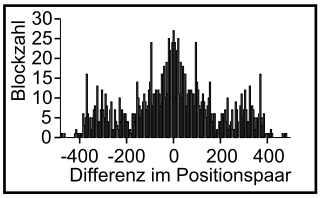


verbundene 100er-Blöcke

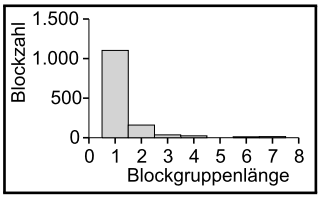
Abb. 3.27: Dotplotausschnitt und Diagramme zur Punktzahl in Diagonalen in einem Textpaar mit vielen häufigen Formulierungen (Kistner,Form. 1584 und Zwengel,Form. 1568; vgl. S. 212 und 215)



Dotplot

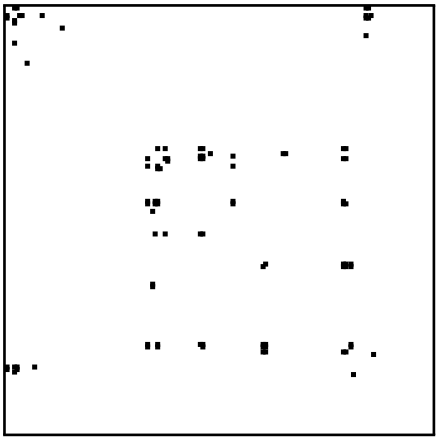


Punktzahl in Diagonalen

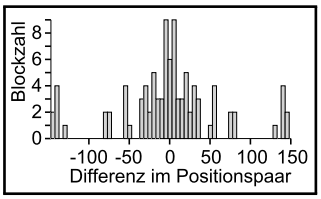


verbundene 100er-Blöcke

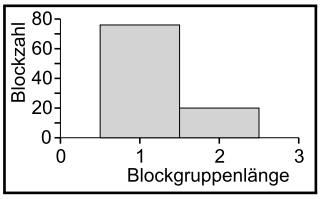
(a) Übereinstimmungen innerhalb von Kistner,Form. 1584



Dotplot

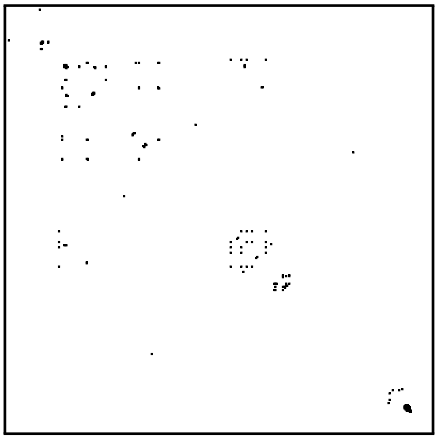


Punktzahl in Diagonalen

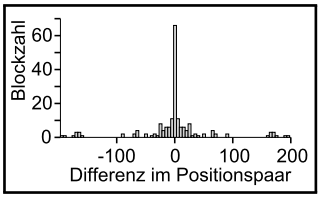


verbundene 100er-Blöcke

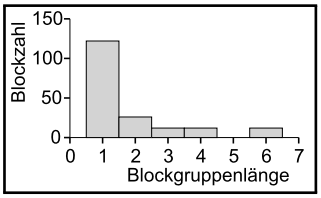
(b) Übereinstimmungen innerhalb von KrainLGO. 1535



Dotplot



Punktzahl in Diagonalen



verbundene 100er-Blöcke

(c) Übereinstimmungen innerhalb von Pölmann,Urteil 1577

Abb. 3.28: Dotplots und Diagramme zur Punktzahl in Diagonalen bei der Verzeichnung von wörtlichen Übereinstimmungen in einem einzigen Text (vgl. S. 212 und 216)

3.4.4 Projektionsdotplots

Die in Unterkapitel 3.4.3 beschriebene Visualisierung von Matches in Form von Dotplots hat den Vorteil, dass sie einen guten Überblick ermöglicht, wo Matches zu finden sind und inwieweit sie im Hinblick auf ihren Kontext zusammenhängen, sofern es sich nicht um nur ganz kurze Übereinstimmungen handelt und nicht um Texte mit einer Fülle von Mehrfachübereinstimmungen, in denen solche Zusammenhänge optisch untergehen. Damit lassen sich wesentlich leichter und sicherer Aussagen über kurze, aber signifikante Übereinstimmungen machen als nach der in Unterkapitel 3.4.1 vorgestellten Analyse auf der Basis einer Auszählung der zu Matches gehörenden Zeichen beziehungsweise der in Unterkapitel 3.4.2 gezeigten Visualisierung der Auszählungsergebnisse. Dem stehen als Nachteile der in der Summe erhebliche Aufwand bei der Generierung von Dotplots für jedes einzelne Textpaar gegenüber, wenn eine größere Zahl von Texten in dieser Weise verglichen werden soll, sowie insbesondere das Fehlen eines Überblicks über Zusammenhänge zwischen Textgruppen.

Eine Möglichkeit, sowohl den Aufwand erheblich zu reduzieren als auch einen solchen Überblick zu gewinnen, besteht darin, das Dotplotverfahren so zu variieren, dass in den verschiedenen Spalten der Matrix zwar weiterhin die Matchpositionen in einem bestimmten Text verzeichnet werden, in den verschiedenen Zeilen hingegen nicht die Positionen in einem weiteren Text, sondern nur jeweils die Existenz eines Matchs in einem bestimmten Text. Eine solche Zeile entspricht also der Projektion eines Dotplots für ein Textpaar auf eine waagerechte Linie. Dementsprechend soll für die gesamte Matrix beziehungsweise die entsprechende Darstellung hier der Terminus „Projektionsdotplot“ verwendet werden.[607] Das Verfahren bietet sich für einen Textvergleich auf der Basis von Matches einer nicht zu geringen Mindestlänge an, weil diese Matches insgesamt so selten auftreten, dass es auch bei der Abbildung der Punkte eines Dotplots im üblichen Sinn auf eine einzige Zeile in vielen Fällen eine gute Unterscheidung zwischen signifikanten Mustern und einer eher zufälligen Verteilung ermöglicht. Dabei ändert sich gegenüber normalen Dotplots allerdings das Erscheinungsbild der Muster: Aus diagonalen Linien werden waagerechte Linien.

Natürlich kann es sein, dass dabei auch Matchbereiche zu einer einzigen Linie zusammengefasst werden, die auf Matches des primär betrachteten Textes mit ganz unterschiedlichen Stellen des anderen Textes beruhen, und insbesondere sind Einschübe im anderen Text als solche nicht mehr zu erkennen. Die Aussagekraft für die Matches eines bestimmten Textpaars ist also natürlich wesentlich geringer als die eines normalen Dotplots. Dem steht aber als erheblicher Gewinn gegenüber, dass sich bei Gruppen von Texten mit einer signifikanten Häufung von Matches

[607] Die Projektion von Dotplots wird auch in Landès/Hénaut/Risler 1993 beschrieben, dient dort allerdings als Grundlage für eine mathematisch komplexe Auswertung. Vgl. dazu oben S. 73.

Matchzeichenzahl mindestens	Blockgröße 1	Blockgröße 100	Blockgröße 200	Blockgröße 300	Blockgröße 400
50	3638	7384	7873	8146	8317
100	1492	4572	5019	5265	5476
150	852	3240	3669	3936	4099
200	542	2560	2902	3095	3249
250	367	2080	2384	2550	2696
300	269	1763	2040	2183	2301
350	205	1565	1800	1953	2044
400	164	1379	1605	1734	1836
450	144	1243	1439	1571	1664
500	132	1147	1325	1436	1515

Tab. 3.15: Zeilen in Projektionsdotplots mit Mindestanzahl von Matchzeichen in einer zusammenhängenden Gruppe von Blöcken, die Matches enthalten

oft unmittelbar erkennen lässt, inwieweit diese Matches bezogen auf den jeweils primär betrachteten Text an denselben Stellen zu finden sind.

Tabelle 3.15 soll zeigen, wie sich das Projektionsverfahren auf die Bildung zusammenhängender Blockgruppen auswirkt. Dazu wird wie in Tabelle 3.14 auf S. 217 verzeichnet, wie viele Texte bei verschiedenen Blockgrößen bestimmte Schwellenwerte hinsichtlich der Zahl der in einer Blockgruppe enthaltenen Matchzeichen überschreiten. Der Vergleich der beiden Tabellen zeigt, dass die Zahl der dem jeweiligen Auswahlkriterium genügenden Textpaare durch die Projektion zwar nicht unerheblich, aber doch insgesamt relativ moderat steigt.

Auch für die Untersuchung von Textgruppen mithilfe von Projektionsdotplots sollen einige Beispiele vorgestellt werden.[608] In den Abbildungen 3.29–3.31 wird in jeder Zeile am linken Rand zunächst durch ein kleines mehr oder weniger dunkles Quadrat visualisiert, wie hoch der Anteil der Blöcke in dieser Zeile ist, für die eine Übereinstimmung festgestellt wurde; danach folgt die Sigle des Textes, um den es in der Zeile geht, und schließlich die Projektion des Dotplots für den Vergleich dieses Texts mit dem gemeinsamen Vergleichstext aller Zeilen des Projektionsdotplots. Die Übereinstimmungsbereiche werden zwar generell schwarz dargestellt, die Matches beim Vergleich mit demselben Text allerdings zur Differenzierung grau, da Übereinstimmungen in dieser Zeile natürlich in aller Regel (soweit es sich nicht um Textsammlungen handelt) gerade nicht als Anzeichen für Textübernahmen im hier intendierten Sinn zu werten sind, sondern vielmehr einen Eindruck vermitteln, inwieweit ein Text durch Formulierungswiederholungen geprägt ist. Der Beginn der eigentlichen Dotplotdarstellung rechts von den Siglen ist durch eine senkrechte Linie gekennzeichnet. Übereinstimmungen sind jeweils durch Rechtecke bezeichnet, deren Höhe in etwa der Schrifthöhe entspricht, um das gehäufte Auftreten von

[608] Die Projektionsdotplots basieren ebenfalls auf den entsprechend Tabelle 3.9 gefilterten Matches der Codierung *ohne A/I/U/B/F* ohne Leerzeichen mit der MEM-Mindestlänge 18.

Matches in einzelnen Spalten leichter erkennbar zu machen. Die unterschiedliche Färbung des quadratischen Blocks links von der Sigle soll nur dazu dienen, Siglen in Zeilen mit höherem Schwärzungsgrad etwas hervorzuheben, um den Überblick zu erleichtern.[609] Es handelt sich dabei nicht um eine elaborierte Bewertung. Insbesondere ist zu betonen, dass es in einer Zeile durchaus auffällige gänzlich oder fast durchgängig schwarz gefärbte Stücke geben kann, ohne dass diese einen hohen Anteil der Zeile füllen und damit zu einer deutlichen Färbung des Quadrats links von der Sigle führen.

Abbildung 3.29 auf S. 230 zeigt den Projektionsdotplot zur Bambergischen Halsgerichtsordnung (auch bekannt als *Bambergensis*) und damit zu einem der bekanntesten Texte des Korpus. Die umfassenden Übernahmen aus diesem Text beziehungsweise aus der von ihm abhängigen *Peinlichen Gerichtsordnung* (der *Carolina*) wurden schon oben auf S. 37 zusammenfassend beschrieben.

Die dort aufgeführten Texte treten – soweit die genannten Texte zum Untersuchungskorpus gehören – ebenso wie die revidierte Fassung von 1580 im Projektionsdotplot deutlich hervor; ebenfalls auffällig ist die Zeile zum *Rechten Spiegel* von Justin Gobler, wenn auch die Lücken darin größer sind; und das ihm zugeschriebene *Statuten Bůch* weist offenbar zwar nicht ganz so umfassende, aber doch recht erhebliche Übereinstimmungen auf.

Natürlich kann aufgrund des hier gewählten Verfahrens, nur jeweils die Existenz mindestens eines Matchs der angesetzten Minimallänge innerhalb von jeweils 100 Zeichen zu verzeichnen, aus einer durchgehenden Linie nicht auf eine vollständige Übereinstimmung geschlossen werden. So ist zwar zum Beispiel erkennbar, dass bei der Revision der *Bambergensis* von 1580 keine größeren Stücke ausgelassen oder völlig abgeändert wurden, aber nicht, welche kleineren Textabweichungen festzustellen sind. Und da hier nur auf die Positionen in der Bambergischen Halsgerichtsordnung von 1507 Bezug genommen wird, lässt sich ebenso wenig ablesen, ob es in der revidierten Fassung Einschübe von einer etwas größeren Länge gibt, die in einem normalen Dotplot für dieses Textpaar – oder auch im Projektionsdotplot zur Fassung von 1580 – auch bei der hier angewandten Unschärfe hinsichtlich der Matchpositionen durch eine deutliche Verschiebung der jeweils folgenden einander entsprechenden Positionen (beziehungsweise im eben angegebenen Projektionsdotplot durch eine Lücke) erkennbar wären.

Man kann auf der Basis des Projektionsdotplots teilweise recht plausible Vermutungen über textuelle Abhängigkeitsverhältnisse anstellen. So entspricht die

[609] Da eine Einfärbung auch bei einem minimalen Entsprechungsanteil wenig aussagekräftig und zudem im Druck nicht gut darstellbar wäre, wurde der ermittelte Entsprechungsanteil zunächst auf die zweite Stelle nach dem Komma abgerundet. Um eine deutliche Hervorhebung auffälliger Werte zu erreichen, wurde anschließend – entsprechend der Gammakorrektur zum Beispiel in der Bildbearbeitung – die Quadratwurzel gezogen (also ein Gammawert von 0,5 zugrunde gelegt).

Verteilung der Bereiche mit und ohne Übereinstimmungen in der *Kölner Reformation* offenbar der in der *Peinlichen Gerichtsordnung*, und das beruht auch tatsächlich darauf, dass diese Ordnung für das Kölner Erzstift übernommen wurde. Bei einem nicht so sicheren Befund kann der Vergleich der Dotplots für die verschiedenen Textpaare möglicherweise ein klareres Bild ergeben. Zudem kann anhand der Matchpositionen überprüft werden, ob es Textstücke gibt, zu denen sich nur in einem der als Vorlage in Betracht zu ziehenden Texte Entsprechungen finden und ob diese zum Beispiel nach den in Unterkapitel 3.3.2 beschriebenen Kriterien als signifikant zu betrachten sind. Im hier betrachteten Fall ist anhand des Projektionsdotplots wohl keine Einschätzung sinnvoll, ob die steirische *Landt vnd Peindlich Gerichts Ordnung* auf die *Carolina* zurückzuführen ist. Die gleiche Frage stellt sich für den *Rechten Spiegel* und das *Statuten Bůch*. Bei den Dotplots für die einzelnen Textpaare zeigen sich aber für den Vergleich mit der *Peinlichen Gerichtsordnung* etwas stärker zusammenhängende diagonale Linien, so dass die auch rechtshistorisch plausible Vermutung, dass in diesen Texten die *Carolina* als Vorlage verwendet wurde, bestärkt wird.

Es gibt noch einige weitere Texte, bei denen etwas breitere zusammenhängende Blöcke geschwärzt sind, so dass es naheliegt zu überprüfen, ob dies auf einer punktuellen Übernahme beruht. Daneben ist deutlich zu erkennen, dass in einigen Spalten gehäuft Übereinstimmungen verzeichnet sind, insbesondere auch Übereinstimmungen mit Texten, die offenbar keine umfangreicheren Passagen von der Bambergischen Halsgerichtsordnung von 1507 als Vorlage genutzt haben. Hier lässt sich vermuten, dass diese Übereinstimmungen auf gängige Formulierungen zurückzuführen sind und für die Ermittlung von Textabhängigkeiten vernachlässigt werden können.

Abbildung 3.30 und 3.31 auf S. 231 und 232 können aufgrund der Größe der primär betrachteten Texte und der Vielzahl der Texte mit Entsprechungen nur in starker Verkleinerung präsentiert werden. Dadurch treten aber senkrechte und waagerechte Linien mit einer Vielzahl von Schwärzungen umso deutlicher hervor.

In Abbildung 3.30 ist insbesondere die Häufung von Übereinstimmungen etwa am Ende des ersten Textdrittels bemerkenswert. In der *Frankfurter Reformation* von 1578, deren Matches mit anderen Texten diesem Projektionsdotplot zugrunde liegen, sind in diesem Textbereich die der Reichsstadt von drei Kaisern verliehenen Appellationsprivilegien zu finden und damit Texte, bei denen schon aufgrund bestimmter üblicherweise in Privilegien vorhandener Bestandteile wie etwa der *Intitulatio* mit einer Vielzahl von Übereinstimmungen im Korpus zu rechnen ist. Inwieweit die ermittelten Matches in diesem Bereich nur auf solche allgemeinen Formulierungsmuster der kaiserlichen Kanzlei oder insbesondere auch auf ein einigermaßen feststehendes Formular speziell für Appellationsprivilegien zurückzuführen sind, lässt sich aus dem Dotplot nicht klar entnehmen, es ist aber auffällig, dass sich die

Matches in vier kleineren Abschnitten in diesem Bereich konzentrieren und die Stücke dazwischen nur in vergleichsweise wenigen Texten Entsprechungen haben. Das legt die Vermutung nahe, dass den vier kleineren Abschnitten Einleitungs- und Schlusspassagen der Privilegien zuzuordnen und vor allem Übereinstimmungen in den Bereichen dazwischen als Hinweis auf einen ähnlichen Privilegienhaupttext zu deuten sind.

Daneben fällt in diesem Projektionsdotplot insbesondere eine zu einem vergleichsweise großen Teil schwarze Linie auf. Sie stellt die Übereinstimmungen mit der *Liberey Keyserlicher / Auch Teutscher Nation Landt vnd Statt Recht* von Noe Meurer (hier in der Bearbeitung von Nikolaus Rücker) dar, also mit einem Werk, das zu einem erheblichen Teil umfängliche Zitate aus verschiedenen Normtexten enthält. Und es gibt auch einige weitere Texte, mit denen die Entsprechungen häufig und konzentriert genug sind, um im Projektionsdotplot den Eindruck einer trotz starker Unterbrechungen zusammenhängenden waagerechten Linie zu erwecken, und einige Matchbereiche, die aufgrund der Matchlänge oder der nahen Nachbarschaft mehrerer Matches trotz der Verkleinerung als auffällig geschwärzt herausstechen.

In Abbildung 3.31, die aufgrund ihres Formats um 90° gedreht gezeigt wird, zeichnet sich etwa zu Beginn des letzten Viertels ein Bereich ab, der in der Mehrzahl der Zeichenblöcke Übereinstimmungen mit bestimmten Texten aufweist. Bei einer Vergrößerung des Projektionsdotplots ist erkennbar, dass es sich dabei um die schon betrachtete Gruppe der Bambergischen Halsgerichtsordnung und der davon abhängigen Texte handelt. Offensichtlich nicht in diesen Zusammenhang einzuordnen ist die fast durchgängige schwarze Linie, die im Anschluss an den Bereich der ausgeprägten Übereinstimmungen mit der *Bambergensis* in einer einzigen Zeile verzeichnet ist. Eine etwas nähere Betrachtung soll in Kapitel 4.1 auf S. 259 erfolgen.

Abbildung 3.32 auf S. 233 soll neben den auch hier erkennbaren Entsprechungen zwischen bestimmten Textgruppen insbesondere verdeutlichen, welcher Informationsverlust mit der Projektion der Matches eines Textpaars auf eine einzige Zeile verbunden ist. Dazu sind zwei Projektionsdotplots zweier verwandter Texte in gleicher Skalierung nebeneinander gestellt. Schon aus der Breite lässt sich der unterschiedliche Textumfang erkennen. Während der linke Dotplot die Vermutung nahelegt, ein recht großer Teil der Coburger Hofgerichtsordnung von 1581 sei weitgehend zusammenhängend mit nur kürzeren Auslassungen oder Änderungen in die Coburger Hofgerichtsordnung von 1598 übernommen worden, ergibt sich aus der Darstellung rechts für dieses Textpaar eine viel weniger ins Auge fallende Linie, in der die geschlossenen schwarzen Abschnitte kürzer und durch größere Bereiche ohne Entsprechungen unterbrochen sind. Beide Texte weisen insbesondere zur Jenaer Hofgerichtsordnung von 1566 starke Übereinstimmungen auf. Während die entsprechende Zeile im linken Dotplot aber weitgehend

in denselben Bereichen geschwärzt ist wie in der Zeile, die die Matches mit der Coburger Hofgerichtsordnung von 1598 dokumentiert (wobei die Linie noch etwas stärker zusammenhängt als rechts), zeigt sich im rechten Dotplot deutlich, dass die Coburger Hofgerichtsordnung von 1598 wesentlich stärkere Ähnlichkeiten mit der Jenaer Ordnung von 1566 aufweist als mit der Coburger Ordnung von 1581. Auch die Übereinstimmungen mit den übrigen sächsischen Hofgerichtsordnungen (der Leipziger von 1529, der Leipziger von 1548 und der Wittenberger von 1550) lassen sich nach dem Bild beider Dotplots wohl gut mit einer über die Jenaer Ordnung laufende Traditionslinie erklären. Das schließt allerdings natürlich nicht aus, dass es auch direkte Übernahmen aus den älteren Texten gibt – aufgrund der Zusammenfassung der Matchpositionen zu Zeichenblöcken lässt sich hier ja nicht erkennen, wie umfassend die Übernahmen tatsächlich sind und ob sich an manchen Stellen vielleicht ausgeprägtere Übereinstimmungen mit anderen Texten dieser Familie feststellen lassen.

Daneben lassen sich für beide Coburger Ordnungen auch noch Matches mit einer Reihe weiterer Texte feststellen, wobei sich teilweise zusammenhängende Bereiche erkennen lassen, die auch an Positionen zu finden sind, für die insgesamt nur wenige Entsprechungen verzeichnet sind. Insbesondere betrifft das auch Passagen der Coburger Hofgerichtsordnung von 1598, zu denen sich keine Übereinstimmungen mit der Jenaer Hofgerichtsordnung feststellen lassen, so dass es nicht unwahrscheinlich erscheint, dass bei der Abfassung dieses Textes verschiedene Vorlagen herangezogen wurden. Eine Bestärkung dieser Vermutung ergibt sich bei der Überprüfung anhand von Dotplots für die einzelnen Textpaare. Darin zeigt sich, dass die Übereinstimmungen zum Beispiel mit der Braunschweig-Wolfenbütteler Hofgerichtsordnung von 1571 zwar überwiegend relativ kurz sind, dass sie sich aber zum großen Teil einer in etwa diagonalen Linie zuordnen lassen. Damit ist zwar nicht sichergestellt, dass die entsprechenden Stellen in der Coburger Hofgerichtsordnung von 1598 tatsächlich auf diesen Text oder auch einen mit ihm verwandten zurückzuführen sind, aber zumindest kann wohl auch ohne eine Betrachtung der einzelnen Stellen davon ausgegangen werden, dass es sich überwiegend um Formulierungen handelt, die inhaltlich signifikant und in einer Hofgerichtsordnung am ehesten in bestimmten Passagen zu erwarten sind.

Zwar springt es in einem Projektionsdotplot teilweise ins Auge, dass die Bereiche mit Entsprechungen in zwei Zeilen einander jedenfalls in etwa gleichen oder die darin verzeichneten Positionen in einer Teil- beziehungsweise Obermenge-Relation zueinander stehen, wenn diese Zeilen aber durch einen größeren Abstand getrennt sind und kein sehr deutliches Muster aufweisen, lassen sich solche Ähnlichkeiten rein visuell und ohne Hilfsmittel nur begrenzt erkennen – die vorgestellten Projektionsdotplots dürften dies hinreichend verdeutlichen. Schon aus diesem Grund legt es sich nahe, die Untersuchung der Übereinstimmungen zwischen Zeilen mit

einem automatischen Verfahren durchzuführen. Als Datenstruktur bieten sich Bitvektoren an, in denen jeder Bereich mit Matches durch ein gesetztes Bit an der entsprechenden Position verzeichnet wird. Damit lässt sich sehr effizient ermitteln, welche Übereinstimmungen und welche Abweichungen es gibt, und jedenfalls bei den hier untersuchten Textmengen ist es problemlos möglich, jede Zeile mit jeder anderen zu vergleichen.[610]

Da dabei jeweils drei Texte in Beziehung zueinander gesetzt werden (wobei die beiden durch Zeilen im Projektionsdotplot repräsentierten Texte allerdings nicht direkt, sondern nur im Hinblick auf ihre Entsprechungen zum Basistext des Dotplots verglichen werden), können hieraus Hinweise auf Abhängigkeitsverhältnisse über die Bildung von Textpaaren hinaus gewonnen werden. Allerdings ist die Interpretation nicht trivial. Insbesondere ist – jedenfalls bei den hier zugrunde gelegten Matchkriterien – je nach Grad der Formelhaftigkeit mit einem vielleicht nicht ganz geringen Anteil an zufälligen oder jedenfalls für die Ermittlung direkter Textbeziehungen nicht aussagekräftigen Übereinstimmungen zu rechnen. Dass also zum Beispiel eine Traditionslinie von einem Text über einen anderen zu einem dritten führt, wird nicht schon dadurch widerlegt, dass der dritte auch Übereinstimmungen mit dem ersten aufweist, die im zweiten keine Entsprechung haben. Wenn diese Übereinstimmungen allerdings umfangreicher und nicht durch andere Quellen zu erklären sind oder wenn sie im Zusammenhang der Matches mit dem ersten Text stehen, ist das wohl ein Grund für die Annahme, dass der dritte Text direkt auf dem ersten (oder einem anderen diesem nahe stehenden) basiert. Nur wenn es auch ähnlich signifikante Passagen gibt, die nur mit dem zweiten Text übereinstimmen, legt es sich nahe, von beiden Texten als Quellen für den dritten auszugehen oder aber von einem weiteren Text als Vorlage für alle drei.

Darüber hinaus spielt natürlich auch zum einen die zeitliche Beziehung eine Rolle, zum anderen die Tatsache, dass im Normalfall nicht damit zu rechnen ist, dass alle verwendeten Quellen der zu untersuchenden Texte auch selbst im ausgewerteten Korpus enthalten sind. Während sich unmittelbar ergibt, dass ein jüngerer Text nicht die Quelle eines älteren Textes sein kann, lässt sich aus der zeitlichen Abfolge keine positive Aussage über Abhängigkeitsverhältnisse ableiten. Vielmehr ist grundsätzlich davon auszugehen, dass Übereinstimmungen auch auf eine gemeinsame Quelle, eventuell vermittelt über andere dazwischen liegende Texte, zurückzuführen sein können.

[610] In *Perl* steht dafür das Modul *Bit::Vector* zur Verfügung, vgl. *http://search.cpan.org/~stbey/Bit-Vector-7.4/Vector.pod*.

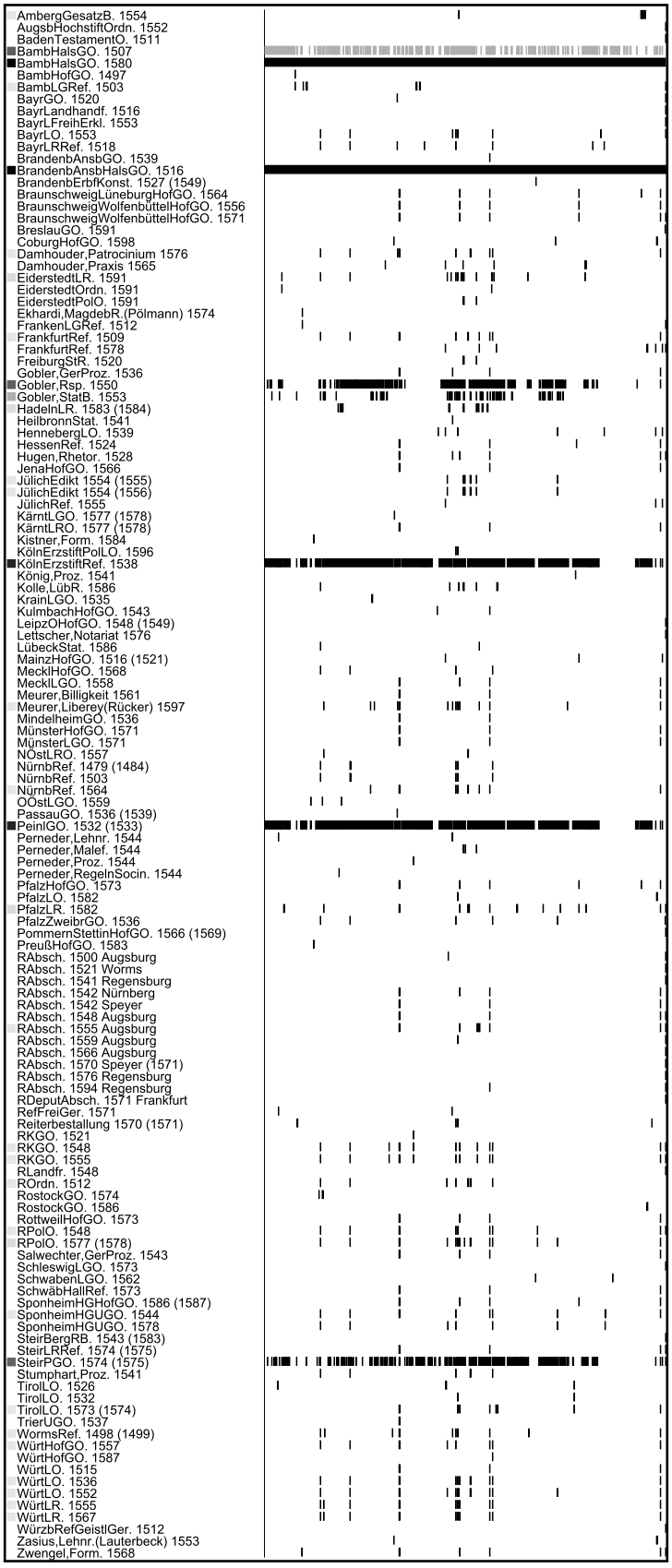


Abb. 3.29: Projektionsdotplot zu BambHalsGO. 1507 (vgl. S. 225 f.)

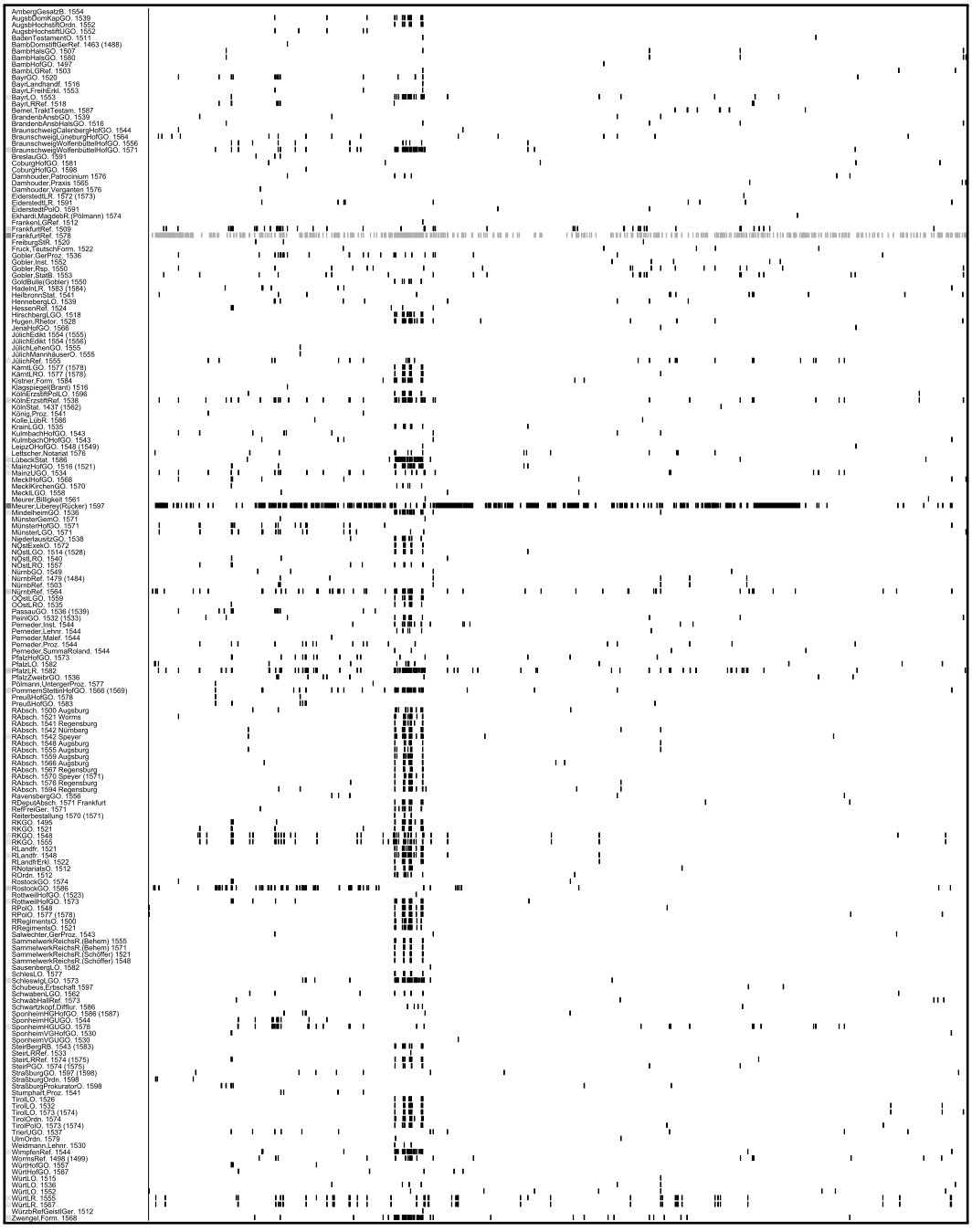


Abb. 3.30: Projektionsdotplot zu FrankfurtRef. 1578 (vgl. S. 226 f.)

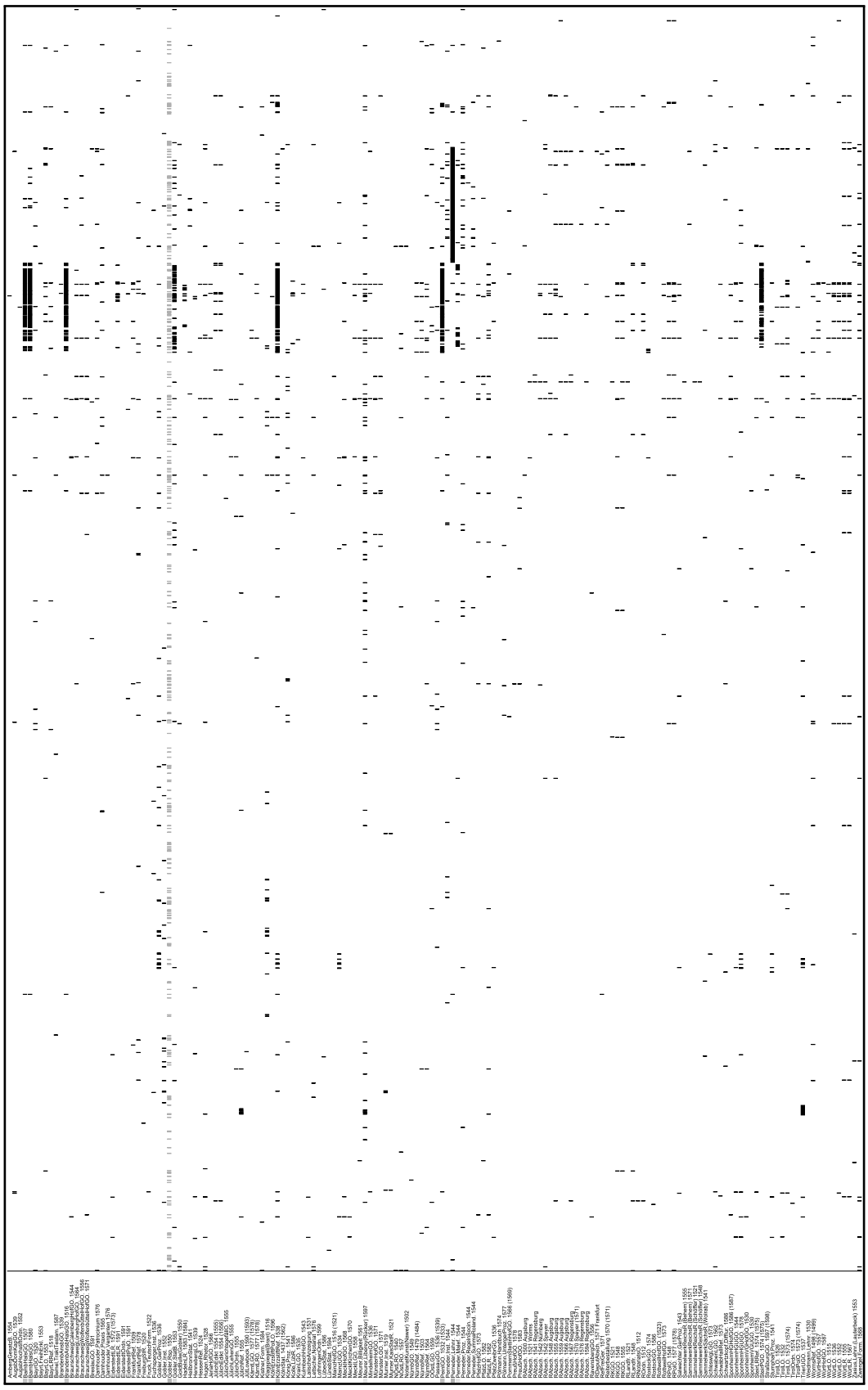
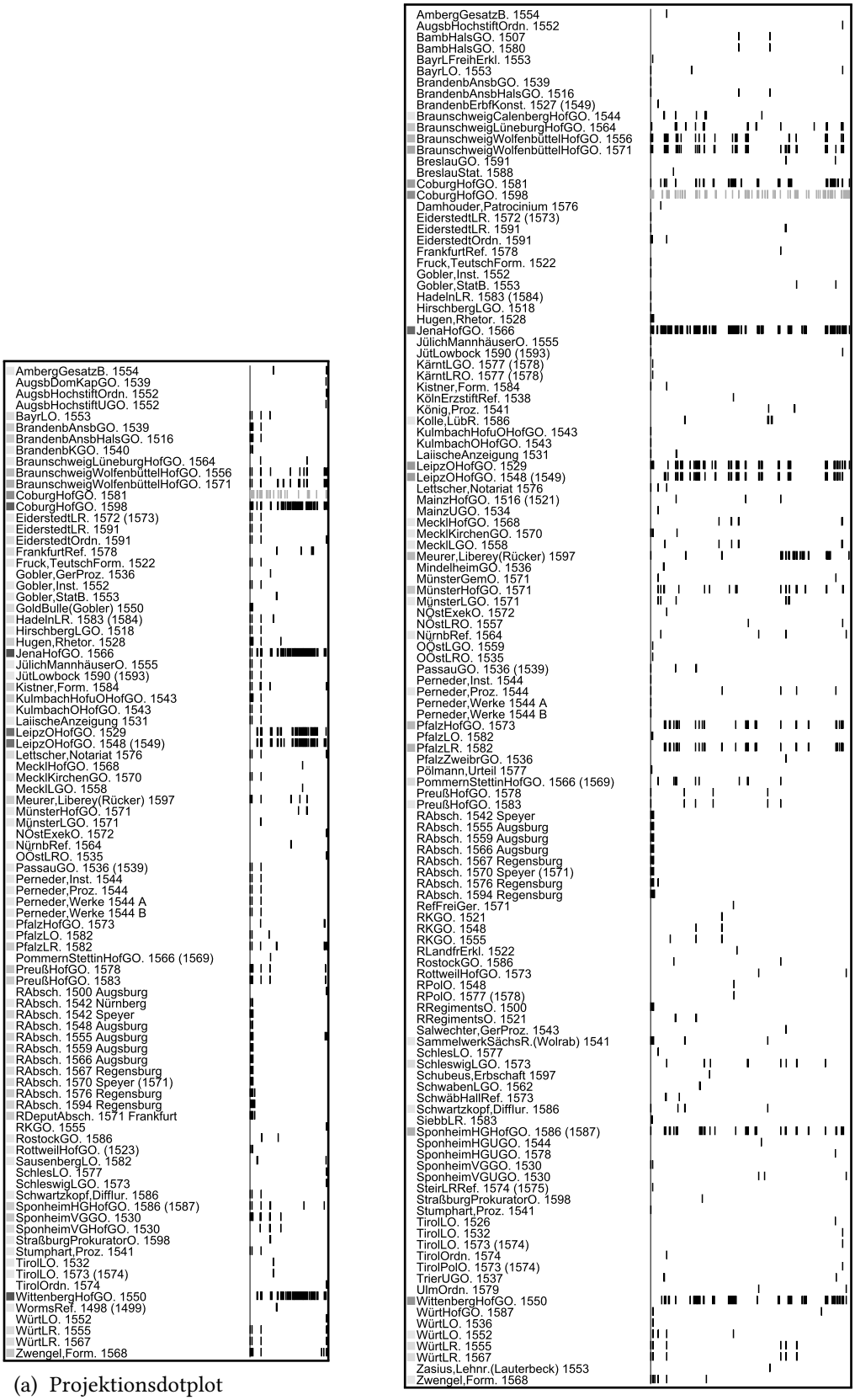


Abb. 3.31: Projektionsdotplot zu Gobler,Rsp. 1550 (vgl. S. 227)



(a) Projektionsdotplot zu CoburgHofGO. 1581

(b) Projektionsdotplot zu CoburgHofGO. 1598

Abb. 3.32: Projektionsdotplots zweier verwandter Texte (vgl. S. 227 f.)

3.4.5 Feinvergleich und Textparallelisierung

Neben der eben besprochenen zusammenfassenden Untersuchung von Übereinstimmungen zwischen Textpaaren und Textgruppen steht die Betrachtung der Textstücke, in denen sich solche Übereinstimmungen finden lassen. Dabei kann die Untersuchung durch unterschiedliche Fragestellungen und Ziele motiviert sein. Insbesondere kann es um einzelne Texte beziehungsweise Textpaare oder auch -gruppen gehen, oder bestimmte kürzere Stücke, die vielleicht eine weitere Verbreitung gefunden haben, stehen im Zentrum des Interesses.

Bei einer Rückbindung der ermittelten Matches an die originalen Stellen bieten sich wohl vor allem zwei Betrachtungsebenen an. Um festzustellen, worin die Übereinstimmungen genau bestehen und wo sich Unterschiede feststellen lassen – oder auch, wo Schreibungsvarianten im Umfeld der Matches durch die angewandte Varianzreduktion nicht vereinheitlicht wurden und ob es sich bei den ermittelten Matches tatsächlich um wörtliche Entsprechungen handelt –, ist es offenkundig erforderlich, eine Abbildung der ermittelten Matchpositionen auf die ihnen zugeordneten Stücke in den originalen Texten vorzunehmen. Daneben steht die genaue Verortung beziehungsweise Identifikation der betreffenden Passagen, die nicht nur für den Quellennachweis bei einer Präsentation von Textparallelen erforderlich ist, sondern auch Aussagen über Textstellen ermöglicht, ohne diese selbst zu zitieren. Eine entsprechende Auflistung von Fundstellen bietet sich an, wenn ein größerer Überblick gewonnen werden soll, als bei einer synoptischen Textdarstellung möglich ist, ohne dabei den Bezug zu den Textstellen gänzlich zu verlieren, wie dies etwa bei einer Dotplotvisualisierung der Fall ist.

Oben in Kapitel 3.1 wurde bereits dargestellt, dass eine Identifikation von Textstellen in XML-Dateien über *XPath*-Ausdrücke möglich ist und dass sowohl die Extraktion des eigentlichen Textes aus einer XML-Datei als auch die Reduktion von Textvarianz so erfolgen kann, dass eine Zuordnung zwischen den einzelnen einander entsprechenden Wörtern vor und nach der Transformation beziehungsweise zwischen *XPath*-Ausdrücken und zugehörigem Text protokolliert wird, so dass es nur geringen Aufwand verursacht, anschließend zu einem Stück der varianzreduzierten Textfassung den zugehörigen Originaltext beziehungsweise den entsprechenden *XPath*-Ausdruck zu ermitteln.

Wenn nicht einfach alle ermittelten und gegebenenfalls nach einfachen Filterkriterien ausgewählten und/oder gekürzten Matches ausgegeben werden sollen, sondern das Ziel eine möglichst gute Parallelisierung ganzer Texte oder größerer Abschnitte ist, ist außerdem zum einen eine weitgehend richtige Auswahl der einander zuzuordnenden Stellen aus der gerade bei kurzen Matchlängen und formelhaften Texten oft nicht unerheblich höheren Gesamtzahl von Matches wichtig, zum anderen die Prüfung auch der im Umfeld der ausgewählten Matches stehen-

den Passagen, die zwar keine exakte Entsprechung der geforderten Mindestlänge enthalten, wohl aber nicht selten kürzere oder durch Zwischentext unterbrochene übereinstimmende Formulierungen.

Deshalb soll in diesem Unterkapitel zunächst betrachtet werden, wie sich aus den ermittelten Matchpositionen eine solche Auswahl treffen lässt, dass sich daraus eine möglichst plausible 1:1-Zuordnung von Textstücken zueinander ergibt. Primär geht es dabei um Textpaare, bei denen sich in großen Teilen Entsprechungen finden und auch für Passagen ohne wörtliche Übereinstimmungen einigermaßen sicher entschieden werden kann, was sinnvollerweise in einer Synopse nebeneinander gezeigt oder als Einschub beziehungsweise Auslassung kenntlich gemacht werden sollte.

Im Idealfall handelt es sich dabei um Textpaare, die keine größeren Umstellungen aufweisen, sondern in denen die zu parallelisierenden Matches jeweils in der gleichen Reihenfolge zu finden sind. Auch unter dieser Annahme ist die Ermittlung einer optimalen Zuordnung nicht selbstverständlich. Insbesondere ist damit zu rechnen, dass eine Stelle in einem Text bei der Matchermittlung mehreren ganz unterschiedlich positionierten Stellen im anderen Text zugeordnet wurde oder dass die entsprechende Passage sogar in beiden Texten wiederholt vorkommt und sich daraus eine Vielzahl von Matches ergibt. Außerdem können sich auch Matches, deren Anfangspositionen in beiden Texten gleich geordnet sind, vor allem aufgrund von Stoppwörtern im Randbereich überlappen, so dass entschieden werden muss, zu welcher Übereinstimmung das doppelt zugeordnete Stück tatsächlich gehört.

Für das zuerst genannte Problem von Textstücken mit Entsprechungen an mehreren Stellen lässt sich eine recht gute Lösung finden, wenn eine Filterung der Matches nach der genannten Annahme erfolgt, dass keine Umstellungen vorliegen. Dann geht es im Kern darum, eine längste Sequenz von Matches beziehungsweise Matchzeichen zu ermitteln, die in beiden Texten stetig ansteigende Positionen aufweisen. Dies entspricht dem oben in Unterkapitel 2.1.2 beschriebenen Problem der längsten gemeinsamen Teilsequenz. Die dafür entwickelten Algorithmen lassen sich nutzen, wenn den Matches jeweils eine Identifikation (ID, zum Beispiel eine laufende Nummer) zugeordnet wird und die IDs in zwei unterschiedliche Sortierungen gebracht werden, die jeweils der Reihenfolge in einem der beiden Texte entsprechen. Um zu verhindern, dass zwei Matches in die ermittelte längste gemeinsame Teilsequenz aufgenommen werden, die in einem der Texte an derselben Position anfangen, kann außerdem als sekundäres Sortierkriterium die Position im jeweils anderen Text in absteigender Reihenfolge berücksichtigt werden. Das schließt allerdings nicht aus, dass Matches in die Sequenz übernommen werden, die einander in einem Text weitgehend überlappen, aber geringfügig unterschiedliche Anfangspositionen haben; wenn die Positionen im anderen Text aber weit voneinander entfernt sind,

bestehen gute Chancen, dass nur eines von ihnen in die Sequenz übernommen wird, weil sich andere Matches besser in die Sequenz einfügen.[611]

Da die Matches eine sehr unterschiedliche Länge haben können, legt es sich nahe, diese Länge auch bei der Ermittlung einer längsten gemeinsamen Teilsequenz zu berücksichtigen. Dies kann dadurch geschehen, dass die IDs in den beiden nach der Textreihenfolge sortierten Gesamtsequenzen entsprechend der Zeichenzahl, eventuell modifiziert zum Beispiel durch Berücksichtigung eines Divisors, wiederholt verzeichnet werden.[612]

Allerdings ist nochmals darauf hinzuweisen, dass es im Regelfall eine Vielzahl von längsten gemeinsamen Teilsequenzen gibt, so dass keineswegs einfach angenommen werden kann, dass die mithilfe eines bestimmten Algorithmus ermittelte Teilsequenz tatsächlich die beste Zuordnung darstellt – ganz abgesehen davon, dass es zum Beispiel durchaus auch der Fall sein kann, dass mehrere auf mehr oder weniger zufälliger Übereinstimmung beruhende Matches zusammen eine größere Länge haben als ein inhaltlich wesentlich aussagekräftigeres Match, das nicht in die unter Einbeziehung dieser Matches gebildete Teilsequenz passt. Insbesondere bei stark von formelhaften Wendungen bestimmten Texten ist damit zu rechnen, dass eine Matchauswahl, die allein auf dem Kriterium einer möglichst großen Gesamtlänge beruht, in die Irre führt.

Eine bessere Annäherung an eine optimale Textstellenzuordnung dürfte in vielen Fällen dadurch zu erreichen sein, dass zunächst einmal nicht sämtliche Matches für die Ermittlung einer längsten gemeinsamen Teilsequenz berücksichtigt werden, sondern nur solche, bei denen die Wahrscheinlichkeit hoch ist, dass sie tatsächlich auf einer textuellen Beziehung zwischen den zugehörigen Textstellen basieren. In einem zweiten Schritt können dann die zuvor ausgeschlossenen Matches darauf geprüft werden, ob sie sich in die ermittelte Teilsequenz einfügen lassen. Auch bei diesem zweiten Schritt kann es allerdings von Vorteil sein, nicht alle zunächst zurückgestellten Übereinstimmungen einzubeziehen, sondern insbesondere die, die auf Formulierungen mit einer hohen Vorkommenshäufigkeit beruhen, auszuklammern. Eine Berücksichtigung solcher Passagen mit einer Vielzahl von entsprechenden Stellen ist insbesondere dann plausibel, wenn sie sich im näheren Umfeld von Textstücken befinden, die schon als einander mit hoher Wahrscheinlichkeit zuzuordnen

[611] Dies ist natürlich abhängig von der Zahl und Verteilung der übrigen Matches.

[612] Oben auf S. 63 wurde in Anm. 247 auf den mit der Zeichenkettenlänge stark ansteigenden Zeitaufwand für die Ermittlung einer längsten gemeinsamen Teilsequenz hingewiesen. Das gilt zwar prinzipiell auch in diesem Fall, allerdings lässt sich die Ermittlung wesentlich schneller durchführen, wenn – wie bei dem hier vorausgesetzten ID-System – jedes zu untersuchende potentielle Sequenzglied zwar gegebenenfalls mehrfach nacheinander, aber darüber hinaus an keiner weiteren Stelle vorkommt. In Testläufen wurden zum Beispiel für die Ermittlung einer längsten gemeinsamen Teilsequenz mit 38.422 Gliedern in zwei Sequenzen der Länge 48.532 ca. 34 Sekunden benötigt. Für eine längste gemeinsame Teilsequenz mit 76.528 Gliedern in zwei Sequenzen der Länge 96.361 waren ca. 150 Sekunden und für eine längste gemeinsame Teilsequenz mit 7.033 Gliedern in zwei Sequenzen der Länge 107.570 ca. 20 Sekunden erforderlich.

erkannt wurden. Da es aber ohnehin sinnvoll ist, von den im Feinvergleich berücksichtigten Matches ausgehend den jeweiligen Kontext auf Übereinstimmungen zu prüfen, die bei der Matcherkennung aufgrund zu geringer Länge nicht erkannt werden konnten, können innerhalb der dabei untersuchten Textbereiche auch zuvor übergangene Matches als Entsprechungen erkannt werden. Die Kontextprüfung wird unten ab S. 238 noch näher beschrieben.

Oben in Unterkapitel 3.3.2 wurden bereits verschiedene Kriterien benannt und verschiedene Bewertungsformeln dargestellt, die für eine Einschätzung der Aussagekraft eines Matchs im Hinblick auf textuelle Beziehungen genutzt werden können. Wenn es – wie eben beschrieben – nur darum geht, zunächst eine Vorauswahl zu treffen, um die übrigen Matches anschließend auf dieser Basis erneut zu prüfen, kann wohl relativ streng gefiltert werden. So legt es sich insbesondere nahe, Textstücke, die zwei oder mehreren Stellen im anderen Text zugeordnet sind, zunächst auszuklammern. Eine Überlappung im Randbereich sollte allerdings wohl nicht als Ausschlusskriterium gelten, da dies ein häufiger Fall ist, wenn die Matchermittlung ohne die Berücksichtigung syntaktischer Grenzen erfolgt. Vielmehr bietet es sich an, nach der Ermittlung einer möglichst langen Teilsequenz von Matches, die auf im Kernbereich jeweils singulären Passagen beruhen, eine Kürzung um Randwörter vorzunehmen, die in einem der Texte zugleich den Abschluss eines Matchs und den Beginn des folgenden darstellen.

Das bisher beschriebene Verfahren beruht auf der Annahme des Idealfalls, dass die einander tatsächlich entsprechenden Passagen in beiden Texten in der gleichen Reihenfolge zu finden sind. Das trifft natürlich in einer Vielzahl von Fällen nicht oder nur teilweise zu. Soweit das Grundmodell einer synoptischen Textdarstellung zweier Texte aber überhaupt adäquat ist, kann man anhand einer längsten gemeinsamen Teilsequenz von als signifikant eingestuften Matches, die vielleicht noch einer gewissen Optimierung unterzogen wird,[613] eine Grundzuordnung vornehmen und die übrigen Matches, die nicht als zufällige Übereinstimmungen eingestuft werden, als Umstellungen betrachten und entsprechend präsentieren.

Eine umfassende Parallelisierung beziehungsweise die Wertung von Übereinstimmungen als Umstellungen ist aber nur für einen Teil der Textpaare adäquat, für die sich textuelle Beziehungen (unter Umständen über Zwischenglieder) feststellen lassen. Als anderes Grundmodell lässt sich eine punktuelle Zuordnung beschreiben, die sich zwar natürlich auch synoptisch darstellen lässt, aber jedenfalls nicht das Ziel hat, zwei Texte möglichst *in extenso* in Beziehung zueinander zu setzen, sondern vielmehr zu einzelnen Passagen eines Haupttextes oder auch nur zu einer

[613] So lassen sich nicht zur ermittelten Teilsequenz gehörige Matches zum Beispiel darauf prüfen, ob sie anstelle eines ausgewählten Matchs ebenfalls in die Teilsequenz passen würden und aufgrund einer größeren Nähe zu einem der in der Teilsequenz benachbarten Matches vermutlich vorzuziehen sind. Daneben sind natürlich auch kompliziertere Fälle denkbar, etwa dass der Austausch mehrerer Matches *en bloc* zu einer Verringerung der Matchabstände führt.

einzigen Stelle Übereinstimmungen in einer oder auch mehreren anderen Texten aufführt.[614] Zwar haben Texte wie die hier untersuchten nicht selten eine durch den Sachzusammenhang beziehungsweise eine etablierte Struktur bedingte Ordnung, die dazu führen kann, dass auch in einem solchen Fall Matches in zwei miteinander verglichenen Texten überwiegend oder sogar gänzlich in der gleichen Reihenfolge auftreten, aber grundsätzlich ist hier doch mit einem höheren Anteil an signifikanten Matches zu rechnen, die sich nicht in eine längste gemeinsame Teilsequenz einordnen lassen.

Mit dem in dieser Arbeit vorgestellte Verfahren einer Stellenzuordnung auf der Basis von MEMs einer bestimmten Mindestlänge kann natürlich in einer Vielzahl von Fällen nur ein Teilbereich der tatsächlichen wörtlichen Übernahmen erkannt werden, da sich oft durch kleinere Formulierungs- oder Schreibungsänderungen auch in den stark varianzreduzierten codierten Textfassungen Abweichungen in der Zeichenfolge ergeben. Dementsprechend ist es sinnvoll, das Umfeld der ermittelten Matches darauf zu überprüfen, ob sich darin ebenfalls übereinstimmende Wörter finden.

Dabei stellt sich zunächst einmal die Frage, welche Grenzen für eine solche Prüfung gezogen werden sollen. Soweit zwei Matches in beiden Texten in gleicher Reihenfolge mit wenig Abstand aufeinander folgen, liegt es recht nahe, einfach die Zwischenstücke zu vergleichen. Bei größerem Abstand lässt sich wohl kaum sicher feststellen, bis zu welchem Punkt ein Vergleich sinnvoll ist, ohne ihn tatsächlich durchzuführen, da aufgrund von Auslassungen beziehungsweise Einschüben prinzipiell auch mit größeren Bereichen ohne Entsprechung zu rechnen ist. Ein vollständiger Vergleich, wie er im Folgenden noch beschrieben wird, kann aber für längere Textstücke unverhältnismäßig aufwendig sein.[615] Wenn jedoch berücksichtigt wird, dass die primäre Matcherkennung schon geleistet ist und es nur darum geht, gegebenenfalls auch im unmittelbaren Umfeld eine Zuordnung von Entsprechungen vorzunehmen, lässt sich eine entsprechende Abgrenzung mit einem relativ einfachen heuristischen Ansatz vornehmen.

Eine Möglichkeit dafür soll zunächst für Textstücke beschrieben werden, die auf ein Match folgen. Das Grundschema besteht aus folgenden Schritten, wobei vorausgesetzt wird, dass der dabei verwendete Algorithmus zur Erkennung einer längsten gemeinsamen Teilsequenz jeweils Elemente mit niedrigerer Ordnungsnummer vorzieht, wenn es mehrere Auswahlmöglichkeiten gibt[616]:

[614] Ein Beispiel für eine gedruckte Edition, die einen Haupttext und daneben die identifizierten Textpassagen aus unterschiedlichen Vorlagen zeigt, ist – aus einem anderen Fachgebiet – Vogl 2007. Während sich eine solche Präsentation in einer digitalen Edition natürlich auch für den Vergleich eines Haupttextes mit mehreren weiteren in vielen Fällen gut umsetzen lässt, ist dies in einer gedruckten Edition aber durch den vergleichsweise starren Seitenaufbau wesentlich schwieriger und dementsprechend eher die Verzeichnung in den Anmerkungen oder im Kommentar typisch.

[615] Oben S. 63, Anm. 247 sind Messwerte für die Ermittlung von längsten gemeinsamen Teilsequenzen genannt.

1. Setze in beiden Texten die Leseposition auf das Ende des Matchbereichs.
2. Lies jeweils eine bestimmte (nicht zu kleine, aber auch nicht zu große) Anzahl von Zeichen oder auch Wörtern ein (aber maximal bis zum folgenden Match).
3. Ermittle in den darin enthaltenen Wörtern (unter Ausschluss von Stoppwörtern oder anderweitig als wenig signifikant eingestuften Wörtern) eine längste gemeinsame Teilsequenz.
4. Prüfe, ob die ermittelte Teilsequenz ab einem bestimmten Punkt größere Lücken enthält, ohne dass die noch folgenden Sequenzelemente so auffällig wären, dass trotzdem eine Zuordnung erfolgen sollte. Wenn das der Fall ist, dann kürze die Sequenz entsprechend.
5. Wenn die Teilsequenz auch nach einer solchen Kürzung Elemente enthält, dann setze die Leseposition jeweils auf den mit dem letzten Glied der Teilsequenz erreichten Endpunkt in den Texten und prüfe die folgenden Textstücke erneut (wiederhole das Verfahren ab Schritt 2). Wenn nichts gefunden wird, dann brich das Verfahren ab und setze als Bereichsgrenze jeweils die Position an, an der die in diesem Durchgang verglichenen Textstücke beginnen.

Für den Vergleich von Textstücken, die einem Match vorangehen, ergeben sich natürlich die entsprechenden Modifikationen. Insbesondere ist dabei wichtig, dass die für die Teilsequenz ausgewählten Wörter möglichst nahe am Ende stehen. Das lässt sich dadurch erreichen, dass die Wort- oder Zeichenreihenfolge für die Ermittlung der längsten gemeinsamen Teilsequenz umgedreht wird; dann muss für die Ermittlung der tatsächlichen Positionen im Text anschließend wieder eine Umrechnung erfolgen. Und um Überlappungen mit einem zuvor für das vorangehende Match ermittelten Folgebereich zu vermeiden, sind gegebenenfalls die dabei erreichten Positionen als frühestmögliche Anfangspunkte zu berücksichtigen.

Die eben gewählte Beschreibung der einzelnen Schritte ist in einigen Punkten gezielt vage gehalten, da sie sich wohl kaum allgemeingültig festlegen lassen, sondern von den Rahmenbedingungen abhängen, insbesondere, welche Merkmale genutzt werden können, um eine Wortform als signifikant zu werten und in welchem Maße Lücken in der Zuordnung toleriert werden sollen. Die Begrenzung eines Vergleichsdurchlaufs auf eine bestimmte Zeichen- oder Wortzahl kann man in Beziehung zu der Toleranz gegenüber Unterbrechungen sehen: Während die in Schritt 4 genannte Lücke dann keine Rolle spielt, wenn danach noch als wichtig eingestufte Entsprechungen gefunden werden, wird durch Schritt 2 die tatsächliche Obergrenze für eine Lücke festgesetzt.[617]

[616] Das *Perl*-Modul *Algorithm::Diff* (vgl. *http://search.cpan.org/~tyemq/Algorithm-Diff-1.1902/lib/Algorithm/Diff.pm*) arbeitet mit einem solchen Auswahlprinzip.

[617] Für die im Folgenden noch gezeigten Textvergleiche werden in jedem Abgrenzungsdurchlauf 300 Zeichen berücksichtigt und dabei Wörter mit weniger als drei Zeichen in der codierten Form ausgeschlossen. Die Teilsequenz wird gekürzt, wenn in beiden Texten zusammen mehr als dreißig Wörter (ohne Zählung der ausgeschlossenen Wörter) dazwischen liegen, es sei denn,

Wenn die Bereiche ermittelt sind, die näher miteinander verglichen werden sollen, können für diesen Vergleich verschiedene Verfahren eingesetzt werden. Insbesondere liegt es dabei wohl nahe, über eine längste gemeinsame Teilsequenz in einem einzigen Durchgang eine möglichst umfassende Zuordnung vorzunehmen. Allerdings ergeben sich dabei verschiedene Probleme. Zunächst einmal stellt sich die Frage, ob der Vergleich auf Wort- oder Zeichenebene erfolgen soll. Einerseits können bei einer Teilsequenz von Einzelzeichen auch Wortformen in Beziehung zueinander gesetzt werden, die zwar zu erheblichen Teilen, aber doch nicht völlig übereinstimmen. Andererseits ist – insbesondere bei einer Codierung mit einer geringen Zahl unterschiedlicher Zeichen – damit zu rechnen, dass die ermittelte längste Teilsequenz zu einem erheblichen Teil aus Einzelzeichen besteht, die in beiden untersuchten Textstücken zufällig in der gleichen Reihenfolge (aber mit Lücken) zu finden sind. Bei einem Vergleich auf Wortebene werden zwar nur komplette Wortformen einander zugeordnet, wenn dabei aber auch Stoppwörter beziehungsweise wenig signifikante Formen einbezogen werden, ist nicht unbedingt gesagt, dass die ermittelte längste gemeinsame Teilsequenz tatsächlich gerade die bedeutungstragenden Entsprechungen enthält. Zudem ist zu berücksichtigen, dass jedenfalls der hier verwendete Algorithmus nicht primär zusammenhängende Elementfolgen auswählt, sondern vielmehr solche Elemente, die jeweils möglichst nahe am Anfang stehen.

Ein anderer Ansatz besteht darin, nach dem gleichen Prinzip wie bei der grundlegenden Matcherkennung nach möglichst langen, in diesem Fall aber die im vorherigen Erkennungsschritt angesetzte Mindestlänge unterschreitenden Übereinstimmungen zu suchen. Das hat gegenüber der Ermittlung einer längsten gemeinsamen Teilsequenz den Vorteil, dass gegebenenfalls übereinstimmende zusammenhängende Wortfolgen nicht auseinandergerissen werden. Ebenso wie bei der Erkennung längerer Matches werden natürlich auch in diesem Fall zunächst einmal nur Zeichenfolgen gefunden, die nicht unbedingt vollständigen Wörtern entsprechen, so dass für die Erkennung gänzlich übereinstimmender Wörter noch eine entsprechende Anpassung der Abgrenzung erfolgen muss. Auch wenn sich dadurch die Matchlänge noch einmal verkürzt, ist die Wahrscheinlichkeit im Umfeld längerer Matches immer noch recht hoch, dass es sich tatsächlich um eine plausible Zuordnung handelt. Insbesondere dürfte sie für darin enthaltene Stoppwörter wesentlich höher sein, als dies bei einer Zuordnung auf der Basis einer längsten gemeinsamen Teilsequenz der Fall ist. Ein offensichtlicher Nachteil einer Zuordnung nur von Matches mit einer bestimmten Mindestlänge liegt darin, dass auch vergleichsweise seltene Einzelwörter oder kurze Wortfolgen, die diese Mindestlänge nicht erreichen, nicht erkannt werden können.

---

darauf folgt noch ein zusammenhängendes Stück von mehr als fünf übereinstimmenden Zeichen (wiederum ohne Berücksichtigung dazwischen liegender ausgeschlossener Wörter).

Die vorgestellten Ansätze lassen sich natürlich auch kombinieren. Ausgehend von der Annahme, dass in aller Regel die Wahrscheinlichkeit einer korrekten Zuordnung mit zunehmender Länge ansteigt, liegt den hier in Teil 4 präsentierten Feinvergleichsergebnissen folgende Abfolge zugrunde, wobei die Schritte 1–3 die Texte in der codierten Form vergleichen:

1. Ermittle alle Übereinstimmungen mit einer Mindestlänge von zehn Zeichen, passe sie an die Wortgrenzen an und bilde daraus eine möglichst lange Sequenz, so dass sich die Matches nicht überlappen und in beiden Texten in der gleichen Reihenfolge zu finden sind.
2. Ermittle in jedem Stück zwischen den Matches von Schritt 1 eine längste gemeinsame Teilsequenz von Wörtern mit einer Mindestlänge von drei Zeichen, wobei Wörter, die im Zwischenstück wiederholt vorkommen, nicht berücksichtigt werden.[618]
3. Ermittle in jedem Stück zwischen den Matches von Schritt 2 eine längste gemeinsame Teilsequenz auf Zeichenebene und berücksichtige alle sich dabei ergebenden Wortzuordnungen, wenn die beiden Wortformen vollständig miteinander übereinstimmen oder wenn mindestens drei Zeichen und im Durchschnitt mindestens 75 % der Zeichen beider Wortformen einander zugeordnet wurden.
4. Ermittle in jedem Stück zwischen den in Schritt 3 verzeichneten Wort-Matches und -Ähnlichkeiten eine längste gemeinsame Teilsequenz der in Kleinschreibung umgewandelten Buchstaben der originalen Textfassungen und prüfe, ob sich dadurch Wortzuordnungen ergeben, bei denen die Formen gleich sind oder bei denen eine Form mit einem Punkt endet und mit dem Anfang der anderen Form übereinstimmt, so dass sie als Abkürzung interpretiert werden kann. Prüfe nach diesem Verfahren auch für die in Schritt 3 als ähnlich erkannten Wortformen, ob sich ihre Unterschiede in dieser Weise erklären lassen.
5. Führe eine Klassifikation der einander zugeordneten und der übrigen Textstücke nach folgendem Schema auf der Basis der codierten Formen durch:
   - Völlige Übereinstimmung: gleich
   - Erfolg von Schritt 4: Abkürzung
   - kurze Stücke mit einer Levenshtein-Distanz von maximal einem Drittel der Länge des kürzeren der beiden Textstücke: ähnlich[619]
   - Übereinstimmung bis auf ein zusätzliches Zeichen am Ende eines der Textstücke: ähnlich[620]

[618] Diese Zusatzbedingung soll die Wahrscheinlichkeit einer falschen Zuordnung minimieren und damit die einer guten Alinierung im nächsten Schritt verbessern.

[619] Die Levenshtein-Distanz wird oben auf S. 59 kurz beschrieben. Da aufgrund der Einzelschritte des Feinvergleichs mit einer Zuordnung weitgehend auf Wortebene gerechnet werden kann, ist die Berechnung der Levenshtein-Distanz für längere Textstücke überflüssig und würde nur überflüssigen Rechenaufwand verursachen.

[620] Damit sollen Formen, die aufgrund zu geringer Länge nicht von der vorherigen Regel abgedeckt werden, sich aber mit einer gewissen Wahrscheinlichkeit nur aufgrund einer Flexionsendung unterscheiden, als weitgehende Entsprechung erkannt werden

- Textstück ohne Entsprechung: Zusatz[621]
- sonst: unterschiedlich

Die Berücksichtigung von Abkürzungen in Schritt 4 stellt eine Optimierung dar, die vor allem für die Autorenwerke des hier untersuchten Korpus von Interesse ist, weil viele dieser Texte Verweise insbesondere auf das *Corpus Iuris Civilis* und die zugehörige Literatur enthalten und dabei die Stellenangaben für das *Corpus Iuris Civilis* nach einem etablierten Schema über die mit einer gewissen Variabilität abgekürzten Anfangswörter der betreffenden Abschnitte erfolgen.[622] Diese Optimierung ist hier nur rudimentär entwickelt und könnte durch zusätzliche Vereinheitlichungen wie etwa zwischen Zahlen in römischen und in arabischen Ziffern verbessert werden. Primär soll sie hier als Beispiel dafür dienen, dass sich manche Entsprechungen mit einem Matching auf der Basis einer Codierung nach lautlichen Ähnlichkeiten natürlich nicht oder nicht sicher feststellen lassen, dass sich aber das hier vorgestellte Zuordnungsverfahren mit einem zusätzlichen Vergleich auf der Basis von dafür passenden Sonderregeln kombinieren lässt.

Auch für Textstücke ohne Abkürzungen kann sich Schritt 4 auswirken, da hier weitgehend die originalen Zeichenfolgen zugrunde gelegt werden und sich deshalb teilweise andere Zuordnungen ergeben als bei der Ermittlung einer längsten gemeinsamen Teilsequenz auf Basis der codierten Textfassungen. Insbesondere erhalten dabei Wortformen ein höheres Gewicht, bei denen die Codierung zu einer überdurchschnittlich hohen Reduzierung der Zeichenzahl führt.

Generell ist zu betonen, dass die genannten Schritte keineswegs eine auch nur annähernd optimale und vollständige Lösung für das Problem einer Erkennung von Textentsprechungen in relativ kurzen Passagen sein sollen. Dies ist eine durchaus eigenständige Aufgabe, die erhebliche Schwierigkeiten aufweist, insbesondere zum einen aufgrund der Möglichkeit von Umstellungen, die hier völlig ausgeklammert wurden,[623] zum anderen jedenfalls für das hier untersuchte Korpus aufgrund der Schreibvarianz, die sich nicht durch allgemeingültige Regeln auffangen lässt, ohne zugleich einen erheblichen Verlust an *Precision* in Kauf zu nehmen.[624] Diese Auf-

[621] Diese Klassifikation ist in vielen Fällen wenig aussagekräftig oder sogar irreführend, weil sie sich einfach daraus ergibt, dass nur in einem der Texte zwischen zwei einander zugeordneten Textstücken noch etwas steht. Wenn eine ermittelte Entsprechung zum Beispiel nur auf einer zufällig gleichen Codeform beruht, lässt sich daraus natürlich nicht folgern, dass ein Zwischenstück durch eine Ergänzung oder Streichung in einem der Texte zu erklären ist. Außerdem werden tatsächlich als Ergänzungen oder Streichungen zu betrachtende Textänderungen dann nicht als solche erkannt, wenn zwei Entsprechungen in beiden Texten nicht unmittelbar aufeinander folgen (auch dann nicht, wenn zum Beispiel das Zwischenstück in einem der Texte das Ende des vorangehenden Abschnitts darstellt und im anderen Text den Anfang des nächsten). Bei Passagen, die einander weitgehend entsprechen, liefert die Klassifikation aber recht gute Ergebnisse.

[622] Vgl. Kantorowicz 1933/1987, insbesondere S. 72 f. und 75.

[623] Oben auf S. 104 wurde schon kurz erwähnt, dass die Ermittlung eines optimalen Editierskripts unter Berücksichtigung von Umstellungen ein NP-vollständiges Problem darstellt, und die dem Programm *MEDITE* zugrunde liegende heuristische Lösung des Problems kurz vorgestellt.

gabe fällt in den Bereich der automatischen Kollationierung.[625] Im Rahmen der vorliegenden Untersuchung soll nur ein rudimentäres Verfahren gezeigt werden, mit dem sich trotz der unvermeidlichen Schwächen einer Matcherkennung auf der Basis exakter Übereinstimmung eine weitgehend plausible Zuordnung auch für Textstücke erreichen lässt, die nicht die für die primäre Erkennung erforderliche Mindestlänge haben.

Bei der Prüfung eines Feinvergleichs auf der Basis der Codierung *ohne A/I/U/B/F* ohne Leerzeichen kann man allerdings feststellen, dass diese radikale Form der Varianzreduktion, die sich für die Ermittlung längerer Übereinstimmungen im hier untersuchten Korpus gut bewährt hat, für eine Zuordnung auch sehr kurzer Zeichenfolgen problematisch ist. Recht offensichtlich zeigt sich das darin, dass in dieser Codierung zum Beispiel die Präfixe „auf“ oder „ab“ einfach gelöscht werden. Und in nicht wenigen Fällen kommt es zur Abbildung ganz unterschiedlicher Wörter auf dieselbe Codeform. Insbesondere wenn bei der Ermittlung einer längsten gemeinsamen Teilsequenz nur einzelne Wörter gefunden werden, die voneinander durch deutlichen Abstand getrennt sind, ist die Wahrscheinlichkeit nicht gering, dass es sich um eine falsche Zuordnung handelt.

Das Problem lässt sich immerhin reduzieren, wenn eine Codierung zugrunde gelegt wird, die stärker differenziert. So kann zum Beispiel die Codierung *ohne A/I/U* ohne Leerzeichen verwendet werden, bei der *b* und *p* durch ein *B* repräsentiert werden und *f* durch ein *F*. Daraus ergibt sich zwar, dass einige Wortübereinstimmungen nicht erkannt werden können – insbesondere bei Wörtern, die den Laut */f/* enthalten, da diesem in der Schreibung ein *f*, *v* oder auch ein *u* entsprechen kann[626] –, aber da der Feinvergleich nach dem vorgestellten Verfahren auch dann Wortformen einander zuordnen kann, wenn diese sich geringfügig unterscheiden, ergibt sich daraus nicht unbedingt eine Verringerung des *Recalls*.[627]

[624] Das in dieser Untersuchung genutzte Codierungssystem ist eigentlich nicht für eine Zuordnung von – oft sogar nur kurzen – Einzelwörtern gedacht, sondern vielmehr für die Ermittlung längerer Übereinstimmungen, bei denen die Wahrscheinlichkeit einer rein zufälligen Gleichheit zu vernachlässigen ist. Vgl. oben S. 125 ff.

[625] Vgl. oben Kapitel 2.6.

[626] Dass in der Codierung *ohne A/I/U/B/F* nicht nur *f*, sondern auch *b* und *p* einfach gestrichen werden, basiert (neben den Abgrenzungsproblemen zwischen Fortes und Lenes) auf der Überlegung, dass ein *f* beziehungsweise *v* oder *u* auch einem *b* der Hochlautung entsprechen kann. Dies betrifft aber nur dialektal geprägte Schreibungen aus bestimmten Sprachräumen und spielt für den Großteil der hier untersuchten Texte keine Rolle.

[627] Eine solche Zuordnung von Formen als ähnlich kann natürlich auch zum Beispiel für Wörter mit den eben beschriebenen Unterschieden hinsichtlich eines Präfixes erfolgen, ist in diesem Fall aber wohl auch als sachlich angemessen zu betrachten. Jedenfalls ergeben sich bei einer Verlängerung der Codeformen der Wörter gegebenenfalls auch längere Übereinstimmungen auf Zeichenebene und damit auch eine höhere Wahrscheinlichkeit, dass die betreffenden Zeichenfolgen in die ausgewählte längste gemeinsame Teilsequenz aufgenommen werden, sowie

Allerdings kann es natürlich der Fall sein, dass eine wörtliche Übereinstimmung aufgrund eines solchen Schreibungsunterschiedes in der primären Matcherkennung nicht die angesetzte Mindestlänge erreicht und deshalb gar nicht als solche erkannt wird – wenn sie dann nicht im Umfeld eines Matchs liegt und über die Kontextprüfung im Feinvergleich gefunden wird, geht diese Zuordnung bei einer stärker differenzierenden Codierung also verloren. Um das zu vermeiden, kann man aber in den verschiedenen Erkennungsschritten mit unterschiedlichen Codierungen arbeiten und die in einer stark vereinheitlichenden Codierung wie *ohne A/I/U/B/F* ohne Leerzeichen ermittelten Matchpositionen (beziehungweise ihre Anfangs- oder Endpunkte) als Anker verwenden, um darauf aufsetzend die eigentlichen Zuordnungen über den Feinvergleich vorzunehmen. Dabei müssen die Matchpositionen, die sich auf die Codierung für die primäre Matchermittlung beziehen, natürlich auf die für den Feinvergleich abgebildet werden, es müssen also – nach der hier angewandten Positionszuordnung – zunächst die den Matches entsprechenden Positionen im Originaltext und dazu dann die Positionen in der stärker differenzierenden Codierung festgestellt werden.[628]

Eine Einschätzung, wie viel sich mit den vorgestellten Schritten erreichen lässt, ist schwierig, da sehr viel vom konkreten Einzelfall abhängt. So kann die Erkennung von Textentsprechungen bei einem Textpaar, in dem sich die Wörter schon über die Ermittlung etwas längerer Matches zum allergrößten Teil einander zuordnen lassen, über einen Feinvergleich natürlich nur vergleichsweise geringfügig verbessert werden. Bei völlig isoliert stehenden Matches hingegen könnten zwar theoretisch auch in einem recht großen Umfeld kürzere Entsprechungen zu finden sein, es ist aber wohl wahrscheinlicher, dass die isolierten Matches auf mehr oder weniger zufälligen Übereinstimmungen beruhen, jedenfalls wenn sie nicht aufgrund erheblicher Länge oder der Verwendung seltener Wörter als vermutlich signifikant einzustufen sind. Dementsprechend mag es zwar in ihrem unmittelbaren Kontext teilweise gleiche oder ähnliche Einzelwörter oder auch kurze Wortfolgen geben, die etwa durch den Formulierungszusammenhang bedingt sind, umfangreiche Passagen, die einander tatsächlich im Wesentlichen entsprechen, dürften aber relativ selten zu finden sein. Ein stärkerer Effekt lässt sich von einem Feinvergleich erhoffen, wenn sich Matches noch soweit in Nachbarschaft zueinander befinden, dass eine zusammenhängende, leicht variierende Übernahme zunächst einmal nicht abwegig erscheint, dabei aber doch einen etwas größeren Abstand haben als im zuerst beschriebenen Fall der fast vollständigen Entsprechung.

eine höhere Wahrscheinlichkeit, dass die angesetzte Mindestlänge für die Klassifikation der betreffenden Wortformen als ähnlich erreicht wird.

[628] Die im Folgenden präsentierten Zahlen wurden ohne diese Optimierung ermittelt; sowohl die primäre Matchermittlung (mit einer Mindestlänge von 18 Zeichen) als auch der Feinvergleich wurden auf der Basis der Codierung *ohne A/I/U/B/F* ohne Leerzeichen durchgeführt.

	gleich	ähnlich	Abkürzung	Zusatz	unterschiedlich
ohne Schritt 1–4	208.491 bzw. 210.130	1.623 bzw. 1.635		4.029 bzw. 377	3.131 bzw. 3.421
ohne Schritt 2–4	209.020 bzw. 210.671	1.548 bzw. 1.560		4.053 bzw. 472	2.653 bzw. 2.860
ohne Schritt 3–4	209.319 bzw. 210.955	1.631 bzw. 1.650		4.258 bzw. 658	2.066 bzw. 2.300
ohne Schritt 4	209.581 bzw. 211.211	1.608 bzw. 1.621		4.296 bzw. 765	1.789 bzw. 1.966
alle Schritte	209.592 bzw. 211.222	1.870 bzw. 1.914	22 bzw. 22	4.296 bzw. 853	1.494 bzw. 1.552

(a) Vergleich von BambHalsGO. 1507 mit BrandenbAnsbHalsGO. 1516

	gleich	ähnlich	Abkürzung	Zusatz	unterschiedlich
ohne Schritt 1–4	204.776 bzw. 204.591	4.906 bzw. 4.879		13.090 bzw. 208	38.667 bzw. 21.045
ohne Schritt 2–4	207.443 bzw. 207.272	5.169 bzw. 5.129		17.960 bzw. 881	30.867 bzw. 17.441
ohne Schritt 3–4	208.613 bzw. 208.444	5.253 bzw. 5.221		19.341 bzw. 1.396	28.232 bzw. 15.662
ohne Schritt 4	209.437 bzw. 209.274	5.694 bzw. 5.499		20.351 bzw. 1.930	25.957 bzw. 14.020
alle Schritte	209.583 bzw. 209.419	6.082 bzw. 5.891	52 bzw. 51	20.699 bzw. 2.132	25.023 bzw. 13.230

(b) Vergleich von BayrGO. 1520 mit PassauGO. 1536 (1539)

	gleich	ähnlich	Abkürzung	Zusatz	unterschiedlich
ohne Schritt 1–4	165.653 bzw. 176.998	4.940 bzw. 5.326		19.592 bzw. 37.787	345.270 bzw. 492.060
ohne Schritt 2–4	180.329 bzw. 192.687	6.486 bzw. 6.925		64.053 bzw. 76.109	284.587 bzw. 436.450
ohne Schritt 3–4	186.092 bzw. 198.862	6.296 bzw. 6.808		71.372 bzw. 78.428	271.695 bzw. 428.073
ohne Schritt 4	190.680 bzw. 203.582	7.289 bzw. 7.950		76.956 bzw. 80.199	260.530 bzw. 420.440
alle Schritte	191.292 bzw. 204.195	7.989 bzw. 8.712	193 bzw. 216	79.255 bzw. 81.796	256.726 bzw. 417.252

(c) Vergleich von WormsRef. 1498 (1499) mit Gobler,StatB. 1553

	gleich	ähnlich	Abkürzung	Zusatz	unterschiedlich
ohne Schritt 1–4	60.592 bzw. 60.815	1.673 bzw. 1.611		12.827 bzw. 190	668.631 bzw. 649.555
ohne Schritt 2–4	69.057 bzw. 69.387	2.989 bzw. 2.788		57.403 bzw. 278	614.274 bzw. 639.718
ohne Schritt 3–4	72.090 bzw. 72.522	2.789 bzw. 2.588		62.179 bzw. 353	606.665 bzw. 636.708
ohne Schritt 4	75.249 bzw. 75.732	3.752 bzw. 3.421		68.946 bzw. 448	595.776 bzw. 632.570
alle Schritte	75.574 bzw. 76.052	3.330 bzw. 3.172	2.189 bzw. 1.641	71.967 bzw. 726	590.663 bzw. 630.580

(d) Vergleich von Perneder,Proz. 1544 mit Gobler,StatB. 1553

Tab. 3.16: Auswirkung verschiedener Feinvergleichsschritte (Zählung der verschiedenen Kategorien zugeordneten Zeichen in den Originaltexten; vgl. S. 241–242 und 245–246 sowie Anm. 628)

Tabelle 3.16 soll einen ungefähren Eindruck vermitteln, in welchem Maße mit einer Verbesserung der Erkennungsrate durch einen Feinvergleich nach dem hier vorgestellten Schema zu rechnen ist. Dabei sind relativ zufällig vier Textpaare ausgewählt, die sich zumindest in etwa der eben beschriebenen ersten beziehungsweise letzten Fallkonstellation zuordnen lassen. Auch hier liegen den Daten die entsprechend Tabelle 3.9 gefilterten MEMs der Mindestlänge 18 in der Codierung *ohne A/I/U/B/F* ohne Leerzeichen zugrunde.

Die Brandenburgische Halsgerichtsordnung von 1516 beruht fast völlig auf einer Übernahme des Textes der Bambergischen Halsgerichtsordnung von 1507. Die in Tabelle 3.16a verzeichneten Zusätze und Unterschiede sind zum großen Teil darauf zurückzuführen, dass die *Brandenburgensis* nicht die mit Bildunterschriften versehenen Holzschnitte der *Bambergensis* enthält und dass insbesondere die Bezeichnungen für das jeweilige Territorium, den beziehungsweise die Landesherren,

die Amtsträger und Ähnliches natürlich angepasst wurden.[629] Die Tabelle verdeutlicht, dass trotz dieser sehr engen Textverwandtschaft ohne einen Feinvergleich immerhin etwa 1,5 % als unterschiedlich eingestuft werden, wozu noch die als Zusatz klassifizierten Stücke kommen, die – wie es der Sachlage entspricht – vor allem für die *Bambergensis* festgestellt werden. Zugleich lässt sich erkennen, dass durch den Feinvergleich die Länge der als gleich gewerteten Textpassagen zwar insbesondere im Verhältnis zur Länge der schon durch die primäre Matcherkennung entsprechend zugeordneten Stücke nur relativ wenig ansteigt, dass sich aber der Anteil der Stücke, die als unterschiedlich betrachtet werden, auf etwa die Hälfte des Ausgangswertes reduziert.

Tabelle 3.16b betrifft mit dem Vergleich der Bayrischen Gerichtsordnung von 1520 und der Passauer Gerichtsordnung von 1536 ebenfalls ein Textpaar, das durch eine den Großteil des Textes umfassende Übernahme verbunden ist, allerdings gibt es hier doch zum Teil auch relativ umfangreiche Änderungen, was sich schon an der um mehr als ein Zehntel geringeren Länge der Passauer Gerichtsordnung erkennen lässt. Die anhand der äußeren und enthaltenen Wortgrenzen gefilterten Matches decken aber auch in der Bayrischen Gerichtsordnung mehr als 78 % des Ausgangstextes ab, bei Einbeziehung auch von als ähnlich klassifizierten sehr kurzen Zwischenstücken werden über 80 % zugeordnet. Durch den Feinvergleich lässt sich dieser Wert auf über 82 % steigern.

Wesentlich größer ist der Effekt für den Vergleich des Justin Gobler zugeschriebenen, 1553 publizierten *Statuten Bůchs* mit der *Wormser Reformation* von 1498 beziehungsweise mit dem 1544 erschienenen *Gerichtlichen Process* von Andreas Perneder. Hier machen die übernommenen Stücke jeweils nur den kleineren Teil der Gesamttexte aus, und es finden sich neben größeren Änderungen auch viele kleinere Varianten im Wortlaut beziehungsweise in der Schreibung. Der Feinvergleich erhöht in beiden Fällen den Umfang der erkannten Entsprechungen recht deutlich, nämlich – wenn man wiederum die als ähnlich klassifizierten Wörter mit berücksichtigt – im Verhältnis zu den über Matches mit Minimallänge zugeordneten Stücken um fast 17 beziehungsweise über 26 %.[630] Beim Vergleich mit dem Werk von Perneder spielt auch die Zuordnung aufgrund von vermuteten Abkürzungen eine – wenn auch geringe – Rolle, da diese beiden Texte eine Vielzahl von Verweisen auf das *Corpus Iuris Civilis* beziehungsweise die zugehörige mittelalterliche Literatur enthalten.

[629] Außerdem lassen sich neben kleineren Änderungen von Einzelwörtern, die teilweise auch als Wort- oder Schreibvarianten erklärt werden können (wie etwa die Änderung von „vnfursichtlicher“ in „vnfursichtiger“ in Artikel 172) etwas umfangreichere beziehungsweise sachlich relevante Änderungen im zweiten als Nr. 258 gezählten Artikel sowie in Artikel 261 und 262 feststellen.

[630] Im Verhältnis zum gesamten Ausgangstext ergibt sich für die *Wormser Reformation* eine Steigerung von knapp 32 auf über 37 %, für den *Gerichtlichen Process* eine Steigerung von über 8 auf über 10 %.

Selbstverständlich kann nicht einfach davon ausgegangen werden, dass die Zahlen für diese Konstellationen typisch sind oder Durchschnittswerten entsprechen. So lassen sich auch in einem Textpaar mit fast vollständiger Übereinstimmung des Haupttextes vielfach erhebliche Abweichungen in Textstücken davor oder danach, zum Beispiel in der Vorrede, feststellen, und solche Stücke fallen aufgrund ihres unterschiedlichen Umfangs natürlich auch unterschiedlich stark ins Gewicht. Und in noch stärkerem Maße ist bei Textpaaren mit größeren Lücken zwischen den Matches mit sehr unterschiedlichen Werten zu rechnen. Insgesamt ist aber zum einen erkennbar, dass sich durch einen Feinvergleich eine je nach Textpaar teilweise nicht geringe Ausweitung der einander zugeordneten Textstücke erreichen lässt, aber auch, dass jedenfalls für die betrachteten Textpaare der deutlich überwiegende Teil der Zuordnung auf gefilterten MEMs basiert und damit die Annahme eine zusätzliche Bestätigung erfährt, dass dieses Matchermittlungsverfahren für Textbeziehungen, wie sie hier untersucht werden, recht aussagekräftige Ergebnisse liefert.

4 Untersuchungen zu Texten des Korpus

Teil 4 soll an einigen Beispielen aufzeigen, wie sich mit den in Teil 3 vorgestellten Methoden Erkenntnisse über textuelle Abhängigkeitsverhältnisse gewinnen lassen, und zugleich einen Beitrag zur Erforschung der dabei betrachteten Texte leisten. In Kapitel 4.1 geht es um Justin Gobler, der in der wissenschaftlichen Literatur verschiedentlich als Plagiator dargestellt worden ist, und drei von ihm verfasste beziehungsweise ihm zugeschriebene Kompendien, in Kapitel 4.2 um den Vergleich zweier Werke Thomas Murners und in Kapitel 4.3 um zwei Beispiele für Normtexte, bei deren Abfassung verschiedene Vorlagentexte herangezogen wurden, nämlich um die Heilbronner Statuten von 1541 und die Henneberger Landesordnung von 1539.

4.1 Textübernahmen in Werken Justin Goblers

Justin Gobler ist einer der Hauptautoren der frühneuhochdeutschen Literatur zum römischen Recht. Neben handbuchartigen Werken – die nicht alle unter seinem Namen veröffentlicht sind, ihm aber jedenfalls in der rechtshistorischen Literatur auch weiterhin zugeschrieben werden[631] – ist er auch mit mehreren Übersetzungen im Korpus von DRQEdit vertreten; daneben hat er eine Vielzahl weiterer Texte verfasst, übersetzt beziehungsweise herausgegeben.[632] Für die vorliegende Untersuchung konnten die erste Fassung des *Gerichtlichen Processes* von 1536, der *Rechten Spiegel* von 1550, das *Statuten Bůch* von 1553 sowie die erstmals 1552 publizierte Übersetzung der *Institutiones* des *Corpus Iuris Civilis* ausgewertet werden; das zuletzt genannte Werk spielt aber für die in diesem Kapitel behandelten Fragen fast keine Rolle.

Es ist bekannt, dass die zuerst genannten drei Werke zu einem erheblichen Teil auf wörtlichen Übernahmen beruhen, und in verschiedenen Publikationen vor allem des 19. Jahrhunderts wird Gobler deshalb mit kritischen Worten bedacht. Carl Georg von Wächter setzt sich mit den strafrechtlichen Abschnitten auseinander und beschreibt sie als „mit Ausnahme nur ganz weniger Stellen Abschriften und Auszüge aus der Carolina (nicht Erläuterungen) und *reine Plagiate aus Schriften Anderer*“.[633] Otto Stobbe schreibt, dass Gobler „sich durch seine Vielschreiberei, aber auch zugleich durch die unverschämte Art auszeichnet, mit welcher er andere Arbeiten benützt und ausschreibt“ und dass es „seinen Arbeiten grossentheils

[631] In Deutsch 2009, der wohl letzten Publikation zu Gobler, wird die Autorschaft nicht in Zweifel gezogen. Der *Gerichtliche Proceß* wird auch im VD16 Gobler (dort unter dem Namen *Göbler*) zugeschrieben, nicht aber das *Statuten Bůch*, das dort mit der Urheberbezeichnung „Deutschland, Gesetze“ geführt wird.

[632] Einen Überblick über seine Werke bietet Deutsch 2009, teilweise etwas detailliertere Beschreibungen finden sich in Spangenberg 1825, S. 441–451 sowie Stintzing 1880, Bd. 1, S. 583–585.

[633] Wächter 1836, S. 127 (die Hervorhebung ist dort nicht kursiv, sondern gesperrt gedruckt).

an jedem selbstständigen Verdienst gebricht";[634] den ausdrücklichen Hinweis in Goblers Widmung des *Rechten Spiegels* an Kaiser Karl V., er habe das Werk „auß den Rechtsbůchern vnnd Scribenten / auch andern Lerern (welcher namen hie zumelden vnnǒtig) zusamen getragen",[635] kommentiert Stobbe: „als Plagiarius hielt er es für zweckmässig, seine Gewährsmänner zu verschweigen."[636] Theodor Muther schreibt über die Zusammenarbeit mit dem Drucker Egenolff, bei dem die genannten wie auch weitere Schriften Goblers erschienen: „Gobler [...] schweisste fremde Arbeiten zu voluminösen Compilationen zusammen, oder lieferte auch, je nach Umständen, ohne weitere Zuthaten fremde Bücher unter die Presse".[637] Johann August Roderich von Stintzing meint: „Gobler's literarische Thätigkeit ist ohne jede wissenschaftliche Bedeutung, im Ganzen auf untergeordnete Bedürfnisse berechnet"[638] und beschreibt zum Beispiel den *Rechten Spiegel* als „zum großen Theil nachweislich mechanische Compilation aus dem *Klagspiegel* und Perneder's Schriften".[639] Und in der ersten Auflage des *Handwörterbuchs zur deutschen Rechtsgeschichte* wird Goblers Gesamtwerk von Hans Erich Troje zwar nicht nur kritisch beschrieben, aber es ist doch von seiner „Gewohnheit, seinen statt des Urhebers Namen zu nennen", die Rede.[640]

Im Hinblick auf den Plagiatsvorwurf und gerade auch auf das zuletzt angeführte Zitat ist immerhin bemerkenswert, dass die Werke teilweise anonym publiziert wurden, so dass man wohl nicht behaupten kann, dass Gobler in diesen Fällen eine Autorschaft beanspruche. Für den *Gerichtlichen Proceß* lässt sich dabei immerhin eine Entwicklung feststellen: In den ersten Auflagen (auch nach einer einen großen Teil des Textes betreffenden Überarbeitung) gibt es zwar Vorreden, in denen sich der Verfasser des Werks äußert, aber er tritt dennoch nicht namentlich in Erscheinung. In der Auflage von 1562 heißt es aber auf dem Titelblatt, das Werk sei „vffs newe ersehen vnd gebessert Durch Herrn Justinum Goblerum, der Rechten Doctorn", und in der Widmung erklärt Gobler, er habe seinem Freund Egenolff (dem Drucker des Werks) „vor etlichen verschienen Jaren / zu gegenwertigem Teutschen Gerichtlichen Proceß vrsach geben / vnd gerathen", sowie wenig später, er habe ihm dazu „gerathen / vrsach geben / vnd verholffen".[641] Nach diesen Formulierungen erscheint Gobler also als zwar irgendwie an der Publikation beteiligt, aber ohne

[634] Stobbe 1864, S. 174 f.
[635] Gobler,Rsp. 1550, Blatt * ij v. Als Ziel des Werks gibt Gobler in diesem Zusammenhang an, „dem Leyen vnd Gemeynem vngelerten man / ein kurtze klare anweisung inn Gerichtshåndeln vnnd Rechtlichen / auch sunst gemeynen Burgerlichen Sachen zuthůn", und weist dabei auch auf Ulrich Tengler und andere Autoren mit ähnlicher Intention hin.
[636] Stobbe 1864, S. 176.
[637] Muther 1876, S. 336.
[638] Stintzing 1880, S. 583.
[639] Ebd. S. 585.
[640] Troje 1970, Sp. 1728.
[641] Gobler,GerProz. 1562 Blatt * ij r/v.

dass seine Rolle wirklich greifbar würde. Jedenfalls erklärt er hier nicht einfach, er habe den Text verfasst.[642]

Allerdings ist auch festzuhalten, dass der anonyme Autor in der Vorrede der ersten Fassung des *Gerichtlichen Processes* zwar erklärt, er habe eine Reihe von deutschsprachigen Darstellungen des Gerichtsverfahrens verwendet, dass er dort aber sein eigenes Werk als deutlich überlegen darstellt. So heißt es unter anderem:

> Dieweil es nun dahin kommen / daß solche Teutsche Gerichtliche Proceß[643] allenthalben dermassen eingerissen haben / daß die gemeynen Procuratores sich alleyn deren behelffen / vnd jre Sachen innhalts derselbigen handlen / Vnd aber gedachte Proceß an vil orten mangelbar vnnd vnuolkommen befunden werden / hab ich [...] alle solliche Proceß / auch lo̊bliche vnd rechtmessige Gerichts ordnung ettlicher Stett vnnd Herrschafften im heyligen Reich / zusamen verglichen / vnnd auß iedem das best vnd gegründtest gezogen / auch was in einem gemangelet / auß dem andern erstattet / Darzů durch mich selbst / wo es mich von no̊ten hat angesehen / das jenig auß den Rechten vnnd Rechtgelerten hinzů gesetzt / so den Procuratorn zuwissen gepůret.[644]

Ebenfalls ohne Nennung eines Autors wurde das *Statuten Bůch* publiziert. Als Anhaltspunkt für die Zuschreibung an Gobler dient hier – neben der Überlegung, dass die Art der Kompilation den anderen Werken Goblers entspreche – die Nennung von „vnserm Rechtenspiegel“,[645] dem einzigen der drei Werke, das von der ersten Auflage an Gobler als Verfasser nennt. Das Argument scheint plausibel – ob es zwingend ist oder die Formulierung vielleicht auch anders zu verstehen sein könnte,[646] spielt für die vorliegende Untersuchung keine Rolle. Entsprechend dem wissenschaftlichen Konsens wird im Folgenden der Einfachheit halber die Autorschaft Goblers sowohl für dieses Werk als auch für den *Gerichtlichen Proceß* vorausgesetzt.

Die drei Werke sollen hier jeweils näher betrachtet werden, zunächst und am ausführlichsten der *Gerichtliche Proceß*. Dieses Werk wurde in verschiedenen Fassungen publiziert, die sich ganz erheblich voneinander unterscheiden. Für die Untersuchung im Rahmen dieser Arbeit konnte nur die Fassung der ersten Auflage von 1536 ausgewertet werden. 1542 erschien eine Fassung, die von den drei Teilen der ersten Fassung den ersten und dritten als neuen ersten Teil übernahm und den

[642] Dass Stintzing 1880, S. 584 frühere Auflagen mit Namensnennung aufführt, scheint auf Fehlern zu beruhen, vgl. Bedenbender 2013, S. 324, Anm. 40 (über die Frage der Autorschaft für dieses Werk, mit denselben Zitaten).

[643] Gemeint sind prozessrechtliche Werke, wie sich aus dem Textzusammmenhang ergibt.

[644] Gobler,GerProz. 1536, Blatt a ij r.

[645] Gobler,StatB. 1553, Bl. 9 r. Wächter 1836, S. 138 weist auf diese Stelle hin. Stobbe 1864, S. 176 begründet die Zuschreibung nur mit der „Art der Arbeit“.

[646] Eventuell vorstellbar scheinen mir ein Bezug des Possessivpronomens auf die Druckerei von Egenolff beziehungsweise auf die Gesamtheit derer, deren Recht in diesem Werk dargestellt wird. Inwieweit die Verwendung des Plurals anstelle des Singulars bei Eigenaussagen eines Autors in dieser Zeit üblich war, kann hier nicht untersucht werden.

bisherigen zweiten Teil wegließ, dafür aber einen anderen zweiten Teil ergänzte.[647] In der Auflage von 1549 ist zudem ein weiterer, nicht gezählter Teil unter der Überschrift „Der Gerichtlich Proceß / inn Peinlichen / Criminal vnnd Malefitz Sachen vnnd handlungẽ [...]“[648] enthalten.[649]

Es ist bekannt, dass insbesondere der zweite Teil der Erstfassung des *Gerichtlichen Processes* auf dem *Klagspiegel* basiert.[650] Darüber hinaus gibt es in der hier herangezogenen Literatur keine Hinweise auf weitere Vorlagen, die in diese Erstfassung eingeflossen sind, und über den Plagiatsvorwurf hinaus keine Erörterung, inwieweit man von einer eigenen Leistung des Verfassers sprechen kann.[651]

Im Projektionsdotplot[652], der in Abbildung 4.1 aufgrund des Textumfangs und der Vielzahl von Texten mit Entsprechungen nur in einer im Hinblick auf die Beschriftung kaum lesbaren Größe gezeigt werden kann, lassen sich zum einen verschiedene Textbereiche mit stark unterschiedlichen Übereinstimmungsmustern erkennen, zum anderen treten einige Zeilen hervor, in denen ausgeprägtere Entsprechungen in bestimmten dieser Bereiche verzeichnet sind.

Übereinstimmungen sind vor allem für die knappe erste Hälfte des Textes zu verzeichnen, in der zweiten Hälfte hingegen gibt es weite Bereiche, die lange oder zumindest dicht aufeinander folgende Entsprechungen zum *Klagspiegel* aufweisen (im Projektionsdotplot die längsten zusammenhängenden Linien), während entspre-

[647] So wird es in der Vorrede 1542 beschrieben (Gobler,GerProz. 1542, Blatt 2 r).

[648] Gobler,GerProz. 1549, Bl. 192 v.

[649] In der Literatur wird verschiedentlich eine Neubearbeitung erst für 1549 festgestellt (so Stintzing 1880, S. 584 und Deutsch 2009, Sp. 439), das ist aber nicht der Fall, wie sich schon aus einem Vergleich der Inhaltsverzeichnisse sowie aus der Vorrede von 1542 ergibt. Der Fehler erklärt sich vermutlich daraus, dass die Vorrede der Ausgabe von 1549 allem Anschein nach wörtlich mit der von 1542 übereinstimmt und dass darin die eben beschriebenen Änderungen als Neuerungen gegenüber früheren Auflagen dargestellt werden, während der tatsächlich in dieser Ausgabe angehängte Teil unerwähnt bleibt.

[650] Vgl. unter anderem Wächter 1836, S. 128 und Deutsch 2004, S. 433 f. Der *Klagspiegel* ist ein in der ersten Hälfte des 15. Jahrhunderts verfasstes Werk, das ebenfalls seinen Autor nicht nennt (vgl. zur Verfasserfrage Deutsch 2004, S. 79–222) und das vor allem durch einen Druck von 1516 (Klagspiegel(Brant) 1516) bekannt wurde. Dass dieser Druck laut seinem Titelblatt „Durch doctorem Sebastianum Brandt wider durchsichtiget unnd zům teyl gebessert“ sein soll, ohne dass tatsächlich eine Überarbeitung festgestellt werden kann (vgl. Deutsch 2004, S. 16–26 und ähnlich Deutsch 2010, S. 90–97), weist vielleicht gewisse Parallelen zur Nennung Goblers auf den Titelblättern späterer Ausgaben des *Gerichtlichen Processes* auf.

[651] Die Fassung von 1549 wird in der Literatur verschiedentlich als stark abhängig von zwei 1544 publizierten Werken Andreas Perneders beschrieben (so Wächter 1836, S. 129 f. und Stintzing 1880, S. 584), das kann sich aber nach der Chronologie nur auf Textpassagen beziehen, die in der Fassung von 1542 noch nicht enthalten sind. Dies trifft auf das von Wächter erwähnte „Schlußcapitel“ (vor dem angehängten ungezählten Teil, Bl. 191 v–192 v) zu. Inwieweit der Text von 1542 in der Fassung von 1549 noch anderweitig geändert wurde, kann hier nicht untersucht werden. Theodor Muther beschreibt die Ausgabe von 1578, also wohl die Fassung von 1549 beziehungsweise 1542, als weitgehend vom (erstmals 1541 publizierten) *Process* Kilian Königs abhängig – also von einem Werk, das tatsächlich auch 1542 schon benutzt worden sein konnte – und zitiert als Beleg eine wörtliche Übernahme (Muther 1860, S. 60 sowie Muther 1876, S. 154).

[652] Vgl. oben Unterkapitel 3.4.4.

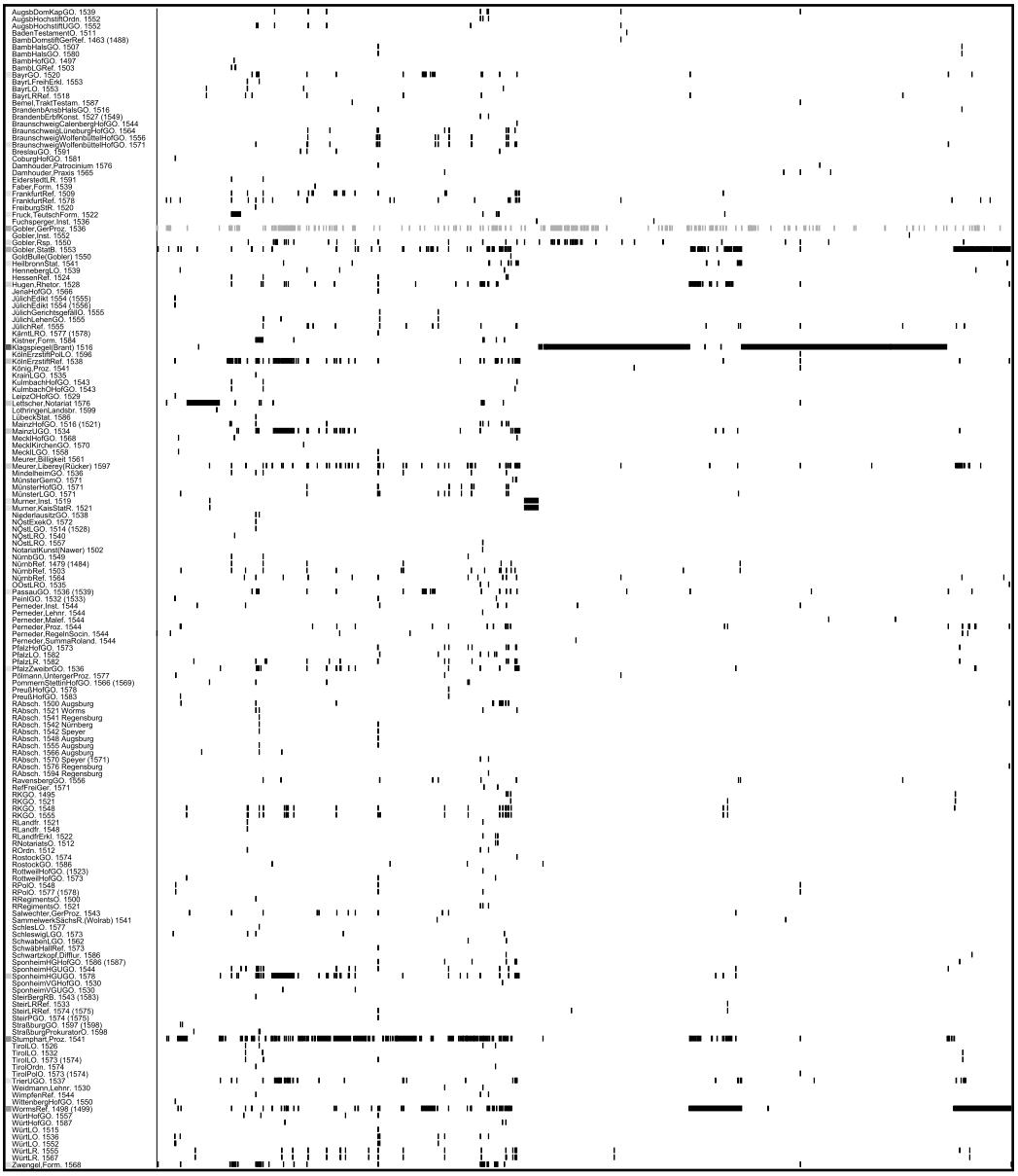


Abb. 4.1: Projektionsdotplot zu Gobler,GerProz. 1536

chende Passagen in anderen Texten kaum zu finden sind.[653] Daneben fällt ein etwas kürzerer Bereich auf, der zwischen der knappen ersten Hälfte und dem Bereich mit Entsprechungen zum *Klagspiegel* liegt und für den fast ausschließlich Übereinstimmungen mit den beiden von Thomas Murner verfassten Werken *Instituten* von 1519 und *Der keiserlichen stat rechten ein ingäg* von 1521 verzeichnet sind, sowie

[653] Auffällig sind in diesen Bereichen nur noch Formulierungswiederholungen im *Gerichtlichen Proceß* selbst, die durch die Zeile mit grauen Blöcken repräsentiert werden, sowie einige Übereinstimmungen mit Goblers *Rechten Spiegel*, die aber im Dotplot für dieses Textpaar zu einem großen Teil auf einer waagerechten Linie liegen. Das deutet, wie oben auf S. 215 bereits dargestellt wurde, darauf hin, dass ihnen wiederholt vorkommenden Formulierungen zugrunde liegen, und jedenfalls beruhen sie nicht auf einer zusammenhängenden Übernahme.

ein Bereich zwischen den beiden *Klagspiegel*-Blöcken und ein weiterer am Ende des Werks, die insbesondere der *Wormser Reformation* von 1498 nahestehen.[654]

Die knappe erste Hälfte des Projektionsdotplots, die dem ersten Teil des *Gerichtlichen Processes* entspricht, ist aufgrund der Vielzahl von Matches wesentlich weniger übersichtlich, aber einige Blöcke fallen doch ins Auge. Auch hier gibt es Bereiche mit ausgeprägten Entsprechungen zur *Wormser Reformation*, daneben fallen die Mainzer Untergerichtsordnung von 1534 und von ihr abhängige Gerichtsordnungen sowie als Autorenwerke das *Teutsch Formularj* von Ludwig Fruck aus dem Jahr 1522 und die erstmals 1528 publizierte *Rethorica* von Alexander Hugen auf, letztere beide aber nur mit eher kurzen Stücken. Relativ unauffällig sind in dieser Darstellung die Übereinstimmungen mit der *Frankfurter Reformation* von 1509, die betreffenden Matches geben aber aufgrund einer etwas größeren Länge Anlass zu einer Überprüfung und sind tatsächlich signifikant. Die umfangreichsten Übereinstimmungen in diesem Bereich bestehen mit dem *Teutschen Process* von Friedrich Stumphart, der aber erst 1541 erschien und damit nicht als Quelle für das Gobler zugeschriebene Werk in Betracht kommt, sondern vielmehr offenbar ein Beispiel dafür ist, dass der *Gerichtliche Proceß* selbst als Vorlage für Textübernahmen genutzt wurde. Das gilt auch für das *Notariat bůch* von Samuel Lettscher aus dem Jahre 1578, für das Matches vor allem in einem Stück verzeichnet sind, das in der Zeile für Stumpharts Text leer ist.[655]

Ausgehend von diesem Befund sollen nun die Entsprechungen zu Texten, die als Vorlage für den *Gerichtlichen Process* in Betracht kommen, etwas näher beschrieben werden. Das Ziel ist dabei nicht, eine auch nur annähernd vollständige Analyse zu bieten, welche Textpassage auf welche Quelle zurückzuführen ist – dafür wäre eine wesentlich gründlichere Prüfung und insbesondere auch ein möglichst umfassendes Korpus einschlägiger Texte erforderlich. Aus den in den Fußnoten verzeichneten Stellen lässt sich aber doch zumindest in etwa erkennen, welches Ausmaß die Übereinstimmungen jeweils erreichen.

[654] Später erschienene Texte sind in dieser Beschreibung nicht berücksichtigt.

[655] Ein strikter Beweis kann hier natürlich nicht geführt werden, da immer auch mit der Möglichkeit zu rechnen ist, dass der *Gerichtliche Proceß* in den entsprechenden Passagen selbst auf einer Vorlage basiert, die nicht zum untersuchten Korpus gehört, und dass die genannten Texte ebenfalls direkt oder über andere Zwischenglieder von diesem Text abhängig sind. In Bezug auf das untersuchte Korpus ist der Befund für die Werke von Stumphart und Lettscher aber eindeutig, da es keinen anderen Text gibt, der die jeweils entsprechenden umfänglicheren Passagen ebenfalls enthält. Eine andere Konstellation liegt bei den Übereinstimmungen mit dem Gobler zugeschriebenen *Statuten Bůch* vor, die weitgehend an Stellen zu finden sind, für die auch Entsprechungen mit der *Wormser Reformation* verzeichnet werden. Auch bei ihnen wäre ohne näheren Vergleich natürlich denkbar, dass der *Gerichtliche Proceß* als direkte Vorlage und damit als Zwischenglied für die Rezeption der *Wormser Reformation* gedient haben könnte, der Vergleich der Dotplots für die Textpaare WormsRef. 1498 (1499) – Gobler,StatB. 1553 und Gobler, GerProz. 1536 – Gobler,StatB. 1553 zeigt aber deutlich, dass die Übereinstimmungen des *Statuten Bůchs* mit der *Wormser Reformation* viel umfangreicher sind als die mit dem *Gerichtlichen Process*.

Der erste nicht ganz kurze Block, in dem der *Gerichtliche Proceß* Entsprechungen zu einem älteren Text im hier ausgewerteten Korpus aufweist, besteht in einer Reihe von Textvorlagen für die Formulierung gerichtlicher Ladungen, die sich weitgehend gleich auch im eben genannten *Teutsch Formularj* finden.[656] Inwieweit es sich dabei tatsächlich um die von Gobler genutzte Quelle handelt, kann nicht sicher gesagt werden, vielmehr ist bei Formularbüchern dieser Art mit einer vielfachen Tradierung darin enthaltener Einzeltexte zu rechnen.

Übereinstimmungen mit der *Rethorica* von Alexander Hugen lassen sich an einigen Stellen des Werks finden. Abweichungen im Wortlaut sind dabei insgesamt wohl ausgeprägter als bei den Abschnitten, die anscheinend aus dem *Teutsch Formularj* stammen. Auch hier handelt es sich um Formulare im tradierten Wortsinn, also um Mustertexte. Es ist also zum einen mit der Möglichkeit zu rechnen, dass die *Rethorica* und der *Gerichtliche Proceß* nicht in einem direkten Abhängigkeitsverhältnis zueinander stehen. Zum anderen ist eine Änderung bestimmter Formulierungen und insbesondere auch der damit zum Ausdruck gebrachten Sachverhalte als Anpassung an die dem jeweiligen Autor vertrauten Gegebenheiten und Formulierungstraditionen gut erklärlich.[657]

Auch die Übereinstimmungen mit der Mainzer Untergerichtsordnung von 1534 beruhen auf der Übernahme von darin enthaltenen Formularen, wobei Gobler die jeweils zu den Klagen gehörigen Urteilsformulare weglässt.[658]

Zwischen dem *Gerichtlichen Proceß* und der *Frankfurter Reformation* von 1509 gibt es, wie bereits erwähnt, nur einige zum Teil recht kurze Bereiche mit wörtlichen Entsprechungen, diese enthalten aber zum großen Teil Matches, die die geforderte

[656] Gobler,GerProz. 1536, Bl. 10 r–11 v; Fruck,TeutschForm. 1522, Bl. 1 r–2 r und 12 v–13 v.

[657] Ein generell bei der Übernahme solcher Formulare häufiger Fall ist die Aktualisierung des Herrschernamens (zum Beispiel Gobler,GerProz. 1536, Bl. 10 v gegenüber Fruck,TeutschForm. 1522, Bl. 1 v, wo trotz des Erscheinungsjahrs noch Maximilian genannt wird). Im Vergleich des *Gerichtlichen Processes* mit Hugens *Rethorica* finden sich gelegentlich Änderungen im Hinblick auf den Sachverhalt, um den es im jeweiligen Schriftstück geht. So wird in Gobler,GerProz. 1536, Bl. 86 v zum Beispiel als weitere Möglichkeit für den Gegenstand eines Verkaufs „ein thonn Hering“ genannt und ebd. Bl. 87 r der für „pferd / acker / weingart 7c.“ gezahlte Preis von „xx. güldenn 7c.“ als in diesem Fall zu hoch beschrieben, während in Hugen,Rhetor. 1528, Bl. 87 v von „xxiiij guldin 7c.“ die Rede ist – vielleicht ein Hinweis auf den jeweils unterschiedlichen Münzwert. Die als im Wesentlichen übereinstimmend ermittelten Passagen lassen sich an diesen Stellen finden (zuerst ist jeweils Gobler,GerProz. 1536 genannt, dann Hugen,Rhetor. 1528): Bl. 51 v–52 r — Bl. 113 v–114 v (mit Auslassung zweier Formulare); Bl. 52 v — Bl. 117 v (zwei kürzere Übereinstimmungen in umgekehrter Reihenfolge innerhalb von unterschiedlichem Text); Bl. 85 v–87 r — Bl. 86 v–88 r (wobei der erläuternde Text, der bei Gobler am Anfang steht, bei Hugen am Ende zu finden ist und bei Gobler ein Formular enthalten ist, der bei Hugen fehlt); Bl. 91 r–92 r — Bl. 83 v–86 r (wobei bei Gobler mehrere Texte fehlen, die bei Hugen zu finden sind).

[658] Das in Gobler,GerProz. 1536 auf Bl. 14 r/v stehende Formular entspricht weitgehend Mainz UGO. 1534, Bl. 7 r/v und die Formulare in Gobler,GerProz. 1536, Bl. 16 v–20 r (ohne das zweite auf der letzten Seite beginnende Formular) MainzUGO. 1534, Bl. 26 r–34 r, allerdings ohne die Urteile sowie ohne einen in der Mainzer Ordnung auf Bl. 26 v–27 r stehenden Abschnitt, der kein Formular ist, sondern eine rechtliche Regelung für das Territorium.

Mindestlänge deutlich überschreiten und sich in der Einzelprüfung tatsächlich als signifikant erweisen. Hier handelt es sich teilweise um Eidformeln, ansonsten aber nicht um Formulare, sondern um die Setzung beziehungsweise Beschreibung rechtlicher Regelungen.[659]

Übereinstimmungen mit der *Wormser Reformation* lassen sich an verschiedenen Stellen finden, wobei der Umfang recht unterschiedlich ist und insbesondere die kürzeren einander zuzuordnenden Passagen zum Teil auch etwas stärkere Änderungen im Wortlaut aufweisen. Neben einer Reihe von kleineren Stücken im ersten Teil sind insbesondere ein größeres Stück im zweiten sowie fast der komplette dritte Teil auf die *Wormser Reformation* zurückzuführen.[660]

Der Vergleich der Übereinstimmungen mit den beiden Texten von Murner zeigt nicht nur, dass es sich in beiden Fällen um die gleichen Textbereiche im *Gerichtlichen Proceß* handelt (was schon aus dem Projektionsdotplot hervorgeht), sondern auch, dass dabei die *Instituten* als Vorlage gedient haben.[661] Es handelt sich dabei um die im Wesentlichen wörtliche Übernahme eines Großteils von Teil 4, Kapitel 6 einer Übersetzung der *Institutiones* des *Corpus Iuris Civilis*.[662] Gobler weist in der Überschrift des Abschnitts mit der Formulierung „Auß dem Tittel der Keyserlichen Instituten / De Actionibus gezogen“[663] ausdrücklich auf den der Übersetzung Murners zugrunde liegenden Originaltext hin, nicht aber auf diese Übersetzung. Das überrascht freilich nicht, wenn man von einer Zitierpraxis ausgeht, die primär dazu dient, auf anerkannte autoritative Werke Bezug zu nehmen. Bemerkenswert ist im Hinblick auf den Plagiatsvorwurf in diesem Zusammenhang, dass Gobler – wenn man die Zuschreibung auch der ersten Fassung des *Gerichtlichen Processes* an ihn nicht in Zweifel zieht – also offenkundig Murners Übersetzung der *Institutiones*

[659] Stellen auf den folgenden Seiten können einander zugeordnet werden (die jeweils erste Angabe bezieht sich auf Gobler,GerProz. 1536, die zweite auf FrankfurtRef. 1509): Bl. 22 v — Bl. 5 r (nur kurz); Bl. 23 r — Bl. 7 r; Bl. 28 r/v — Bl. 7 v–8 r; Bl. 28 r — Bl. 10 v; Bl. 45 v — Bl. 17 v–18 r; Bl. 57 r/v — Bl. 40 r (mit stärkeren Abweichungen).

[660] Die einander entsprechenden Stücke finden sich auf folgenden Seiten beziehungsweise in folgenden Abschnitten (jeweils zuerst ist Gobler,GerProz. 1536 und dann WormsRef. 1498 (1499) angegeben): Bl. 2 r — I 26; Bl. 13 r/v — I 5; Bl. 30 v — I 8; Bl. 34 r — I 9; Bl. 38 r — I 10; Bl. 38 r/v — I 11; Bl. 41 r–43 v — III 3, 2–16; Bl. 44 r — I 13; Bl. 49 v — I 16; Bl. 52 v — II 2; Bl. 53 v — II 3; Bl. 54 v — II 3; Bl. 54 v–55 r — II 3; Bl. 55 v–56 v — II 4–11; Bl. 85 v–93 r — III 1, 1–34; Bl. 125 r–134 r — III 2, 1–38.

[661] Dies ist an verschiedenen Stellen zu erkennen, an denen es zwischen Murners Texten kleine Unterschiede gibt und der *Gerichtliche Proceß* den *Instituten* folgt, zum Beispiel „zů veruolgẽ“ (Murner,Inst. 1519, Bl. 118 r), „zů erfolgen / vnd begerẽ“ (Murner,KaisStatR. 1521, Blatt l iii r) und „zuueruolgen“ (Gobler,GerProz. 1536, Bl. 59 r).

[662] Ausgelassen sind §§ 2–15, der zweite Teil von § 37 sowie § 38 nach der modernen Zählung. Daneben gibt es noch einige kleinere Änderungen, zum Beispiel im Hinblick auf Fremdwörter die Ersetzung von „vffsatzung“ durch „Constitution“ (Murner: IV [6 § 21]; Gobler,GerProz. 1536, Bl. 59 v), den Einschub von „bonæ fidei“ (Murner: IV [6 § 23]; Gobler,GerProz. 1536, Bl. 60 r) und die Ergänzung „Seruiana, Quasi Seruiana, &c.“ (Murner: IV [6 § 26]; Gobler,GerProz. 1536, Bl. 60 r).

[663] Gobler,GerProz. 1536, Bl. 59 r.

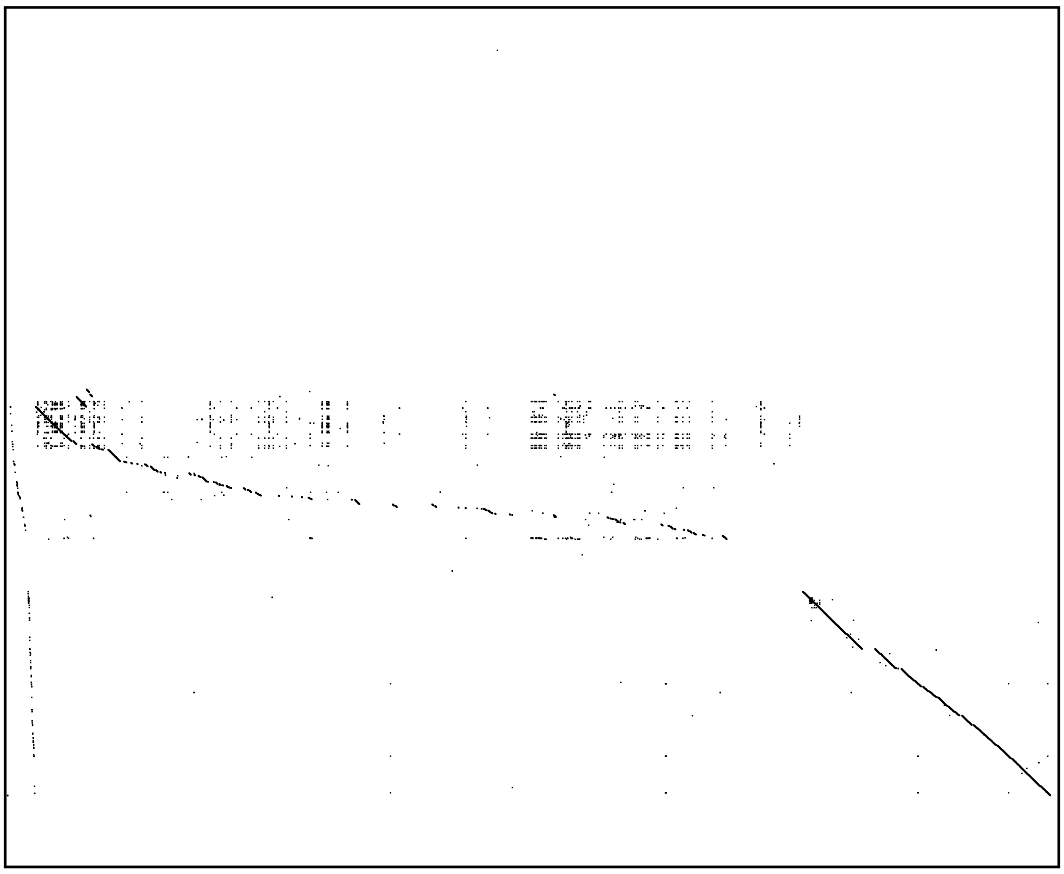


Abb. 4.2: Dotplot: Übernahmen von Klagspiegel(Brant) 1516 in Gobler,GerProz. 1536

kannte, in seiner eigenen Übersetzung und Kommentierung dieses Textes aber allem Anschein nach nicht darauf zurückgegriffen hat.[664]

Die Abhängigkeit des *Gerichtlichen Processes* vom *Klagspiegel* ist, wie bereits erwähnt, in der Literatur wiederholt thematisiert worden, und die Übernahmen aus dieser Quelle sind so umfänglich, dass eine Auflistung der Entsprechungen (oder auch der ausgelassenen Stücke) hier den Rahmen sprengen würde. Der Dotplot in Abbildung 4.2 ermöglicht aber einen Überblick über die von den Übernahmen betroffenen Textbereiche.

In dieser Abbildung steht die x-Achse für den *Klagspiegel* und die y-Achse für den *Gerichtlichen Proceß*. Wie schon festgestellt, finden sich die Übernahmen aus dem *Klagspiegel* nur im zweiten Teil von Goblers Werk; dementsprechend ist der obere Teil des Dotplots fast vollständig leer. In der unteren Hälfte zeigt sich, dass es – wie schon der Projektionsdotplot deutlich macht – im *Gerichtlichen Proceß* einen größeren Bereich zwischen den Übernahmen aus dem *Klagspiegel* und einen weiteren am Ende gibt, die beide nicht auf den *Klagspiegel* zurückzuführen sind – auch

[664] Die mit dem hier eingesetzten Verfahren ermittelten Übereinstimmungen zwischen beiden Übersetzungen sind sehr vereinzelt, kurz und wohl gut erklärlich durch die gemeinsame lateinische Vorlage.

hier finden sich in den entsprechenden waagerechten Bereiche nur ganz wenige Eintragungen, die als zufällige Entsprechungen erklärt werden können.

Im Hinblick auf die Bereiche mit Übernahmen ergibt sich aber ein gegenüber dem Projektionsdotplot wesentlich präziseres Bild. Auf der *x*-Achse gibt es eine auffällige Zweiteilung: Während sich ungefähr für das letzte Viertel eine recht deutliche Diagonale erkennen lässt, die nur einmal durch eine etwas größere Lücke unterbrochen ist und daneben einige kleinere Verschiebungen aufweist, lässt sich im Überblick zwar auch für die ersten drei Viertel eine Zuordnungslinie erkennen, aber diese ist vielfach unterbrochen durch Lücken auf der *x*-Achse, also durch Textstücke, die im *Klagspiegel* stehen, aber nicht im *Gerichtlichen Proceß*. Zudem gibt es in diesem Bereich offenbar Formulierungen, die in beiden Texten vielfach vorkommen, wobei diese Formulierungen im *Gerichtlichen Proceß* vor allem zu Beginn der Übernahmen aus dem *Klagspiegel* zu finden sind – dementsprechend ergibt sich im Dotplot der Eindruck eines waagerechten Bereichs mit einer Konzentration von Matches, und in diesem Bereich kann man ein streifenförmiges Muster erkennen.

Die Zweiteilung des Dotplots auf der *x*-Achse lässt sich sehr gut in Beziehung setzen zur Textstruktur des *Klagspiegels*. Dieser gliedert sich nämlich in zwei Teile, wobei der zweite Teil in etwa den Umfang eines Viertels des Gesamtwerks hat. Der erste Teil enthält neben sachlichen Erläuterungen eine Vielzahl von Formularen für die Klage vor Gericht, die typischerweise in etwa mit der Formulierung „Also formier dein clage. Herr richter ich clag eüch von .N.“[665] anfangen. Der zweite Teil behandelt straf- und strafprozessrechtliche Fragen. Worauf es zurückzuführen ist, dass diese Teile in deutlich unterschiedlichem Maße in den *Gerichtlichen Proceß* übernommen wurden, wäre eine weiterführende inhaltliche Frage, die im Rahmen dieser Untersuchung nicht beantwortet werden kann.

Was ergibt sich nun als Gesamtbild für diesen Text? Die in der Literatur betonte starke Abhängigkeit des Werks vom *Klagspiegel* ist zweifellos zutreffend, aber der erste Teil des *Gerichtlichen Processes* beruht offenbar auf verschiedenen Quellen, und zwar – wie in der Vorrede behauptet – sowohl auf Darstellungen des Prozessrechts als auch auf Normtexten. Die hier ermittelten und als signifikant eingestuften Übereinstimmungen machen allerdings nur einen kleinen Teil des Textes aus, und auch wenn zu vermuten ist, dass die Erweiterung des Untersuchungskorpus die Identifikation weiterer Vorlagen ermöglicht, muss doch zunächst einmal davon ausgegangen werden, dass der Text in einem wohl erheblichen Teil nicht einfach auf Kompilation, sondern auf einer eigenständigen verfasserischen Leistung beruht. Selbst wenn sich diese Einschätzung bei Auswertung eines größeren Korpus als Irrtum erweisen sollte, handelt es sich aber jedenfalls nicht einfach um die Übernahme eines einzigen Werks, sondern um die Zusammenführung von Stücken aus

[665] So zum Beispiel Klagspiegel(Brant) 1516, Bl. III r.

verschiedenen Vorlagen. Wie groß die Leistung bei der Auswahl und Ordnung dieser Passagen ist, kann hier nicht eingeschätzt werden.

Auch der erstmals 1550 erschienene, ausdrücklich von Gobler stammende *Rechten Spiegel* wird verschiedentlich als Plagiat dargestellt. Wohl am ausführlichsten äußert sich Carl Georg von Wächter, der im achten Teil des Werks, der Straf- und Strafprozessrecht zum Inhalt hat, neben umfänglichen Übernahmen aus der *Peinlichen Gerichtsordnung* von 1532 vor allem ein Plagiat der 1544 publizierten *Gerichtlichen Practica aller Malefitz* von Andreas Perneder (bei Wächter in Anlehnung an die originalen Kopfzeilen als „H. G. O.“, also als *Halsgerichtsordnung* bezeichnet) und darüber hinaus einen „Auszug“ aus dem *Klagspiegel* feststellt.[666] Für Letzteres lassen sich – jedenfalls im genannten achten Teil – mit dem hier angewandten Untersuchungsverfahren allerdings keine Belege finden.[667] Für den Rest des Werks vermutet Wächter ebenfalls „großentheils“ einen plagiatorischen Charakter und weist auf die Abhängigkeit des neunten Teils von Perneders *Lehnrecht* hin.[668] Otto Stobbe erklärt außerdem noch, dass der zehnte Teil „eine kurze römische Rechtsgeschichte, dann Auszüge aus den Reichsgesetzen, Uebersetzungen aus dem Corpus juris civilis u. s. w.“ enthalte, die für das Gesamturteil, dass das Werk „zum grossen Theile den Vorgängern des Verfassers an[gehört]“, angeführten Hinweise entsprechen aber ansonsten denen Wächters.[669]

Der Projektionsdotplot zu diesem Text wurde schon oben auf S. 232 gezeigt, und auf S. 227 wurde kurz beschrieben, dass sich darin zum einen ein Bereich mit starken Übereinstimmungen mit der *Bambergensis*, der *Carolina* und verwandten Texten erkennen lässt, zum anderen im Anschluss daran ein weiterer Bereich mit fast durchgehenden Entsprechungen in einem einzigen Text. Dabei handelt es sich um das von Wächter genannte Werk *Der Lehenrecht kurtze vnd aygentliche Verteütschung* Andreas Perneders. Die Übereinstimmungen mit Perneders *Practica aller Malefitz* fallen demgegenüber nur wenig auf; sie sind an einigen Stellen innerhalb des Gesamtbereichs zu finden, der sich den peinlichen Gerichtsordnungen zuordnen lässt, dabei zum Teil an Stellen, an denen zum Beispiel die Zeile, die der *Carolina* entspricht, Lücken aufweist. Darüber hinaus gibt es auch noch einige weitere Texte, zu denen es jeweils in ziemlich kleinen Bereichen des *Rechten Spiegels*

[666] Wächter 1836, S. 132–134.

[667] Wächter gibt die Stelle im *Rechten Spiegel* nicht genau an, sie müsste allerdings nach der Beschreibung kurz nach Beginn des achten Teils zu finden sein. Für den *Klagspiegel* nennt er Bl. 135–142 in der Ausgabe von 1516, die auch dieser Untersuchung zugrunde liegt. Im sechsten Teil des *Rechten Spiegels* gibt es tatsächlich einige, aber nur recht kurze Entsprechungen, vor allem in Formulierungsmustern für bestimmte gerichtliche Klagen. So stimmen als wohl umfangreichste Stücke die dort unter den Überschriften „Ein andere Klag / schlechter einfeltiger Forme“ und „Form der Klag Fauianæ“ (Gobler,Rsp. 1550 Bl. 62 r und 63 r) angeführten Mustertexte weitgehend mit Formularen aus dem *Klagspiegel* (Klagspiegel(Brant) 1516, Bl. 1 r und 2 r) überein.

[668] Wächter 1836, S. 133, Anm. 28.

[669] Stobbe 1864, S. 175 mit Anm. 37, wo auch auf Wächter verwiesen wird.

längere beziehungsweise dicht aufeinander folgende Entsprechungen gibt.[670] Für einen Großteil des Werks lassen sich aber im hier ausgewerteten Korpus keine oder allenfalls kurze wörtliche Übereinstimmungen feststellen, so dass die Annahme, das Werk sei ganz überwiegend als Plagiat einzustufen, bis zum Erweis des Gegenteils wohl zurückgewiesen werden sollte (jedenfalls soweit man als Kriterium dafür weitgehende wörtliche Abhängigkeit von Vorlagen und nicht nur inhaltliche Übereinstimmungen betrachtet).

Ein deutlich anderes Bild zeigt der Projektionsdotplot für das *Statuten Bůch* in Abbildung 4.3.

Auch hier findet sich – in diesem Fall am Ende des Werks – ein Bereich, der durch starke Übereinstimmungen mit der Gruppe der peinlichen Gerichtsordnungen in der Tradition der *Bambergensis* auffällt. Diese Übernahmen werden allerdings deutlich stärker als im *Rechten Spiegel* durch Zwischenstücke unterbrochen, die im Wesentlichen auf Perneders *Practica aller Malefitz* beziehungsweise auf dessen *Institutiones* zurückgeführt werden können. Und in einem Großteil des übrigen Werks zeigt der Projektionsdotplot in verschiedenen Zeilen, insbesondere aber in denen, die für Goblers *Gerichtlichen Proceß* von 1536, die *Wormser Reformation* von 1498 und für Perneders Werke *Gerichtlicher Process* und *Institutiones* stehen, längere zusammenhängende Linienstücke. Hier ergibt sich also tatsächlich der Gesamteindruck eines Werks, dessen Text wohl im Wesentlichen auf die Kombination von Ausschnitten einiger weniger Vorlagen zurückgeführt werden kann.

Trotzdem kann man aber kleinere Änderungen nicht einfach von vornherein als belanglos abtun. So zeigt sich bei den Entsprechungen zur *Wormser Reformation*, dass im *Statuten Bůch* immer wieder Allegationen, also Stellenangaben zum *Corpus Iuris Civilis* beziehungsweise zu der zugehörigen Literatur zu finden sind, die in der *Wormser Reformation* fehlen und auch nicht auf eine andere der hier ausgewerteten Quellen zurückgeführt werden können.[671] Hintergrund mag eine Angleichung an die von Perneder übernommenen Abschnitte sein, die ebenfalls solche Angaben enthalten. Jedenfalls ist aber im Hinblick auf diese Ergänzungen – sofern sich nicht noch eine andere Quelle finden lässt, die den Text der *Wormser Reformation* mit den im *Statuten Bůch* angeführten Allegationen anreichert und als Vorlage gedient haben könnte – von einer eigenständigen Leistung des Verfassers beziehungsweise Kompilators auszugehen.

[670] Als Beispiele seien genannt: Auf Bl. 32 v gibt es eine Passage, die auf Murner,Inst. 1519, Bl. 9 v–10 v zurückgeführt werden kann (wobei es mehrere Auslassungen und kleinere Änderungen gibt). Wohl am umfangreichsten von den nicht schon von Wächter erwähnten Übernahmen ist die aus TrierUGO. 1537, die sich auf Bl. 28 r–30 r findet.

[671] Beispiele hierfür sind: „Spec. in tit. de cita. §. pen." (Bl. 11 v); „Spec. d. c. de Cita." und „Spec. in tit. de Cita. §. sequitur" (Bl. 12 r); „d. l. Sancimus. C. de iudic." (Bl. 15 r); „Auth. Offeratur" und „l. ij in prin. C. eo. de iureiur." (Bl. 15 v). Die entsprechenden Stellen ohne Allegationen lassen sich in WormsRef. 1498 (1499) in den Abschnitten I 1–3 und 5 f. finden.

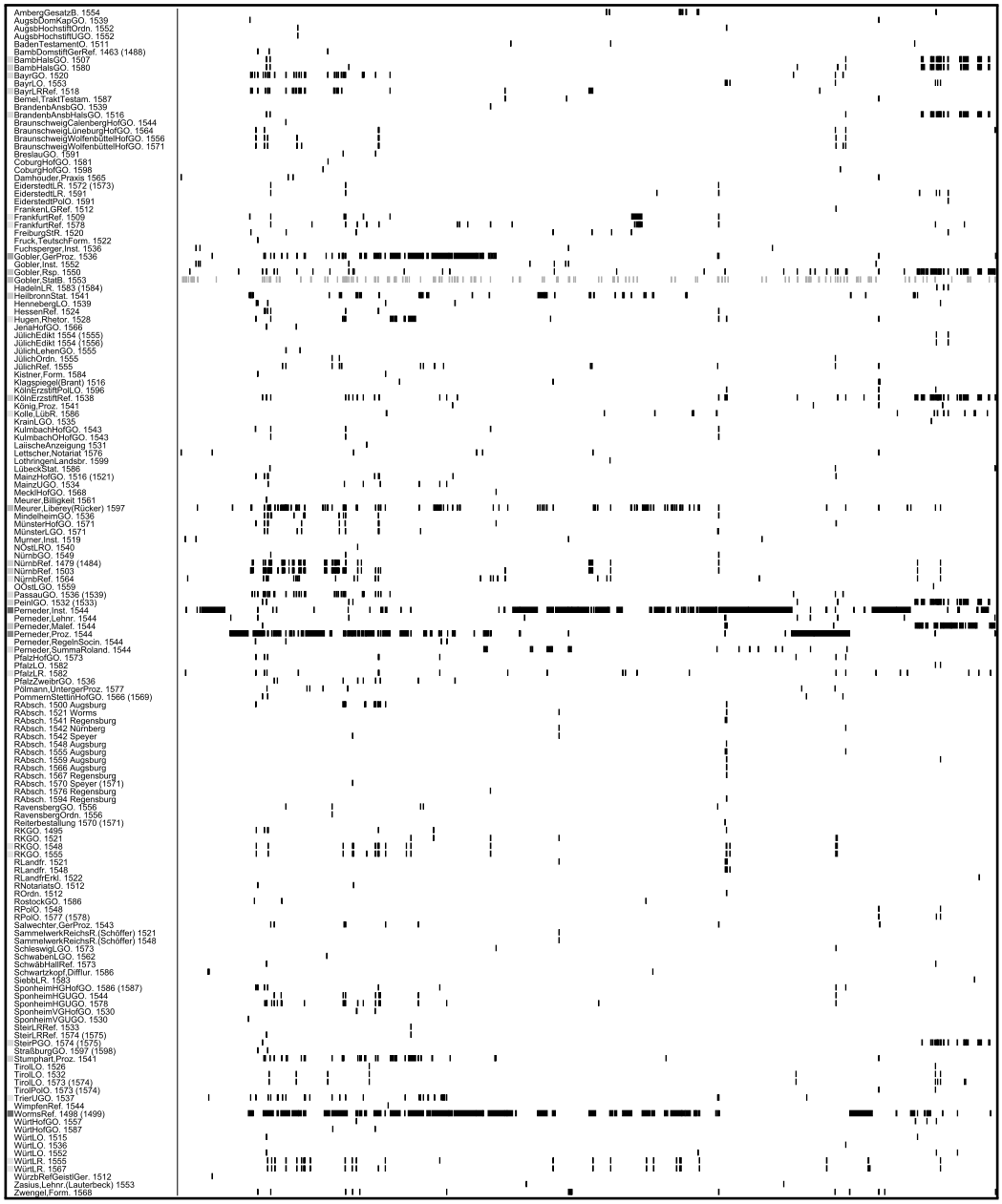


Abb. 4.3: Projektionsdotplot zu Gobler,StatB. 1553

4.2 Thomas Murners *Instituten* und *Ingang*

Thomas Murner ist als humanistischer Autor und Gegner Luthers bekannt. Seine juristischen Publikationen haben in der Forschung vergleichsweise wenig Aufmerksamkeit erfahren. Die wohl einzige Arbeit, die sich etwas vertiefter mit diesen Texten auseinandersetzt, stammt von Adalbert Erler.[672] Er erklärt zwar, Murner

[672] Als spätere Publikationen, die sich mit diesen Werken beziehungsweise mit Murner als Juristen befassen, sind noch Kisch 1962, S. 86–93, der Katalogbeitrag Kaib 1987 sowie die Lexikonar-

könne „nicht den Namen eines *großen* Juristen beanspruchen“,[673] betont aber die „glänzende pädagogische Begabung“[674] und seine Leistung bei der Darstellung des römischen Rechts in deutscher Sprache,[675] die Murner vermutlich auch als erster in einer juristischen Vorlesung verwandte.[676]

Murners juristische Werke stehen teilweise offenbar in engem Zusammenhang mit seiner universitären Lehrtätigkeit[677] und dienen der Vermittlung von Inhalten des *Corpus Iuris Civilis*. Als erste teilweise deutschsprachige Publikation ist *Vtriusq[ue] iuris tituli et regule [...] in Alemanicum traducti* aus dem Jahr 1518 zu nennen, eine Auflistung und Übersetzung insbesondere der Titelrubriken von Werken des römischen und kanonischen Rechts.[678] 1519 folgte mit *Instituten ein warer vrsprung vnnd fundament des Keyserlichen rechtens* die erste deutsche Übersetzung der *Institutiones*, also des zum *Corpus Iuris Civilis* gehörenden Anfängerlehrbuchs. Schließlich erschien 1521 *Der keiserlichen stat rechten ein ingãg vnd wares fundamẽt*, ein Werk, das nach der Vorrede zum Ziel hat,

> mit vertütschten rechten / die fürsichtigen / frummen / weisen / meister vñ rädten in den stetten vnsers keiserthůms zůbegaben / zů fürderung vnnd erlüchtung des rechtens[679].

In dieser Vorrede legt Murner auch dar, dass er sich eigentlich zum Ziel gesetzt habe, „das gantz keiserlich recht das in im sibẽ vnd sübzig bücher verfasset zů verteutschen“, also wohl das *Corpus Iuris Civilis* einschließlich damals als zugehörig betrachteter Teile,[680] und dass er dies „den merẽdeil vollendet“ habe.[681]

> So aber nun das groß mer des Keiserlichen rechtens in siben vnd sibentzig bücher verfasset / so gähenlich nit kan oder mag vberschwümmen vnd angegriffen werden. Hab ich mit groser arbeit dises bůch in siben theil geteilt zů dem andern mal vertütschet / vnd vff ein leichtern verstant geordnet ein waren brun vnd yngang / auch das rechtschuldig fundament aller keiserlichen rechten / meinem grosen werck für wo̊llen lassen gon / darin sich die recht begirigen vor allen dingen erlernen sollen vnd erkünden.[682]

Dies ist wohl so zu verstehen, dass der *Ingang* eine Neuübersetzung der *Institutiones* sein soll, wobei Textreihenfolge und -strukturierung Abweichungen aufweisen.

---

tikel Erler 1981 und Worstbrock 2013, Sp. 339–341 zu nennen, sie basieren aber wohl im Wesentlichen (im Hinblick auf die hier interessierenden Punkte) auf Erlers Monographie.

673 Erler 1956, S. 7 (die Hervorhebung ist im Original gesperrt).

674 Ebd. S. 13.

675 Ebd. S. 23 ff.

676 Vgl. ebd. S. 45–47 und zustimmend Kisch 1962, S. 90.

677 Vgl. Erler 1956, S. 46.

678 Vgl. ebd. S. 20, 28–35 und 39 f.

679 Murner,KaisStatR. 1521, Vorrede, S. [3] der ungezählten ersten Lage.

680 Die Zahl 77 könnte zu erklären sein als Summe von 4 (*Institutiones*) + 50 (Digesten) + 12 (Codex) + 9 (Novellen in der Gliederung, die in der späteren Referenzausgabe von Lyon 1627 zu finden ist, vgl. *http://digi.ub.uni-heidelberg.de/diglit/justinian1627bd5/*) + 2 (*Libri Feudorum*).

681 Murner,KaisStatR. 1521, S. [5].

682 Murner,KaisStatR. 1521, S. [6].

Unklar bleibt dabei sowohl, wie eng sich Murner an den Aufbau der *Institutiones* hält, als auch, inwieweit die Übersetzung auf der in den *Instituten* beruht.

In der wissenschaftlichen Literatur finden sich nur wenige Hinweise zur Charakterisierung des *Ingangs*.[683] Am ausführlichsten sind auch in diesem Punkt die Bemerkungen von Erler:

> In der Tat lehnt sich die Arbeit Murners weitgehend an die Institutionen an. Trotzdem kann man nicht sagen, daß die Schrift „eigentlich nur eine dritte Ausgabe der Institutionen sei" [Fußnote: So v. Liebenau S. 134]. Die einleitenden Ausführungen über Recht und Gerechtigkeit sind frei und im mittelalterlichen Geist behandelt. Gegenüber den Instituten bilden die *Stat Rechte* einen freien Grundriß, der alle archaischen Erinnerungen der Institutionen – etwa jene des Titels 2, 10 – sowie alle auf römische Sonderverhältnisse zugeschnittenen und deshalb nicht mehr praktischen Vorschriften beiseite läßt. Da die Arbeit auch keine Übersetzung darstellt, erheben sich Stil und Inhalt zu einer gewissen Freiheit und Anschaulichkeit, die den *Instituten* bisweilen fehlt.[684]

Um das Verhältnis zu den *Instituten* beziehungsweise den *Institutiones* genauer zu untersuchen, muss zunächst einmal festgestellt werden, welche Textstücke die Vorlage für die einzelnen Abschnitte sind oder zumindest als solche in Frage kommen. Gerade dies ist aber keineswegs einfach. Schon die Gliederung auf der obersten Ebene stimmt nicht überein: Die *Institutiones* haben vier Teile, der *Ingang* hingegen sieben. Generell finden sich im *Ingang* keine Allegationen, die eine Zuordnung ermöglichen würden. Nur in einem kleinen Teil der Überschriften finden sich unmittelbare Entsprechungen. Und auch dort, wo sie vorhanden sind, lassen sich die zugehörigen Abschnitte nicht in jedem Fall aufeinander beziehen.

So lautet zwar die erste Überschrift in Teil 1 des *Ingangs* in wörtlicher Übereinstimmung mit der Überschrift zu Buch 1, Titel 1 der *Instituten* „Von gerechtigkeyt vnd dẽ rechten", tatsächlich handelt es sich bei dem folgenden Absatz aber offenbar um eine partielle Übersetzung des Anfangs von Buch 1, Titel 1 der *Digestae*, der wie Buch 1, Titel 1 der *Institutiones* die Überschrift *De iustitia et iure* trägt, und auch die folgenden Abschnitte lassen sich wohl zu großen Teilen auf diese Quelle zurückführen,[685] sie sind allerdings wohl nur teilweise eine Übersetzung (von ausgewählten Passagen), da Murner anscheinend auch eigene Strukturierungen in die Darstellung einbaut.[686]

[683] Vgl. Kaib 1987, S. 99 und Worstbrock 2013, Sp. 341.

[684] Erler 1956, S. 50 f. Ebd. S. 47–50 werden die Prosa- und Reimvorrede sowie der Titel des Werks behandelt.

[685] Für den Hinweis auf diese Abhängigkeit danke ich Herrn Dr. Heino Speer. Da es auch zwischen Buch 1, Titel 1 der *Institutiones* und Buch 1, Titel 1 der *Digestae* Textübereinstimmungen gibt, ließen sich die ihnen zuzuordnenden Sätze aus dem *Ingang* natürlich prinzipiell auch als Übersetzung der *Institutiones* interpretieren, aber da Murner offenbar den Anfang der *Digestae* für seine einleitenden Abschnitte herangezogen hat, scheint es kaum plausibel, hier einen Wechsel der Vorlage anzunehmen.

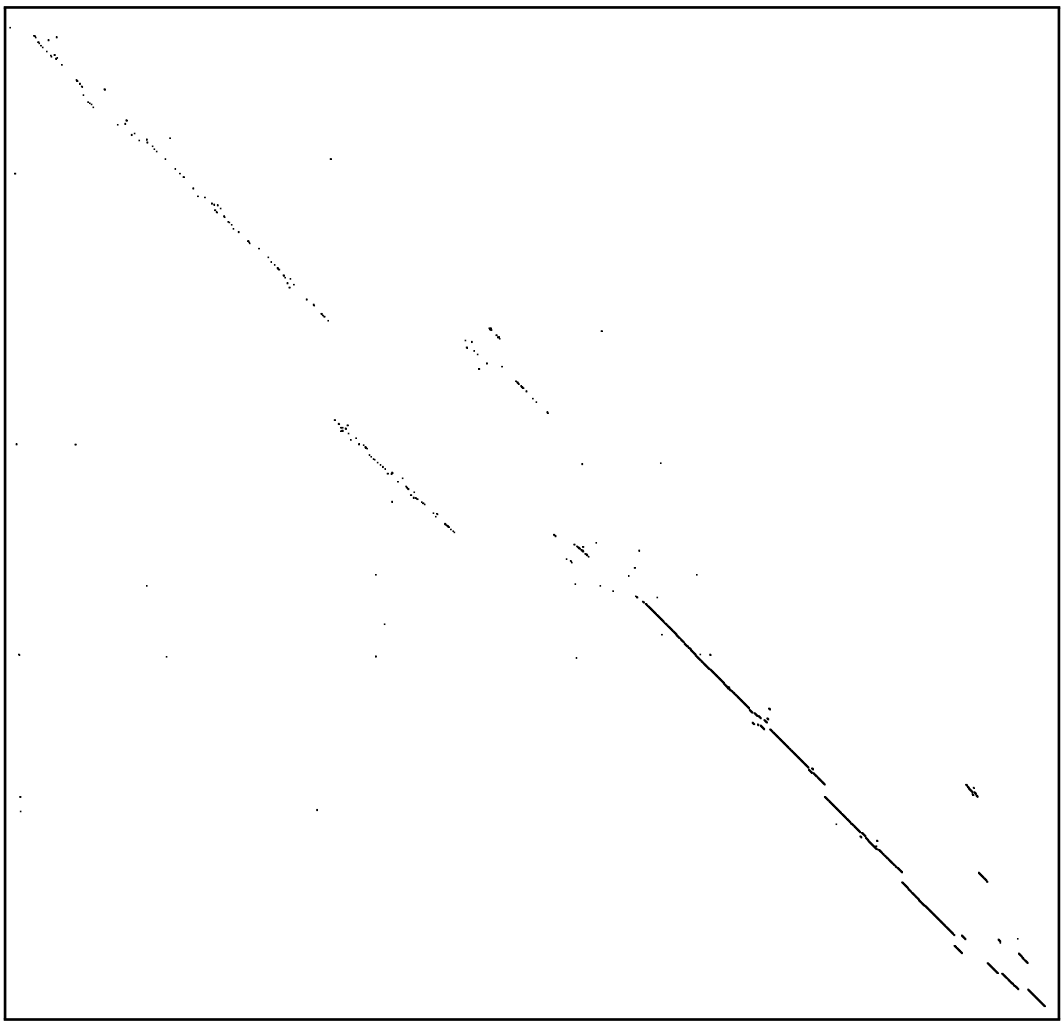


Abb. 4.4: Dotplot Murner,Inst. 1519 – Murner,KaisStatR. 1521

Abgesehen vom Anfang des Werks – der auf Murners ursprüngliches Ziel einer Übersetzung des gesamten *Corpus Iuris Civilis* zurückzuführen sein mag – lässt sich aber, wie Erler dargestellt hat, eine Abhängigkeit von den *Institutiones* beziehungsweise den *Instituten* feststellen. Der Dotplot in Abbildung 4.4 zeigt deutlich, dass sich fast überall Textübereinstimmungen mit den *Instituten* finden, wobei ganz

[686] So führt er im zweiten Absatz unter der Überschrift „Die ander zerteylung des sunderlichen rechten“ auf Blatt A ii v vier „exemplen“ für das natürliche Recht, das die Menschen mit den Tieren teilten, an. In der entsprechenden Passage der *Digestae* (D.1.1.1.3) werden aber (ohne Zählung) nur die ersten drei davon genannt. Zur vierten Aussage, „den mit naturlichem rechten syndt alle menschen anfengklich fry erboren“, findet sich aber eine Entsprechung in D.1.1.4 („cum iure naturali omnes liberi nascerentur“, zitiert nach CICiv.(Krüger/Mommsen)). Inwieweit sich auch andere Passagen, die keine unmittelbare oder vollständige Entsprechung im Digesten-Text haben, durch eine solche Umstrukturierung erklären lassen und inwieweit es sich um eine eigenständige Leistung Murners handelt beziehungsweise um eine Darstellung entsprechend der im universitären Rahmen etablierten Lehre des römischen Rechts, kann hier nicht untersucht werden.

überwiegend die Reihenfolge der einander entsprechenden Passagen in beiden Werken die gleiche ist, aber einige etwas längere Stücke umgestellt wurden. Der Dotplot lässt darüber hinaus eine Zweiteilung erkennen: Die Übereinstimmungen in etwas mehr als der ersten Hälfte des Werkes verteilen sich zwar über den Text, dabei handelt es sich aber jeweils um vergleichsweise kurze Stücke. Im restlichen Text hingegen zeigen sich ziemlich lange zusammenhängende diagonale Linien, die nur an wenigen Stellen voneinander getrennt sind. Hier kann man also von einem wesentlich stärkeren Übereinstimmungsgrad ausgehen.

Über einen Feinvergleich der beiden Texte lässt sich tatsächlich feststellen, dass der Text des *Ingangs* ziemlich genau ab dem Beginn von Blatt [Z iiii] r (mitten in der Darstellung der Verwandtschaftsgrade[687]) weitgehend auf einer wörtlichen Übernahme des Texts der *Instituten* beruht. Auch anschließend finden sich zwar immer wieder kleinere Formulierungsänderungen, Auslassungen und Einschübe sowie – wie schon aus dem Dotplot ersichtlich – an wenigen Stellen etwas umfangreichere Umstellungen, man kann hier aber sicherlich nicht von einem „freien Grundriß“ sprechen, wie Erler es in der oben angeführten Beschreibung tut. Da es sich wohl kaum inhaltlich begründen lässt, warum Murner gerade ab der genannten Stelle keinen Bedarf mehr für eine Neufassung des Texts gesehen haben sollte, legt sich wohl die Vermutung nahe, dass er die vertiefte Überarbeitung in der Mitte des Werks einfach abgebrochen hat, so wie er ja auch seinen in der Vorrede dargelegten Plan, das gesamte *Corpus Iuris Civilis* zu übertragen, trotz der dort gemachten Behauptung, dies „den merẽdeil vollendet“ zu haben, zumindest nicht so weit umgesetzt hat, dass dies zu einem Druckwerk geführt hätte.

Ein nicht unerheblicher Teil der Änderungen in dieser knappen zweiten Hälfte des Textes ist in den Überschriften zu finden, die – wie schon festgestellt – vielfach nicht mit denen der *Instituten* übereinstimmen und zudem insgesamt zahlreicher als diese sind, wobei aber zugleich die in den *Instituten* zusätzlich wiedergegebenen originalen lateinischen Überschriften und Initien entfallen. Als weitere etwas häufigere Änderung ist noch zu verzeichnen, dass Murner an einigen Stellen Zählungen in den Text eingefügt hat.[688]

In der ersten Hälfte des Werks hingegen weicht der Text so stark von dem der *Instituten* ab, dass vielfach jedenfalls ohne vertiefte Betrachtung nicht sicher zu entscheiden ist, in welchem Maße er überhaupt darauf beziehungsweise auf den *Institutiones* basiert und ob Murner hier auf weitere Texte – wie etwa die Digesten

[687] Die entsprechende Textstelle in den *Institutiones* ist Buch 3, Titel 6, § 4 (Zählung nach CICiv.(Krüger/Mommsen)).

[688] So auf Blatt g r/v: „Die erst eigentschafft der beuel“, „Die ander“, „Die drit“ usw. Ähnliches findet sich auf Blatt i r/v und mehrfach auf Blatt [k iiii] v – l v sowie in den ersten Überschriften von Teil 7 (Blatt [k iiii] r – n r). Wesentlich häufiger setzt Murner dieses Strukturierungsmittel in der ersten, freier formulierten Hälfte des Werks ein, so schon gleich in den ersten Abschnitten mehrfach.

oder die mittelalterlichen Glossen und Kommentare zum *Corpus Iuris Civilis* – zurückgegriffen oder auch eigene Strukturierungen vorgenommen hat.

Eine nähere Untersuchung der Übereinstimmungen und Abweichungen kann im Rahmen dieser Arbeit nicht geleistet werden. Tabelle 4.1 bietet aber einen Überblick, welche Abschnitte der *Instituten* sich mit dem hier entwickelten Verfahren welchen Seitenbereichen des *Ingangs* zuordnen lassen.

Die Tabellenzeilen sind nach der Textreihenfolge in den *Instituten* sortiert. Die Stellen dieses Texts sind nach der heute üblichen Zitierweise für die *Institutiones* bezeichnet,[689] um den Fachkundigen eine schnelle Orientierung zu ermöglichen, wo es Umstellungen oder größere Lücken (beziehungsweise noch nicht ermittelte Entsprechungen) gibt.

Aufgrund der ausgeprägten Änderungen und der Umstellungen ist eine Abgrenzung der Entsprechungen schwierig. Im Interesse einer möglichst aussagekräftigen Zuordnung wurden die automatisch ermittelten Bereiche durchgeschaut und präzisiert, mit Ungenauigkeiten ist allerdings zu rechnen.[690]

Insbesondere ist darauf hinzuweisen, dass sich die angegebenen Abschnitte und Seiten, auch wenn sie als Bereiche zusammengefasst sind, in der ersten Hälfte in der Regel nur zu einem relativ kleinen Teil aufeinander abbilden lassen. Sie können durchaus auch nicht ganz kurze Stücke enthalten, in denen sich keine Übereinstimmung feststellen lässt. Eine Aufteilung auf verschiedene Tabellenzeilen wurde bei etwas längeren Passagen ohne Entsprechung sowie bei Umstellungen vorgenommen. Die Tabelle enthält auch die kurzen Entsprechungen zu Beginn des Werks, die wohl auf Textübereinstimmungen zwischen den *Institutiones* und den hier tatsächlich zugrunde liegenden *Digestae* zurückzuführen sind.

Um das Verhältnis der beiden Texte zumindest exemplarisch etwas näher zu betrachten, soll hier schließlich noch das Vergleichsergebnis zum von Erler genannten Titel 2, 10 der *Institutiones* vorgestellt werden. Entgegen Erlers Beschreibung, dass Murner die darin enthaltenen „archaischen Erinnerungen“ weggelassen habe, ist festzustellen, dass der *Ingang* eine Beschreibung von vorjustinianischen Rechtsverhältnissen enthält, die ausdrücklich als veraltet dargestellt werden. Tabelle 4.2 zeigt als Synopse den ersten Abschnitt des Titels in der Übersetzung der *Instituten* sowie den entsprechenden ersten Abschnitt von Teil 4 des *Ingangs*, in denen es zu großen Teilen um dieses ältere Recht geht. Wörtliche Entsprechungen sind dabei

[689] Diese Stellenangaben beruhen auf *milestone-Tags*, die von Dr. Heino Speer in die hier zugrunde gelegte XML-Datei der *Instituten* eingefügt wurden.

[690] Insbesondere für die Erkennung von kurzen Entsprechungen, die sich nicht in gleicher Reihenfolge im Umfeld von Matches mit der geforderten Mindestlänge befinden, ist das hier vorgestellte Verfahren naturgemäß nur sehr begrenzt geeignet. Dementsprechend konnte die Ermittlung der Passagen des *Ingangs*, die dem unten noch etwas näher betrachteten Titel 2.10 der *Institutiones* zuzuordnen sind, teilweise nur durch einen nachträglichen Abgleich erfolgen.

Murner,Inst. 1519 (Zählung nach CICiv. (Krüger/Mommsen))	**Murner,KaisStatR. 1521 (Bogensignatur)**
[Murners Reimvorrede]	[A] r
1.1.2 – 1.2.1	A ii r – A iii r
1.2.3 – 1.2.8	[A iiii] r – B r
1.3.2 – 1.5.1	B v – B ii r
1.8 – 1.9.3	C r – C ii v
1.10 – 1.10.3	C iii r – [C iiii] v
1.12.8 – 1.12.9	D iii v
1.16.7	E iii r
1.17	E iii v
1.19 – 1.20 Überschrift	[E iiii] r
1.22.1 – 1.24.1	F ii r – F iii v
[Abschluss Buch 1] – 2.1.12	[G iiii] r – H ii r
2.1.12 – 2.1.13	H iii r – H iii v
2.1.14 – 2.1.16	H ii r – H iii r
2.1.20 – 2.1.25	[H iiii] r – I r
2.1.27 – 2.1.37	I ii r – [I iiii] r
2.4.2 – 2.5.6	K iii r – [K iiii] v
2.6.pr. – 2.6.3	L r – L ii r
2.6.4 – 2.6.6	L iii r – L iii r
2.6.7	L iii v
2.7.1	[L iiii] r
2.7.2 – 2.8.pr.	[L iiii] r – M ii r
2.8.2	M ii v – M iii r
2.9 Überschrift – 2.9.5	M iii r – N r
2.10.1 – 2.10.5	Q iii v – [Q iiii] r
2.10.6	R r
2.10.7 – 2.10.8	R r
2.10.9 – 2.10.10	R r – R v
2.10.11	R r
2.10.12 – 2.10.14	[Q iiii] v
2.11.pr.	R v – R ii r
2.11.1 – 2.11.2	R ii v – R iii r
2.12.1 – 2.12.5	R iii v – [R iiii] r
2.13.1 – 2.14.pr.	S v – S iii r
2.14.1	S iii v
2.14.5	S iii v
2.15.pr. – 2.16.1	T r – T ii r
2.17.3 – 2.18.1	[T iiii] r – V v
2.18.1 – 2.19.7	V ii r – [V iiii] v
2.20.5	N iii v
2.20.6	[N iiii] v – O r
2.20.22 – 2.20.23	O iii r
2.20.25 – 2.20.33	N ii r – N iii v
2.20.36 – 2.24.3	[O iiii] r – Q iii v

Murner,Inst. 1519 (Zählung nach CICiv. (Krüger/Mommsen))	**Murner,KaisStatR. 1521 (Bogensignatur)**
2.25.pr. – 2.25.2	X r – X v
3.1.9 – 3.1.14	X ii v – [X iiii] r
3.2.4 – 3.2.6	[Y iiii] v – Z r
3.5.pr. – 3.6.4	Z v – [Z iiii] r
3.6.4 – 3.15.1	[Z iiii] r – c iii v
3.15.1	c iii v – [c iiii] r
3.15.1	[c iiii] r
3.15.2 – 3.19.2	[c iiii] r – d iii r
3.19.3 – 3.19.4	d iii r – d iii v
3.19.5	[d iiii] v – e r
3.19.6 – 3.19.10	d iii v – [d iiii] r
3.19.11	e r
3.19.12	[d iiii] r
3.19.13 – 3.19.17	e r – e v
3.19.18 – 3.19.19	[d iiii] v
3.19.20	[d iiii] r
3.19.22 – 3.19.25	d iii r
3.19.25 – 3.25.3	e v – f iii v
3.25.4 – 3.25.8	[f iiii] r
3.25.9	f iii v
3.26.pr. – 3.26.13	[f iiii] v – g ii r
3.26.13 – 4.1.6	[g iiii] r – i r
4.6.40 – 4.7.2a	g ii r – g ii v
4.7.3 – 4.7.4b	g ii v – g iii r
4.7.5	g ii v
4.7.4c	g iii v
4.7.5a	g iii r – g iii v
4.7.6 – 4.7.7	g iii v
4.1.7 – 4.1.8	i v
4.1.8 – 4.1.11	i r – i v
4.1.11 – 4.1.12	i v – i ii r
4.1.13 – 4.1.16	i ii r – i iii r
4.1.17	i ii r
4.1.18 – 4.3.16	i ii r – k ii v
4.4.pr. – 4.6.31	[k iiii] r – m iii v
4.6.33 – 4.6.35	n v – n ii v
4.6.36 – 4.6.40	[m iiii] r – [m iiii] v
4.8.pr. – 4.9.1	k ii v – [k iiii] r
4.9.1 – 4.11.5	[n iiii] r – o v
4.11.5 – 4.12.1	[m iiii] v – n r
4.12.2 – 4.14.4	o v – o iii v
4.14.4 – 4.15.7	n ii v – n iii v
4.15.7 – 4.18.12	o iii v – [o vi] r

Tab. 4.1: Übernahmen aus Murner,Inst. 1519 in Murner,KaisStatR. 1521

TEstamentū ist ein latinsch wort vñ lut so vill zů dütsch alß ein kuntschafft des gemůts. Dz aber nüt der alten rechtẽ vnterlassen werdẽ ist zů wyssen dz vor zytẽ nit mer deñ zweyerley geschlecht der testamẽtẽ in dẽ bruch gewesen sint. vß welchẽ sy dz ein in růw vñ dẽ frydẽ gebrucht habẽ / dz sy nantẽ der gůtẽ versamlung testamẽt. das ander so sy in ein krieg vßziehẽ woltẽ das das gehelich genãt was.	DIe altẽ habend vor zeiten drierlei testament gehabt / dz erst des sie sich in dẽ friden / vnd in in růwen gebruchtẽ Das ander des sy sich in kriegs leüffen bruchtẽ vñ heiß dz gehe vñ schnel testamẽt
Es ist zů letst dz drit dar zů kummẽ das do genãt was durch gelt vnd das gewicht darũb das es durch frylassung vñ erdichtes verkauffen geschahe mit funff zügen / vnd des gewichtes schetzer in gegẽwurdt der Rőmschẽ burger die über die ior worẽdt vnd des der do genant was ein keuffer des geschlechts. Aber die zwey ersten geschlecht der testamẽt in vergangenen alten zyten sindt in ein mißbruch vnd ablassung kũmen / das aber durch gelt vnd gewicht geschahe / wie wol es lenger deñ die andren geweret hat / noch dẽnocht het es eins deyll vff gehőret in dem bruch zu syn. Aber die ob genanten nãmen der testament wardendt dem statt recht heim erkant. dar nach vß vßspruch des richters ist ein andre form vnd gestalt testament zů machẽ erfundẽ wordẽ. Deñ vß des richters pretoris rechtẽ wardt kein frylassung vß vătterlichem gewalt erfordret / sunder es was genůg das syben zügen versigleten / welche zeichẽ oder sygil mit dem statt recht nit not warẽ.	Dz drit ward genãt mit gelt vñ gewiht erkauffet das ist durch hãdschlagũg vß vetterlichẽ gewalt / vñ geschahe mit einẽ erdichtẽ verkauffen in gegẽ würd funffzügen / vñ des gewicht meisters rőmschen burgerẽ vñ vber die iar / vñ auch des dẽ mã nẽnet eĩ keüffer eĩs fremden gesundes Aber die zwei ersten testamẽt sind veraltet vñ durch die lenge d' zeit abgangẽ Dz aber mit dẽ gelt vñ gewiht beschahe wie wol es am lengsten geweret hat / ist es doch eĩs deils vß dẽ bruch kũmen / vñ geschahẽ solche testamẽt in krafft des statrechtẽ Aber nachgons vß dẽ spruch des richters ist eĩ ander frõ testamẽt zů machẽ erfůden worden / dẽ mit solchẽ richterlichẽ rechtẽ ward keĩ von handlassen erfordret sunder es worẽd genůg sibẽ sigel d' zügen / so doch mit statrecht keĩ sigil not was der zügen.

Tab. 4.2: Synopse Murner,Inst. 1519, II 10 [1. Abschnitt] (Inst. 2.10.pr.-2.10.2) – Murner,KaisStatR. 1521 IV [1. Abschnitt]

zur Verdeutlichung unterstrichen und ähnliche Wortformen durch Wellenlinien gekennzeichnet; eine unterbrochene Linie markiert eine wörtliche Übereinstimmung, die in den beiden einander entsprechenden Passagen etwas unterschiedlich positioniert ist.[691]

Aus der Synopse ist leicht zu entnehmen, dass dieser Abschnitt des *Ingangs* einerseits auf der Vorlage der *Instituten* – oder auch auf dem zugrunde liegenden lateinischen Original – basiert, aber andererseits doch auch erhebliche Textänderungen aufweist. So wird die historische Darstellung am Anfang etwas vereinfacht, indem das Testament *per aes et libram* mit den beiden älteren Testamentsformen zusammengefasst wird, während es in den *Institutiones* als spätere Entwicklung von diesen abgegrenzt wird. Auch die übrigen Abschnitte von Titel 2, 10 der *Institutiones* haben eine Entsprechung im *Ingang*. Allerdings hat Murner dabei die Textreihenfolge stark geändert.[692]

[691] Diese Synopse beruht zwar weitgehend auf einer automatischen Analyse mit dem hier vorgestellten Instrumentarium, sie wurde aber an einigen Stellen etwas überarbeitet, um eine möglichst schlüssige Zuordnung zu erreichen.

[692] Dies geht aus Tabelle 4.1 nur teilweise klar hervor, da dort für den *Ingang* jeweils die Druckseite genannt wird; es ist aber durch die Aufteilung auf mehrere Tabellenzeilen angedeutet.

4.3 Normtexte mit mehreren Vorlagen

4.3.1 Die Heilbronner Statuten von 1541

Die 1541 publizierten *Statuten / Satzung / Reformation vnd Ordnung / Burgerlicher Pollicey des Heyligen Reychßstat Haylpronn* sind ein für die Rechtsgeschichte dieser Stadt zentraler Text.[693] Anscheinend sind sie aber bisher weder inhaltlich noch in ihren Abhängigkeitsverhältnissen näher untersucht worden.

Die Angaben zum Inhalt beschränken sich in der hier herangezogenen Literatur auf Übersichten über die Sachbereiche, die in den zehn Teilen der Statuten behandelt werden[694] beziehungsweise auf eine Beschreibung der Inhalte „einzelne[r] und wichtiger erscheinende[r] Bestimmungen", wobei die Kriterien für die Auswahl ungenannt bleiben[695].

Zur Entstehung der Statuten ist bekannt, dass sie in dieser Form vom städtischen Syndikus Dr. iur. utr. Jakob Ehinger[696] auf der Basis älterer Verordnungen erstellt beziehungsweise redigiert wurden. Eine Beteiligung von Ulrich Zasius[697] wird in der Literatur erwähnt, aber wohl nicht belegt oder näher beschrieben.[698]

Als Textvorlage werden insbesondere die Heilbronner Statuten von 1513 genannt, auf denen die Teile II-VI der Statuten von 1541 „fussen" sollen, sowie ältere Ratsverordnungen.[699] Carl Georg von Wächter schreibt: „Sichtbar ist das Statut zum Theile aus denselben Quellen unmittelbar geschöpft, welche bei unsrem ersten Landrechte benützt wurden", führt das aber nicht weiter aus.[700]

Die Statuten von 1513 gehören zwar eigentlich nicht zum hier ausgewerteten Korpus, da sie aber ebenfalls in digitaler Form vorliegen,[701] konnten sie in diese spezielle Untersuchung mit einbezogen werden.

Auch hier soll zunächst der Projektionsdotplot in Abbildung 4.5 einen Überblick verschaffen. Im Interesse der Übersichtlichkeit sind dabei nur die Texte berücksich-

[693] Sie behielten nach der hier herangezogenen Literatur „im wesentlichen, d. h. mit nur geringfügigen Änderungen und Ergänzungen" ihre Geltung, solange Heilbronn als Reichsstadt eigenständig war (so Schrenk/Weckbach 1993, S. 147 – Online-Version: S. 122 f. –, ähnlich Jäger 1828, S. 133).

[694] So in Wächter 1839, S. 691, Anm. 22 und OAHeilbronn^2 I, Pag. 1, S. 171. Die Gliederungsübersicht zu diesem Text in DRQEdit (*http://drw-www.adw.uni-heidelberg.de/drqedit-cgi/zeige?term=HeilbronnStat.%201541&index=glueb*) ermöglicht eine Orientierung über die Inhalte der einzelnen Titel.

[695] Jäger 1828, S. 133–138.

[696] Kurze biographische Hinweise finden sich in Schrenk/Weckbach 1993, S. 16 (Online-Version: S. 8)

[697] Vgl. über ihn zum Beispiel Sturm 1996.

[698] Jäger 1828, S. 133; Titot 1865, S. 222; Schrenk/Weckbach 1993, S. 147 (Online-Version: S. 123).

[699] So die einleitenden editorischen Bemerkungen zu den Statuten von 1513 in UBHeilbronn III, S. 337. Dieser Beschreibung entspricht auch die Darstellung in Schrenk/Weckbach 1993, S. 147 (Online-Version: S. 122 f.).

[700] Wächter 1839, S. 691, Anm. 22.

[701] *http://repertorium.at/qu/1513_heilbronnstat.html.*

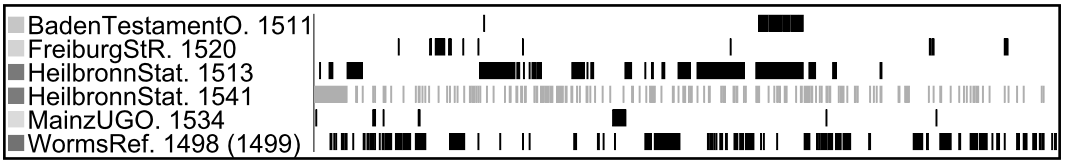


Abb. 4.5: Projektionsdotplot zu HeilbronnStat. 1541 (nur ausgewählte Vergleichstexte berücksichtigt)

tigt, die nach einer Gesamtauswertung signifikante Ähnlichkeiten enthalten und nach der Datierung als Quelle in Frage kommen.

Es ist gut zu erkennen, dass die Statuten von 1513 tatsächlich für einen recht großen Teil der Statuten von 1541 als Vorlage gedient haben. Dies gilt aber offenbar nicht für den gesamten Bereich der Teile II-VI; vielmehr finden sich darin auch etwas längere Abschnitte, die nicht auf diese Quelle zurückzuführen sind. Am umfangreichsten sind daneben offenbar die Übernahmen aus der *Wormser Reformation* von 1498; außerdem gibt es an einzelnen nicht ganz kurzen Stellen Entsprechungen zum Freiburger Stadtrecht von 1520, zur Mainzer Untergerichtsordnung von 1534 sowie – in einem Bereich, der auch weitgehend mit den Statuten von 1513 übereinstimmt – zur Badischen Testamentsordnung von 1511.

Die Tabellen 4.3 und 4.4 ordnen die einzelnen Titel der Statuten von 1541 den verschiedenen Quellen zu.[702] Wie stark der jeweilige Übereinstimmungsgrad ist und inwieweit die genannten Titel auch Passagen enthalten, die ohne Entsprechung sind, ist dabei recht unterschiedlich und aus dieser Darstellung in der Regel nicht zu entnehmen. Wenn ein Titel allerdings signifikante Entsprechungen zu mehreren Vergleichstexten aufweist, sind die Zeilen durch graue Schrift zurückgenommen, die offenbar nicht einem direkten Abhängigkeitsverhältnis entsprechen, da die betreffende Passage deutlich stärkere Übereinstimmungen mit einem anderen Text aufweist.

Insgesamt ist festzuhalten, dass Abschnitte aus den Statuten von 1513 oft mit nur geringen Änderungen übernommen wurden, während der Vergleich mit den übrigen Texten vielfach größere Unterschiede erkennen lässt. Dies trifft allerdings nicht in jedem Fall zu. So entspricht Teil IV, Titel 5 fast vollständig dem Text von Buch IV, Teil 2, Titel 6 der *Wormser Reformation*, lässt sich aber nur zu einem kleinen Teil im vierten Buch der Statuten von 1513 wiederfinden, die offenbar ebenfalls teilweise auf der *Wormser Reformation* basieren.

Diese Übersichten bestätigen den Eindruck, der sich aus dem Projektionsdotplot ergibt. Die genannten Quellen – oder solche, die ihnen in den angeführten Abschnitten nahe stehen – wurden bis auf die Badische Testamentsordnung offenbar tatsächlich bei der Formulierung der Heilbronner Statuten von 1541 herangezogen. Und auch wenn es keinen Anhaltspunkt dafür gibt, dass die Badische Testaments-

[702] Auch hier wurden die automatisch ermittelten Daten von Hand überarbeitet, um inhaltlich möglichst präzise zu sein.

HeilbronnStat. 1541	**weitere Texte**
Vorrede	HeilbronnStat. 1513 Vorrede
I 1	WormsRef. 1498 (1499) I 3
I 2	WormsRef. 1498 (1499) I 25
I 3 – I 4	WormsRef. 1498 (1499) III 1, 33 – III 1, 34
I 5	WormsRef. 1498 (1499) I 23
I 6 – I 7	WormsRef. 1498 (1499) I 22
I 10	FreiburgStR. 1520 I 3
I 11 – I 13	FreiburgStR. 1520 I 9
I 15	FreiburgStR. 1520 I 8
I 16 – I 17	WormsRef. 1498 (1499) II 1 – II 2
I 16	FreiburgStR. 1520 I 11
II 1 – II 4	HeilbronnStat. 1513 II
II 1	BadenTestamentO. 1511 Art. 10
III 1 – III 9	HeilbronnStat. 1513 I
III 12 – III 13	HeilbronnStat. 1513 I
III 12	WormsRef. 1498 (1499) IV 4, 9
III 16 – III 17	HeilbronnStat. 1513 I
III 18	WormsRef. 1498 (1499) V 5, 1
III 20 – III 21	MainzUGO. 1534 [Tit. 44]
III 22	HeilbronnStat. 1513 I
IV 1	HeilbronnStat. 1513 IV
IV 1	WormsRef. 1498 (1499) IV 2, 4
IV 2	WormsRef. 1498 (1499) IV 2, 2
IV 3	WormsRef. 1498 (1499) IV 2, 5
IV 4	WormsRef. 1498 (1499) IV 2, 3
IV 5	WormsRef. 1498 (1499) IV 2, 6
IV 5	HeilbronnStat. 1513 IV
IV 6	WormsRef. 1498 (1499) IV 2, 8
V 1	HeilbronnStat. 1513 III
V 3 – V 9	HeilbronnStat. 1513 III
V 3 (Schluss)	WormsRef. 1498 (1499) IV 2, 9
V 4 – V 9	WormsRef. 1498 (1499) IV 3, 4 – IV 3, 11
V 11	WormsRef. 1498 (1499) IV 3, 14 und IV 3, 12 (Einschub)
VI 1 – VI 4	HeilbronnStat. 1513 V
VI 2 – VI 4	BadenTestamentO. 1511 Art. 6 – Art. 8
VI 1 – VI 3	WormsRef. 1498 (1499) IV 4, 1
VI 4 (Arbor Consanguinitatis)	WormsRef. 1498 (1499) IV 4, 1
VI 5	WormsRef. 1498 (1499) IV 4, 2
VI 6	WormsRef. 1498 (1499) IV 4, 4
VI 7	WormsRef. 1498 (1499) IV 4, 3
VII 2	HeilbronnStat. 1513 I
VII 3 – VII 4	WormsRef. 1498 (1499) V 1, 1
VII 6	WormsRef. 1498 (1499) V 1, 6
VII 7	WormsRef. 1498 (1499) VI 1, 10
VII 13	WormsRef. 1498 (1499) III 1, 13

Tab. 4.3: Übereinstimmungsbereiche zu HeilbronnStat. 1541, Teil 1–7

HeilbronnStat. 1541	weitere Texte
VIII 8	WormsRef. 1498 (1499) V 3, 1
VIII 9	WormsRef. 1498 (1499) III 2, 35 – III 2, 36
VIII 10	FreiburgStR. 1520 I 14
VIII 11	FreiburgStR. 1520 II 9
VIII 12	MainzUGO. 1534 [Tit. 34]
IX 1 – IX 2	WormsRef. 1498 (1499) V 4, 1
IX 4	WormsRef. 1498 (1499) V 4, 1
IX 8 – IX 11	WormsRef. 1498 (1499) V 4, 12 – V 4, 15
IX 12	WormsRef. 1498 (1499) V 4, 19; V 4, 18
IX 13	WormsRef. 1498 (1499) V 4, 25
IX 14	WormsRef. 1498 (1499) VI 1, 17
IX 17	WormsRef. 1498 (1499) VI 1, 14
IX 20 – IX 21	FreiburgStR. 1520 IV 1
X 2	WormsRef. 1498 (1499) VI 1, 5
X 4	WormsRef. 1498 (1499) VI 1, 2
X 6	WormsRef. 1498 (1499) VI 1, 21
X 7	WormsRef. 1498 (1499) VI 1, 16
X 9	WormsRef. 1498 (1499) V 5, 1
X 11	WormsRef. 1498 (1499) VI 2, 21 – VI 2, 21

Tab. 4.4: Übereinstimmungsbereiche zu HeilbronnStat. 1541, Teil 8–10

ordnung selbst benutzt worden ist, zeigt sich doch klar eine Traditionslinie, die auf sie zurückzuführen ist.

Daneben ist festzustellen, dass sich zwar die einzelnen Titel in der Regel recht klar einer bestimmten Vorlage zuordnen lassen, dass aber zwischen den Titeln öfters ein Wechsel des Ausgangstexts erfolgt. Auch ohne nähere Prüfung der jeweils ausgewählten Quellen lässt sich daraus wohl schließen, dass die Statuten trotz ihrer zu erheblichen Teilen unselbständigen Formulierung durchaus auf einem sorgfältigen Redaktionsprozess beruhen.

Dafür ein Beispiel: Die Titel 20 und 21 von Teil III treffen Regelungen für die sogenannte Einkindschaft, durch die bei Wiederverheiratung verwitweter Eltern deren Kinder aus erster Ehe mit Kindern aus der neuen Ehe erbrechtlich gleichgestellt werden konnten.[703] Die Statuten von 1513 gehen auf die Möglichkeit der Einkindschaft anscheinend nicht ein, so dass bei der Redaktion der Fassung von 1541 offenbar Ergänzungsbedarf gesehen wurde. Dabei wurde aber nicht die *Wormser Reformation*, also die andere Hauptquelle, herangezogen, die in Buch IV, Teil 4, Titel 4 und Buch V, Teil 5, Titel 4 die Einkindschaft behandelt, sondern die Mainzer Untergerichtsordnung, die nur noch an einer weiteren Stelle als Vorlage gedient hat. Dies erklärt sich wohl dadurch, dass die Untergerichtsordnung nicht nur die Rechtswirkungen der Einkindschaft darlegt, sondern insbesondere auch Regelungen trifft, um die Kinder aus erster Ehe vor einer Einkindschaft, die zu ihrem Nachteil wäre, zu schützen.

[703] Vgl. Lipp 2007.

4.3.2 Die Henneberger Landesordnung

Die 1539 publizierte Landesordnung der Grafschaft Henneberg[704] wird in der wissenschaftlichen Literatur wiederholt als fast völlig abhängig von der Tiroler Landesordnung von 1532 beschrieben. Diese Darstellung geht wohl auf eine 1831 erschienene Schrift von Karl Ernst Schmid[705] zurück,[706] in der die Nähe beider Texte auf der Basis eines kleinen Textausschnitts gezeigt und darüber hinaus erklärt wird, dass der Verfasser der Landesordnung, der hennebergische Kanzler und promovierte Jurist Johann Gemel[707], bis auf Änderungen in der Wortwahl und gewisse inhaltliche Anpassungen die Tiroler Vorlage weitgehend übernehme:

> Eandem materiarum dispositionem, singulos autem articulos Gemelius ita secutus est, ut plerumque nil nisi verba quaedam mutaret, orationemque verbosiorem redderet. Omisit potissimum quae in Tyrolensi ordinatione de iure nobilium et praelatorum, de cultura vinearum, de venatione dicta sunt aliaque ad politiam pertinentia; pauciora addidit, quae iuris in Comitatu Hennebergico iam antea constituti ratio suadebat.[708]

Dementsprechend heißt es in der wissenschaftlichen Literatur verschiedentlich, die Henneberger Landesordnung sei eine „Umarbeitung“[709] oder „Paraphrase“[710] der Tiroler Landesordnung von 1532 oder sie bestehe „in einer fast wörtlichen Wiederholung“ dieser Quelle[711]. Albert Unger hat allerdings in einem 1889 erschienenen Handbuch festgestellt, dass die Ähnlichkeit beider Texte keineswegs so ausgeprägt ist:

> Wer beide Landesordnungen genau vergleicht, erhält den Eindruck, daß Satz für Satz der Tiroler LO. von den Urhebern der Hennebergischen geprüft, und diejenigen Sätze, deren Inhalt auf die Hennebergischen Verhältnisse paßte oder sich denselben anpassen ließ, entweder wörtlich oder mit größeren oder geringeren Abänderungen in die Hennebergische aufgenommen wurden, besonders solche Sätze, welche exaktes Verfahren vorschrieben oder Verdeutlichungen römischen Rechts enthielten, daß dabei aber mit Vorsicht verfahren, große Theile der Tiroler LO. nicht angenommen sind, und neben dem herübergenommenen Stoff die Henneb. LO. noch eine solche Menge

[704] Vgl. zu ihrem Inhalt Simon 1898. Die Bestimmungen der Landesordnung hatten teilweise bis zur Einführung des Bürgerlichen Gesetzbuchs im Jahr 1900 Rechtsgeltung (ebd. S. 29; vgl. zum Geltungsbereich Unger 1889, S. 94 f.).

[705] Es handelt sich um die Einladungsschrift zur Promotion von August Heinrich Emil Danz (Schmid 1831). Auf dem Titelblatt wird Schmid zwar nicht als Verfasser, sondern als einladender Dekan benannt, es handelt sich aber nicht um die Dissertation des Kandidaten, die ein anderes Thema hatte (vgl. ebd. S. 22 f.).

[706] Diese Quelle wird jedenfalls von Gerber 1846, S. 183, Anm. 55 und von Unger 1889, S. 91 genannt. Die übrigen im Folgenden genannten entsprechenden Äußerungen enthalten keine Hinweise, woher die Information stammt.

[707] Vgl. über ihn Unger 1889, S. 90 (dort weitere Literaturhinweise).

[708] Schmid 1831, S. 15 f.

[709] Gerber 1846, S. 184.

[710] Motloch 1907, S. 353.

[711] Stobbe 1864, S. 218 (ähnlich Krohn 2008 S. 467 mit Verweis auf Stobbe).

> von Bestimmungen, welche theils Hennebergisches Gewohnheitsrecht, theils eigene Gedanken der Urheber wiedergeben, enthält, daß dieselbe im Ganzen nicht als unselbständige Arbeit erscheint.[712]

Diese Darstellung passt gut zu den Ergebnissen eines automatisierten Vergleichs. Tatsächlich lassen sich zahlreiche wörtliche Übereinstimmungen zwischen den beiden Landesordnungen feststellen, sie sind aber keineswegs so umfassend, wie nach der oben zitierten Hochschulschrift zu vermuten wäre. Daneben weisen noch einige weitere ältere Texte in bestimmten Textbereichen signifikante Ähnlichkeiten auf und kommen deshalb als Vorlage in Betracht, nämlich die Bayrische Gerichtsordnung von 1520 sowie die davon abhängige Passauer Gerichtsordnung von 1536, die Reichsnotariatsordnung von 1512, die Reichskammergerichtsordnungen von 1495 und von 1521 (zum Teil in Überschneidung mit der Bayrischen Gerichtsordnung) und schließlich die Tiroler Landesordnung von 1526, die durch die Tiroler Landesordnung von 1532 ersetzt wurde und mit dieser teilweise wörtlich übereinstimmt.[713]

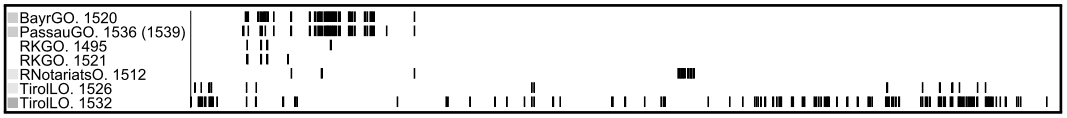


Abb. 4.6: Projektionsdotplot zu HennebergLO. 1539 (nur ausgewählte Vergleichstexte berücksichtigt)

Der Projektionsdotplot in Abbildung 4.6 ist auf die genannten Texte beschränkt. Die ersten beiden Zeilen sind der Bayrischen und der Passauer Gerichtsordnung zuzuordnen, danach folgen die beiden Reichskammergerichtsordnungen, die Reichsnotariatsordnung und schließlich die beiden Tiroler Landesordnungen.

Es ist gut zu erkennen, dass die Passauer Gerichtsordnung und die Tiroler Landesordnung von 1526 wohl kaum als Quellen für die Henneberger Landesordnung genutzt worden sein dürften, da die Übereinstimmungen mit diesen Texten in Bereichen liegen, die auch Übereinstimmungen mit der Bayrischen Gerichtsordnung beziehungsweise der Tiroler Landesordnung von 1532 aufweisen und sich in den beiden zuletzt genannten Texten noch wesentlich mehr Entsprechungen finden.

Dies ist allerdings kein zwingendes Argument: Auch die Übereinstimmungen mit den beiden Reichskammergerichtsordnungen fallen fast vollständig in Bereiche, in denen sich auch Matches mit der Bayrischen Gerichtsordnung finden. Wie der Vergleich von Titel II 2, 11 der Henneberger Landesordnung mit Titel 3, 4 der Bayrischen Gerichtsordnung sowie mit der Reichskammergerichtsordnung von 1495 aber zeigt, folgt die Henneberger Landesordnung zwar auch hier weit-

[712] Unger 1889, S. 92.

[713] Daneben gibt es auch Übereinstimmungen zu späteren Texten, die in Anbetracht der Frage nach den Quellen hier nicht weiter betrachtet werden sollen.

Den vierdten weg approbiert des heyligen Reychs Ordnung / der auch den gemeynen Rechten gemeß / Das der Richter auff des gehorsamen teyls anrůffen / kundtschafft vnd ander fůrbrengen hǒren vnd volfaren / vnnd entlich vrteyl darauff geben soll / Vnd so als dann für den vngehorsamen teyl geurteylt / soll doch der gehorsam kleger der kosten vnd scheden entlediget / vnd seines gehorsams genießlich sein.	Der vierdt weg / den des heyligen Reychs Camergerichts ordnung zůläßt / vnd dem rechten gemäss ist. Nemlich das auf begern des clagers / der richter / demselben Clager / sein kundtschafft vnd ander sein fürpringen / hǒr{e}n / vnd damit in allen terminen als ob sein widertail entgegen wäre volfarn sol lassen / vnd darnach enntlich vrteil darauff geben. Vnd steet in des Clagers willen / der oberzelt{e}n weg ainen fürzenemen / vnd wie auf dieselb{e}n drey weg / jr yeden sonderlich / in recht sol verfarn werden. Ist in hernachuolgenden gesatzen diss Tittls außgedruckt.

(a) Synopse mit BayrGO. 1520, Tit. 3, 4 [§ 5]

Den vierdten weg approbiert des heyligen Reychs Ordnung / der auch den gemeynen Rechten gemeß / Das der Richter auff des gehorsamen teyls anrůffen / kundtschafft vnd ander fůrbrengen hǒren vnd volfaren / vnnd entlich vrteyl darauff geben soll / Vnd so als dann für den vngehorsamen teyl geurteylt / soll doch der gehorsam kleger der kosten vnd scheden entlediget / vnd seines gehorsams genießlich sein.	Wurde auch der antwurter in der Ersten rechtuertigung oder in der Appellacion sach vor be"=ues"=ti"=gũg des kriegs vngehorsam so solte doch auf des clagers anrůeffen . durch das gerichte zů der Acht vnd aberacht . auch zů dem einsatze ex primo decreto wider den vngehorsamen antwurter procedirt werden oder soll das gericht auf begern des clagers kuntschafft vnd ander fůrbringen hǒren vnd volfaren vñ endtlich vrteil geben . welchen weg der clager fürnemen wirdet vnd ob fůr den vngehorsamen teil vrteil gesprochenn wurde so soll doch der gehorsam clager . der Cost vnd scheden entledigt werden.

(b) Synopse mit RKGO. 1495 [Art. 23]

Tab. 4.5: Synopsen zu HennebergLO. 1539, II 2, 11 [§ 5]

gehend der bayrischen Vorlage, ergänzt aber am Ende eine Formulierung, die aus der Reichskammergerichtsordnung stammt (auf die die Bayrische Gerichtsordnung ausdrücklich Bezug nimmt). Die Tabellen 4.5a und 4.5b zeigen den betreffenden letzten Absatz dieses Kapitels der Henneberger Landesordnung jeweils in Synopse mit einer der beiden Vorlagen, auch hier mit Kennzeichnung von wörtlichen Übereinstimmungen durch Unterstreichung und von fast übereinstimmenden Wortformen durch Wellenlinien. Nach diesem Vergleich ist davon auszugehen, dass die Reichskammergerichtsordnung von 1495 – oder ein von ihr abhängiger, hier nicht berücksichtigter Text – als Quelle für die Henneberger Landesordnung verwendet wurde.

Insbesondere ist im Projektionsdotplot zu erkennen, dass sich die Übereinstimmungen mit der Bayrischen Gerichtsordnung, der Reichsnotariatsordnung und der Tiroler Landesordnung von 1532 nicht oder kaum überlappen und dass auch die

Zeile, die die Entsprechungen mit dem zuletzt genannten Text repräsentiert, recht große Lücken aufweist. Auch wenn prinzipiell natürlich zu berücksichtigen ist, dass das hier gewählte Erkennungsverfahren für die Ermittlung von stark verändernden Übernahmen nur sehr begrenzt geeignet ist, ist es wohl recht offenkundig, dass man für die Bereiche, in denen sich Entsprechungen mit anderen Texten konzentrieren, aber keine Übereinstimmungen mit der Tiroler Landesordnung von 1532 zu finden sind, kaum davon ausgehen kann, dass dieser Text zugrunde liegt. Das legt die Vermutung nahe, dass die Henneberger Landesordnung auch ansonsten keineswegs durchgängig auf diese Vorlage zurückzuführen ist, sondern vielmehr mit einem ganz erheblichen Anteil von neu verfassten Abschnitten zu rechnen ist oder auch mit der Benutzung weiterer Quellen, die nicht im hier ausgewerteten Korpus enthalten sind. Dabei mögen natürlich insbesondere die von Unger erwähnten Henneberger Rechtstraditionen eine Rolle spielen.

Auch hier soll in Tabelle 4.6 konkreter angegeben werden, welche Abschnitte sich mit einer gewissen Plausibilität welcher Vorlage zuordnen lassen, und auch hier erscheint die Schrift in grau, wenn allem Anschein nach kein direktes Abhängigkeitsverhältnis vorliegt, sondern vielmehr eine Vermittlung über einen anderen Text (in diesem Fall anscheinend die Bayrische Gerichtsordnung) erfolgt ist.

Die Tabellen lassen allerdings nicht erkennen, dass sich häufig in den betreffenden Abschnitten nur relativ kurze Stücke entsprechen. Und insbesondere ist zu betonen, dass auch für die Teile III-VIII der Henneberger Landesordnung, für die fast nur Übernahmen aus der Tiroler Landesordnung von 1532 verzeichnet sind, keineswegs von einer umfassenden Abhängigkeit von dieser Quelle auszugehen ist, sondern vielmehr von einem weitgehend neu – oder auf der Basis anderer Quellen – verfassten Text. Das schließt natürlich nicht aus, dass trotz unterschiedlicher Formulierungen inhaltliche Beziehungen zwischen bestimmten Passagen bestehen mögen. Jedenfalls ist aber festzuhalten, dass die Henneberger Landesordnung zwar wie viele andere Normtexte der Zeit unter Verwendung von Vorlagen verfasst wurde, dass sie aber wesentlich eigenständiger ist, als es der nach allem Anschein bisher in der wissenschaftlichen Literatur überwiegenden Sicht entspricht.

HennebergLO. 1539	andere Texte
I 1, 1 – I 1, 2	TirolLO. 1532, I 1 – I 2
I 1, 3 – I 1, 4	TirolLO. 1532, I 3 – I 4
I 1, 3	TirolLO. 1532, I 5
I 2, 1 – I 2, 2	TirolLO. 1532, I 6
II 1, 2	BayrGO. 1520, Tit. 1, 2
II 1, 2	RKGO. 1495, [Art. 3]
II 1, 6	TirolLO. 1532, II 2
II 1, 8 – II 1, 14	BayrGO. 1520, Tit. 1, 5 – Tit. 1, 11
II 1, 9	RKGO. 1495, [Art. 5]
II 1, 9	RKGO. 1521, [Tit. 13]
II 1, 11	RKGO. 1495, [Art. 11]
II 1, 11	RKGO. 1521, [Tit. 20]
II 1, 15	TirolLO. 1532, II 14
II 1, 16	RKGO. 1521, [Tit. 28]
II 2, 1 (bis [§ 2])	BayrGO. 1520, Tit. 2, 1
II 2, 1 (ab [§ 5])-II 2, 2	TirolLO. 1532, II 20 – II 21
II 2, 5 – II 2, 6	BayrGO. 1520, Tit. 2, 3 – Tit. 2, 5
II 2, 7 – II 2, 8	BayrGO. 1520, Tit. 3, 1 – Tit. 3, 2
II 2, 9	BayrGO. 1520, Tit. 3, 13
II 2, 10 – II 2, 15	BayrGO. 1520, Tit. 3, 3 – Tit. 3, 8
II 2, 11	RKGO. 1495, [Art. 23]
II 2, 16	BayrGO. 1520, Tit. 3, 10
II 2, 19 – II 2, 20	BayrGO. 1520, Tit. 3, 14 – Tit. 3, 16
II 3, 2 – II 3, 3	BayrGO. 1520, Tit. 4, 3 – Tit. 4, 4
II 5, 15	TirolLO. 1532, II 47
II 7, 2	TirolLO. 1532, II 52
II 8, 1	TirolLO. 1532, II 66 – II 67
II 8, 8 – II 8, 11	TirolLO. 1532, II 80 – II 85
II 9, 5	TirolLO. 1532, VII 13
III 1, 1	TirolLO. 1532, VII 14
III 2, 1; III 2, 3	TirolLO. 1532, III 1
III 2, 6 – III 3, 1	TirolLO. 1532, III 2 – III 3
III 3, 2	RNotarariatsO. 1512, [Abs. 32]
III 3, 3	RNotarariatsO. 1512, [Abs. 31]
III 3, 4	RNotarariatsO. 1512, [Abs. 25] – [Abs. 26]
III 3, 5	RNotarariatsO. 1512, [Abs. 27] – [Abs. 30]
III 3, 6	RNotarariatsO. 1512, [Abs. 32] – [Abs. 33]

HennebergLO. 1539	andere Texte
III 3, 8	TirolLO. 1532, III 22
III 3, 13	TirolLO. 1532, III 43
III 4, 6	TirolLO. 1532, III 17
III 4, 10 – III 4, 12	TirolLO. 1532, III 18 – III 20
III 4, 13 – III 4, 14	TirolLO. 1532, III 23 – III 24
III 4, 16	TirolLO. 1532, III 26
III 4, 17	TirolLO. 1532, III 28
III 4, 18	TirolLO. 1532, III 27
III 4, 21	TirolLO. 1532, III 31
III 5, 1	TirolLO. 1532, III 35
III 6, 1 – III 6, 2	TirolLO. 1532, III 46
III 6, 5 – III 6, 8	TirolLO. 1532, III 48 – III 52
III 6, 10	TirolLO. 1532, III 53
III 6, 11 – III 6, 12	TirolLO. 1532, III 55 – III 56
IV 2, 1	TirolLO. 1532, IV 17
V 1, 1 – V 1, 2	TirolLO. 1532, V 2 – V 3
V 1, 4	TirolLO. 1532, V 7
V 2, 1	TirolLO. 1532, V 24
V 2, 2 – V 2, 3	TirolLO. 1532, V 26 – V 27
V 2, 6	TirolLO. 1532, V 39
VI 2	TirolLO. 1532, VI 26
VI 3, 2	TirolLO. 1532, VI 36
VI 4, 1 – VI 4, 2	TirolLO. 1532, VII 7 – VII 8
VI 4, 3 – VI 4, 4	TirolLO. 1532, VII 15 – VII 16
VII 1, 2	TirolLO. 1532, VIII 70
VII 1, 3	TirolLO. 1532, VIII 79
VII 1, 4 – VII 1, 6	TirolLO. 1532, VIII 2 – VIII 4
VII 2, 1 – VII 3, 1	TirolLO. 1532, VIII 8 – VIII 13
VII 3, 3	TirolLO. 1532, VIII 72 – VIII 73
VII 4, 2	TirolLO. 1532, VIII 78
VII 5, 1 – VII 5, 2	TirolLO. 1532, VIII 53 – VIII 54
VII 5, 3	TirolLO. 1532, VIII 80
VII 5, 5 – VII 5, 6	TirolLO. 1532, VIII 57 – VIII 58
VIII 1, 1	TirolLO. 1532, VII 2
VIII 3, 3 – VIII 3, 7	TirolLO. 1532, VIII 47 – VIII 51
VIII 4, 2 – VIII 4, 3	TirolLO. 1532, VIII 31 – VIII 33
VIII 4, 4	TirolLO. 1532, VIII 35 – VIII 36
VIII 4, 6 – VIII 4, 7	TirolLO. 1532, VIII 45 – VIII 46
VIII 5, 5 – VIII 5, 6	TirolLO. 1532, VIII 18 – VIII 19
VIII 7, 3	TirolLO. 1532, VIII 22
VIII 7, 5	TirolLO. 1532, VIII 26
VIII 7, 6 – VIII 7, 9	TirolLO. 1532, VIII 28 – VIII 30

Tab. 4.6: Übersicht Übereinstimmungen von HennebergLO. 1539 mit verschiedenen Quellen

Schluss

Die vorgelegte Untersuchung ist von einer zweifachen Zielsetzung geprägt: Zum einen ging es zunächst einmal konkret darum, wörtliche Übereinstimmungen im zugrunde gelegten Korpus zu ermitteln und Vorarbeiten für eine Analyse der sich darin zeigenden Textbeziehungen zu leisten. Zum anderen soll die Darstellung auf das Potential derartiger Analysen auch für andere Forschungsgebiete hinweisen und die für die Ermittlung und Auswertung der Übereinstimmungen genutzten und entwickelten Methoden vorstellen.

In Kapitel 1.1 wurden verschiedene geisteswissenschaftliche Forschungsgebiete betrachtet, in denen Abhängigkeitsverhältnisse zwischen Texten thematisiert werden. Offenkundig geht es dabei zum Teil um Beziehungen, die sich nicht oder nicht unbedingt anhand von wörtlichen Übereinstimmungen feststellen lassen. Aber es gibt doch verschiedene Anknüpfungspunkte, und in Unterkapitel 1.1.3 konnte im Überblick gezeigt werden, dass heutige Vorstellungen vom geistigen Eigentum nicht einfach auf frühere Zeiten übertragen werden sollten, dass die Verwendung von Textmaterial aus Vorlagen teilweise durchaus üblich war und dass auch jetzt nicht alle Texte unter urheberrechtlichem Schutz stehen. In diesem Zusammenhang wurden bereits verschiedene Textsorten genannt, bei denen eine EDV-gestützte Untersuchung von wörtlichen Übernahmen ergiebig sein könnte.

Dabei handelt es sich nur um Beispiele, die sich aus Hinweisen in der wissenschaftlichen Literatur ergeben haben. Es ist zu vermuten, dass sich bei umfassenderer Kenntnis noch vieles Weitere anführen ließe. Eine Untersuchungsmethode wie die hier vorgestellte eignet sich zudem auch für Texte, die in der Forschung noch nicht oder jedenfalls nicht näher betrachtet worden sind. Voraussetzung dafür ist freilich, dass die Texte in einer hinreichend guten maschinenlesbaren Fassung vorliegen. Soweit dies gegeben ist, lässt sich mithilfe einer automatisierten Auswertung ein anderer Überblick gewinnen als bisher, und daraus können sich auch andere Fragen und Schwerpunkte ergeben.

Die Ermittlung und Auswertung von wörtlichen Übereinstimmungen lässt sich, wie in Kapitel 2.5 dargestellt wurde, in die *Text-Reuse*-Forschung einordnen, setzt allerdings einen anderen Schwerpunkt als in vielen der diesem Bereich zuzurechnenden Projekte. In dieser Untersuchung geht es primär um im Hinblick auf den Wortlaut exakte Übereinstimmungen, die sich aufgrund ihrer Länge mit relativ großer Sicherheit auf die Verwendung einer schriftlichen Vorlage zurückführen lassen, obwohl sie in den meisten Fällen nicht als Übernahmen kenntlich gemacht sind. Es geht also nicht um Anspielungen, in erheblichem Maße umformulierte Passagen oder gar die Wiedergabe übernommener Ideen in neuer Textgestalt.

Um Textbeziehungen zu finden, die sich nur an solchen schwächer ausgeprägten Übereinstimmungen mehr oder weniger sicher erkennen lassen, sind andere Verfahren als ein auf der Ermittlung exakter Übereinstimmungen basierender Ansatz

sicherlich eher geeignet; sie setzen allerdings voraus, dass die untersuchten Texte in sprachlicher Hinsicht besser erschlossen sind, als es für das hier untersuchte Korpus und auch viele andere Texte insbesondere aus dem Mittelalter und der Frühen Neuzeit zutrifft.

Die Schwierigkeiten, die sich bei dem hier ausgewerteten Textmaterial ergeben, lassen sich ähnlich wohl auch für ältere Sprachstufen verschiedener anderer europäischer Sprachen feststellen; sie betreffen die Satz- und Wortabgrenzung und damit die Bildung grundlegender Analyseeinheiten (wobei die Wortabgrenzung immerhin in der weit überwiegenden Zahl der Fälle einheitlich ist), die Schreibungsvarianz, die oft sehr komplexe Syntax und das Fehlen eines einigermaßen umfassenden Wörterbuchs, anhand dessen eine regelbasierte Lemmatisierung überprüft werden und idealerweise sogar eine Zuordnung von bedeutungsverwandten Wörtern zueinander erfolgen könnte.

Hier wurde nicht der Versuch unternommen, im Hinblick auf die sprachliche Erschließung der Texte irgendwelche Verbesserungen zu erreichen, auch wenn über die Anwendung von Faustregeln vermutlich schon manches möglich wäre. Dieser Untersuchung liegen zwei Annahmen zugrunde, nämlich dass Vorlagen bei der Abfassung der hier untersuchten Texte im Regelfall ohne größere Änderungen in der Formulierung kopiert wurden, so dass die Ermittlung exakter Matches einer etwas größeren Länge eine gute Basis bietet, um textuelle Abhängigkeiten festzustellen, und dass das Hauptproblem bei der Erkennung dieser Übernahmen die Schreibungsvarianz ist.

Diese Annahmen treffen sicherlich nicht generell zu. Soweit das Ziel einer Untersuchung die Ermittlung von Beziehungen zwischen literarischen Werken ist, ist wohl mit wesentlich stärkeren sprachlichen Anpassungen oder überhaupt nur mit der Übereinstimmung einzelner Wörter zu rechnen, und Ähnliches gilt für die Ermittlung verschleierter Plagiate. Für die Entstehungszeit der Texte, um die es hier geht, ist aber nicht mit Verschleierungsmaßnahmen zu rechnen, weil das Kompilieren – nach allem, was sich sagen lässt – eine recht gängige Praxis war. Dementsprechend scheint es durchaus plausibel, auch für andere Sachtexte dieser Zeit unveränderte Übernahmen zu vermuten, soweit überhaupt Vorlagen als Basis für die Textproduktion herangezogen wurden.

Trotzdem lassen sich natürlich auch im hier ausgewerteten Material immer wieder kleinere Änderungen in der Formulierung feststellen. Dahinter kann der Wunsch stehen, etwas sprachlich besser zu gestalten oder Anpassungen aufgrund unterschiedlichen Sprachgebrauchs vorzunehmen, und es kann sich auch um Ungenauigkeiten beim Kopieren handeln, das ja nicht den Zweck hatte, die Vorlage exakt wiederzugeben, sondern auf den Wunsch zurückzuführen ist, eine als gut erkannte Formulierung zu verwenden. Und auch inhaltliche Anpassungen lassen sich nicht selten festellen.

Dass Änderungen der einen oder anderen Art die Erkennung von Übereinstimmungen mit dem hier eingesetzten Verfahren teilweise unterminieren, kann nicht ausgeschlossen werden. Für Texte, die bekanntlich eng verwandt sind, zeigen die vorgestellten Auswertungen aber, dass die Ermittlung exakter Übereinstimmungen insgesamt gut funktioniert, und auch für Textpaare, in denen sich nur für bestimmte Bereiche Übernahmen ermitteln lassen, haben sich exakte Übereinstimmungen als Grundlage für eine Einschätzung textueller Beziehungen und gegebenenfalls als Ausgangsbasis für eine Überprüfung ihres Umfelds auf weitere Entsprechungen bewährt.

Die für diese Untersuchung entwickelte Vorgehensweise basiert insbesondere auf der Kombination zweier Ansätze, nämlich der Ermittlung von *maximal exact matches (MEMs)* einer bestimmten Mindestlänge und der Eliminierung von Schreibungsunterschieden durch eine Abbildung der Texte auf Folgen von Codezeichen, die sich an den zu vermutenden Lautwerten der Buchstaben orientieren und unterschiedliche Laute auf der Basis phonetischer Ähnlichkeit zusammenfassen oder auch ganz streichen – ein Verfahren, das in der Tradition des bekannten *Soundex* steht.

Eine möglichst weitgehende Eliminierung von irrelevanten Unterschieden ist deshalb wichtig, weil es auf dieser Basis möglich ist, wörtliche Übereinstimmungen trotz solcher Unterschiede zu ermitteln, ohne bei jedem Vergleich zweier Textstellen prüfen zu müssen, ob eine Abweichung nach den angesetzten Kriterien akzeptabel ist oder gegen eine Einstufung als wörtliche Entsprechung spricht. Das reduziert die Komplexität enorm – die Ermittlung von MEMs ist mithilfe eines Suffixbaums, eines Suffixarrays oder einer in ähnlicher Weise funktionierenden ausgedünnten Datenstruktur möglich und kann auch für Textmengen, wie sie hier untersucht werden, sehr effizient durchgeführt werden.

Für die MEM-Ermittlung stehen verschiedene Programme zur Verfügung, die im Rahmen der Bioinformatik entwickelt wurden. Ihre Anwendung auf Textdaten ist bisher anscheinend etwas Neues und jedenfalls wohl insofern ungewöhnlich, als dabei nicht Wort-, sondern Zeichenfolgen ermittelt werden. Da aber die Wortabgrenzung in den hier untersuchten Textdaten nicht immer einheitlich ist, kann es gerade von Vorteil sein, dabei auch Abweichungen zu tolerieren. Wie in Unterkapitel 3.3.1 gezeigt wurde, lassen sich die Informationen zur Leerraumsetzung in einem Nachbearbeitungsschritt zur Präzisierung der Matchdaten nutzen.

Die für den Vergleich von Genomsequenzen entwickelten Programme eignen sich nur teilweise für den Vergleich von Textdaten, da diesen ein größeres Alphabet zugrunde liegt. Für vier der untersuchten Programme konnten die sich daraus ergebenden Probleme durch Anpassungen im Quellcode weitestgehend oder vollständig behoben werden; die dabei vorgenommenen Änderungen sind in Unterkapitel 3.2.1 dokumentiert.

Unterkapitel 3.2.2 zeigt die Ergebnisse von Tests zum Zeit- und Speicherbedarf dieser Programme, wenn sie für Textdaten eingesetzt werden. Grundsätzlich ist trotz der zugrunde liegenden effizienten Datenstrukturen mit Problemen insbesondere hinsichtlich des Speicherbedarfs zu rechnen, wenn es um wesentlich größere Korpora geht, allerdings wächst er bei Verwendung einer dieser Datenstrukturen nur im selben Verhältnis wie die Textmenge. Eine sinnvolle Korpuszusammenstellung ist also durchaus relevant, gegebenenfalls ist es aber zum Beispiel – mit steigendem Zeitaufwand – möglich, ein Korpus in Teilkorpora zu untergliedern und diese jeweils miteinander zu vergleichen.

Ein grundsätzliches Problem, das sich bei einem automatisierten Vergleich stellt, ist die Frage, nach welchen formalen Kriterien eine Übereinstimmung als relevant – oder jedenfalls als überprüfungswürdig – einzustufen ist. Dabei ist davon auszugehen, dass auch bei gut gewählten Kriterien nicht alles gefunden wird, was für die jeweilige Forschungsfrage von Interesse ist, wohl aber manches Irrelevante, und dass eine Erhöhung des *Recalls* durch Aufweichung der Kriterien die *Precision* senken kann und bei einer Verschärfung der Kriterien mit dem umgekehrten Effekt zu rechnen ist.

In dieser Untersuchung ist die Grundannahme, dass wörtliche Übereinstimmungen, die nicht ganz kurz sind, ein guter Indikator für textuelle Abhängigkeitsverhältnisse sind. Bei kurzen Übereinstimmungen ist die Wahrscheinlichkeit hoch, dass sie auf etablierten sprachlichen Mustern oder auch auf zufällig gleicher Wortwahl beruhen, und dementsprechend gibt es bei der Ermittlung von *maximal exact matches* mit einer niedrigen Länge im hier untersuchten Korpus enorm hohe Zahlen von Matches. Der Befund kann nicht unbedingt einfach auf andere Korpora übertragen werden, da die hier ausgewerteten Texte teilweise stark durch formelhafte Sprache geprägt und auch Übereinstimmungen etwas längerer Wortfolgen zum Teil darauf zurückzuführen sind. Die Grundregel, dass kurze Matches für sich genommen wenig aussagekräftig sind und nur ihre Häufung als Hinweis auf Abhängigkeitsverhältnisse gewertet werden kann, während gleiche Wortfolgen größerer Länge nicht einfach zufällig sind, gilt aber sicherlich generell.

Eine vertiefte Untersuchung kürzerer Übereinstimmungen wurde hier nicht in Erwägung gezogen, da es hier insbesondere auch um die Ermittlung von Entsprechungen der Wortwahl trotz unterschiedlicher Schreibung geht. Der gewählte Ansatz, Schreibungsunterschiede durch Anwendung einer Codierung aufzufangen, die zu einem erheblichen Informationsverlust führt, bedingt aber, dass der Vergleich sehr kurzer Wortfolgen auf dieser Basis zu einer hohen Zahl von Matches führt, denen keinerlei Entsprechung der Originaltexte zugrunde liegt. Bei etwas längeren Wortfolgen zeigt sich hingegen, dass Matches beim Vergleich codierter Zeichenfolgen zwar möglicherweise im Randbereich auch kleinere Stücke enthalten, denen im Original keine Wortübereinstimmung zugeordnet werden kann, dass weitergehende

Abweichungen aber sehr unwahrscheinlich sind. Die Zahl der nach diesem Verfahren gefundenen Übereinstimmungen ist aber viel höher als bei Zugrundelegung der Originaltexte oder einer durch Umwandlung in Kleinschreibung vereinfachten Textfassung, und sie sind mit viel höherer Wahrscheinlichkeit als kurze Matches im Hinblick auf die Frage nach textuellen Traditionslinien aussagekräftig.

Dass der gewählte Ansatz funktioniert, wurde hier auch dadurch gezeigt, dass zwar in den Unterkapiteln 3.2.3 und 3.3.1 verschiedene Codierungsvarianten miteinander verglichen wurden, aber für die eigentliche Untersuchung des Korpus eine recht extreme Variante zugrunde gelegt wurde, die mit besonders wenigen unterschiedlichen Codezeichen auskommt und nicht nur die Vokale, sondern auch das Leerzeichen und verschiedene Konsonanten letztlich einfach streicht. Diese Codierungsregeln ermöglichen zwar, dass wörtliche Übereinstimmungen auch bei bestimmten Schreibungsunterschieden ermittelt werden können, sie gehen aber mit einer deutlichen Verschlechterung der *Precision* bei der Ermittlung kurzer Matches einher. Für etwas längere Matches hingegen zeigt sich dieses Problem nicht.

Eine Mindestmatchlänge von 18 Zeichen ist nach den durchgeführten Untersuchungen für diese Codierung und dieses Korpus wohl ein guter Wert, um weitgehend sicherzustellen, dass den Matches tatsächlich gleiche Wortfolgen zugrunde liegen. Bei dieser Mindestlänge (die in dieser Codierung umgerechnet etwa sieben Wörtern durchschnittlicher Länge entspricht) ist zudem der Anteil der Matches besonders hoch, bei denen beide einander zugeordneten Stellen im selben Text liegen. Bei Ansetzung dieser Mindestlänge werden also zwar auch viele Übereinstimmungen gefunden, die nicht auf die Verwendung von Vorlagen zurückzuführen sind – das ist innerhalb eines Textes ja eher unwahrscheinlich –, es lässt sich aber vermuten, dass hier individuelle Formulierungsvorlieben zum Tragen kommen, Matches dieser Länge also in vielen Fällen nicht einfach rein zufällig oder durch allgemein gängige Formulierungsmuster bedingt sind.

Trotz der im Vergleich zu anderen Untersuchungen recht hohen Mindestlänge ist ein Großteil der ermittelten Übereinstimmungen für das Anliegen der vorliegenden Untersuchung nicht aussagekräftig. Es lassen sich aber einfache formale Kriterien finden, anhand derer eine Filterung der Matches und eine Einschätzung ihrer Relevanz vorgenommen werden kann – Kapitel 3.3 stellte hierzu Ansätze vor, die neben der Wortabgrenzung insbesondere die Häufigkeit der Zeichenfolgen und die Positionierung der Matches berücksichtigen. Auch dabei stellt sich allerdings das Grundproblem einer Abwägung zwischen *Precision* und *Recall*, und je nach Korpuszusammensetzung und Untersuchungsanliegen kann eine unterschiedliche Gewichtung verschiedener Parameter sinnvoll sein; insbesondere die in Unterkapitel 3.3.2 untersuchten Bewertungsformeln sind deshalb nur als Beispiele zu verstehen.

Auch wenn man auf eine Filterung auf der Basis einer solchen Bewertung verzichtet, bieten die Daten zu Matches der eben angegebenen Mindestlänge in der ausgewählten Codierung eine gute Basis, um einen Überblick über wörtliche Übereinstimmungen zwischen Einzeltexten und über Textgruppen zu gewinnen. Kapitel 3.4 stellte verschiedene Möglichkeiten dafür vor, die unterschiedlich detailliert sind, insbesondere Visualisierungen.

Graphdarstellungen auf der Basis quantitativer Merkmalen können verdeutlichen, welche Textpaare etwas stärker ausgeprägte Übereinstimmungen aufweisen; dabei können auch Textgruppen erkennbar werden. Allerdings wird die Darstellung bei einem größeren Korpus schnell unübersichtlich, wenn auch Matches mit einem geringen Gesamtumfang berücksichtigt werden.

Für die schnelle Orientierung über Matches zwischen zwei Texten eignet sich das aus der Bioinformatik bekannte Dotplot-Visualisierungsverfahren, das einen Eindruck von der Positionierung der Matches vermittelt und Bereiche mit einer Häufung von Übereinstimmungen leicht erkennbar macht.

Um in ähnlicher Weise auch ein Gesamtbild der Matches eines Textes mit anderen Texten zu erhalten, lassen sich die in einzelnen Dotplots enthaltenen Daten auf die Angaben zur Position der Matches in diesem Text reduzieren und dadurch jeweils in einer Zeile zusammenfassen. Diese anscheinend – trotz gewisser Ähnlichkeiten mit einem am Ende von Unterkapitel 2.2.2 beschriebenen Auswertungsverfahren für Matchdaten – neue Darstellungsform wird hier als *Projektionsdotplot* bezeichnet. Obwohl sie gegenüber den Dotplots für einzelne Textpaare mit einem erheblichen Informationsverlust behaftet ist, eignet sie sich bei den hier gewählten Einstellungen gut, um Textbereiche zu ermitteln, in denen es im jeweils primär betrachteten Text vermutlich signifikante Übereinstimmungen gibt, und zugleich können Ähnlichkeiten zwischen der Matchverteilung in verschiedenen Textpaaren und damit Gruppenzusammenhänge zwischen Texten sichtbar werden.

Die vorliegende Darstellung ist aus dem Anliegen erwachsen, eine Basis für die Analyse von Abhängigkeitsverhältnissen innerhalb der im maschinenlesbaren Volltext vorliegenden Quellen des Projekts DRQEdit zu schaffen. Eine einigermaßen umfassende Auswertung und Interpretation der Daten wäre in einem anderen Rahmen zu leisten, aber anhand einiger in Teil 4 vorgestellter Beispiele konnte gezeigt werden, dass sich auf der Basis der ermittelten Matchdaten neue Erkenntnisse über die Beziehungen zwischen den Texten gewinnen und Fehleinschätzungen korrigieren lassen.

Anhang

Glossar

Das Glossar enthält kurze Erläuterungen zu Fachtermini und Abkürzungen, die im Haupttext beziehungsweise an anderen Stellen im Glossar als bekannt vorausgesetzt werden. Diese Erläuterungen sollen vor allem den nicht primär an EDV-Fragen interessierten Lesern eine ungefähre Vorstellung vermitteln, worum es geht, nicht aber Definitionen im Sinne einer technischen Dokumentation liefern. Querverweise innerhalb des Glossars sind durch Verweispfeile gekennzeichnet.

Alinierung Als Alinierung (oder Alignierung, englisch *alignment*) wird die Ermittlung beziehungsweise Darstellung von Entsprechungen zwischen zwei →Strings (oder anderen Elementfolgen) bezeichnet, wobei (zumindest im hier zugrunde liegenden Verständnis) die einander zugeordneten Elemente jeweils gleich sind und in beiden Zeichenketten in der gleichen Reihenfolge vorkommen. Die Bewertung einer Alinierung als optimal kann nach verschiedenen Kriterien erfolgen, zum Beispiel nach der Zahl der einander zugeordneten Zeichen oder mit einer besonderen Gewichtung von Lücken zwischen den Entsprechungen. Alinierungen lassen sich zum Beispiel so darstellen, dass die untersuchten Zeichenfolgen in zwei Zeilen übereinander stehen und spezielle Füllzeichen eingeschoben werden, um zu erreichen, dass die einander zugeordneten Elemente jeweils direkt übereinander stehen.

Array Ein (eindimensionales) Array ist eine Datenstruktur zur Verzeichnung von mehreren Einträgen in einer solchen Form, dass diese über ihre Ordnungsnummer im Array adressiert (also ausgelesen und gegebenenfalls auch verändert) werden können. Ein zweidimensionales Array ist eine entsprechende Struktur, wobei hier für die Adressierung jeweils zwei Zahlen erforderlich sind. Es lässt sich in Tabellenform darstellen, wobei der jeweils ersten Zahl die Zeile in der Tabelle zuzuordnen ist und der zweiten Zahl die Spalte (oder immer umgekehrt).

ASCII Der *American Standard Code for Information Interchange (ASCII)* ist ein Zeichensatz, der →*Codepoints* für 128 Zeichen (die Buchstaben des lateinischen Alphabets in Groß- und Kleinschreibung, die arabischen Ziffern, das Leerzeichen sowie einige Satz- und Sonderzeichen und nicht druckbare Steuerzeichen) vorsieht. Da für die Codierung nur 7 Bits benötigt werden, die Zeichen aber in Bytes (also jeweils 8 Bit) gespeichert werden, gibt es Erweiterungen, die den vom 7-Bit-ASCII nicht belegten *Codepoints* Zeichen zuordnen, die in ASCII nicht vorgesehen sind, zum Beispiel die im Deutschen verwendeten Umlaute und das *ß*. Da die Belegung dieser *Codepoints* aber je nach verwendeter Codierung variiert, ist einer Textdatei nicht ohne Weiteres anzusehen, welche Zeichen in ihr gespeichert sind. Vgl. →UTF-8.

Baum Ein Baum ist eine Datenstruktur aus dem Bereich der Graphentheorie, er besteht also aus →Knoten und →Kanten. Kennzeichnend für einen Baum ist, dass es genau einen Wurzelknoten gibt, mit dem alle übrigen Knoten durch genau einen Pfad von Kanten (gegebenenfalls mit dazwischen liegenden Knoten) verbunden sind. Das bedeutet, dass außer der Wurzel jeder Knoten genau einen Elternknoten hat (der dem Wurzelknoten einen Schritt näher steht) und dass jeder Knoten über den Verbindungspfad zur Wurzel

identifiziert werden kann. Knoten können prinzipiell beliebig viele Kindknoten haben. Knoten ohne Kindknoten werden als Endknoten oder als Blätter bezeichnet. Wofür in einem Baum ein Knoten beziehungsweise eine Kante steht, hängt von der jeweiligen Verwendung ab, ebenso, inwieweit die Knoten und Kanten über ihre Verknüpfung hinaus zusätzliche Informationen speichern. Ein Baum kann zum Beispiel zur Beschreibung von Inklusionsbeziehungen – etwa der Elemente in einem →XML-Dokument – dienen oder auch zur Strukturierung von Alternativen.

Codepoint Ein *Codepoint* ist ein numerischer Wert, der in einem Zeichensatz ein bestimmtes Zeichen repräsentiert.

Codierung Als Codierung wird in dieser Untersuchung insbesondere die Anwendung von Transformationsregeln auf einen Text bezeichnet, im engeren Sinne bezogen auf Transformationen, deren Resultat eine nicht mehr lesbare Folge von Codezeichen ist, im weiteren Sinne (um die Darstellung zu vereinfachen) auch unter Einschluss von Transformationen, die den Buchstabenbestand im Wesentlichen unverändert lassen. Je nach Zusammenhang kann damit auch das Resultat der Transformation bezeichnet sein. In Bezug auf Zeichensätze steht der Begriff hingegen für die Repräsentation von Zeichen durch bestimmte numerische Werte beziehungsweise Bitmuster.

DTD Eine *Document Type Definition (DTD)* dient dazu, formale Regeln über die Elemente und Attribute eines →XML-Dokuments festzuhalten. Sie beschreibt in einer bestimmten Notation, welchen Inhalt die einzelnen Elemente haben können oder müssen (dabei kann es sich um weitere Elemente oder um Text handeln) und welche Attribute für sie zulässig oder vorgeschrieben sind. Wenn eine XML-Datei den Regeln einer DTD entsprechen soll, wird diese vor dem Wurzelelement des Dokuments angegeben (wobei sie vollständig eingefügt, aber auch auf eine lokal oder online verfügbare DTD verwiesen werden kann). Die generelle Syntax und Semantik von DTDs ist Teil der *W3C Recommendation* zu XML (*http://www.w3.org/TR/xml/#sec-prolog-dtd*). Anstelle von DTDs werden für die Beschreibung der Dokumentstruktur zunehmend →XML-Schemasprachen verwendet, die genauere Angaben über zulässige Elementinhalte enthalten können.

GUI Der englischen Bezeichnung *graphical user interface (GUI)* entspricht auf deutsch *graphische Benutzeroberfläche* – die englische Abkürzung wird hier vor allem um der Kürze willen verwendet. Ein GUI bietet den Benutzern des jeweiligen Programms in aller Regel vielfältige Möglichkeiten der Interaktion (zum Beispiel über Menüpunkte oder graphische Elemente wie Icons und Buttons). Programme ohne graphische Oberfläche können vom Benutzer nur – soweit im Programm vorgesehen – über beim Aufruf übergebene Parameter, über Konfigurationsskripte oder auch gegebenenfalls durch Beantwortung von Fragen, die während des Programmablaufs gestellt werden, gesteuert werden. Das bedeutet gegenüber den Möglichkeiten eines GUI zwar eine Einschränkung, erleichtert aber zugleich die wiederholte Verwendung für gleiche oder ähnliche Aufgaben, da die entsprechenden Programmaufrufe selbst gespeichert werden können.

Hashtabelle Hashtabellen verzeichnen Daten in Paaren von Schlüsseln und zugeordneten Werten und bieten die Möglichkeit, in mehr oder weniger konstanter Zeit nach einem Schlüssel (und dem zugeordneten Wert) zu suchen. Dabei wird über eine Hashfunktion die

Speicheradresse auf der Basis des Schlüssels berechnet. Eine gut gewählte Hashfunktion ermöglicht nicht nur eine schnelle Berechnung der gesuchten Adresse, sondern führt auch in aller Regel zu einer relativ gleichmäßigen, niedrigen Belegung der von der Hashtabelle genutzten Adressen, so dass mit einer geringen Zahl von Rechenoperationen der dem Schlüssel zugeordnete Wert ermittelt oder auch einfach die Existenz eines Schlüssels überprüft werden kann.

Kanonische Referenz Als kanonische Referenz (beziehungsweise als *canonical reference*) im Sinne der →TEI wird eine etablierte Form des Verweises auf eine Textstelle (oder einen Abschnitt) bezeichnet.[714]

Kante Eine Kante (im Sinne der Graphentheorie) ist eine (gerichtete oder bidirektionale) Verbindung zwischen zwei →Knoten, kann also zum Beispiel Bestandteil eines →Baums sein.

Knoten Ein Knoten (im Sinne der Graphentheorie) ist eine elementare Einheit, die bei einer graphischen Darstellung durch einen Punkt repräsentiert werden kann. Er kann über →Kanten mit anderen Knoten verbunden sein. Er kann zum Beispiel Bestandteil eines →Baums sein.

Markup Was zum *Markup* eines →XML-Dokuments zählt, wird in der XML-Spezifikation unter Punkt 2.4[715] aufgeführt. Neben →*Tags* zählen auch Entitäts- und Zeichenreferenzen, XML-Kommentare, *Processing Instructions* und Ähnliches dazu.

OCR *Optical character recognition (OCR)* steht für die Ermittlung des auf einer eingescannten Vorlage stehenden Textes durch ein Computerprogramm. OCR weist auch nach dem derzeitigen Stand eine nicht unerhebliche Fehlerrate auf, die insbesondere dann steigt, wenn die Schrift der Vorlage von den heute gängigen Schrifttypen stärker abweicht und auch in sich eine größere Varianz aufweist sowie wenn die Sprache und Orthographie des Textes nicht den dem OCR-Programm bekannten Regeln entspricht.

Precision* und *Recall Die Begriffe *Precision* und *Recall* dienen zur Beschreibung der Qualität einer nach bestimmten Kriterien erfolgenden Auswahl von Elementen einer Gesamtmenge, zum Beispiel von Dokumenten, die zu einer Suchanfrage passen. Mit *Precision* wird angegeben, in welchem Maße die gefundenen Elemente tatsächlich den Kriterien entsprechen, mit *Recall*, wie hoch der Anteil der gesuchten Elemente ist, die tatsächlich gefunden werden. Im Idealfall werden für beides hohe Werte erreicht, es passiert aber leicht, dass eine Verbesserung der *Precision* aufgrund strengerer Auswahlkriterien zu einer Verschlechterung des *Recalls* führt und umgekehrt.

Regulärer Ausdruck Ein regulärer Ausdruck ist ein Suchmuster, in dem in einer bestimmten Notationsform festgehalten ist, welche Bedingungen in einem zu untersuchenden String erfüllt sein müssen, damit das Muster passt. Dabei ist es insbesondere möglich, Teile des Musters als fakultativ, als beliebig oft wiederholt oder als Varianten zu kennzeichnen, hinzu kommen (je nach Implementierung) zum Teil noch viele weitere Möglichkeiten.[716]

[714] Vgl. *http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SACR.*
[715] *http://www.w3.org/TR/2006/REC-xml11-20060816/#syntax.*
[716] Vgl. Friedl 2004.

Sequenz Der Begriff *Sequenz* wird hier teilweise im Sinne der Bioinformatik verwendet. Dort geht es um Nukleotid- sowie Aminosäure-Sequenzen, also um die zusammenhängenden Ketten von Bausteinen, die Nukleinsäuren und Proteine bilden. Hier kann er daneben auch allgemeiner eine geordnete Folge von Elementen bezeichnen. Vgl. →Teilsequenz.[717]

Teilsequenz Eine Teilsequenz setzt sich aus Elementen zusammen, die in einer →Sequenz in derselben Reihenfolge, aber mit einer beliebigen Zahl von Unterbrechungen zu finden sind.

Serialisierung *Serialisierung* bedeutet die Übertragung von strukturierten Daten in eine (jeweils bestimmte) lineare Form, insbesondere die Übertragung von Informationen aus dem Arbeitsspeicher in Dateien, die später in einer entsprechenden Deserialisierung wieder eingelesen werden können. Damit wird es zum einen ermöglicht, diese Datenstrukturen auch sitzungsübergreifend zu speichern, zum anderen lassen sich Informationen aus dem Arbeitsspeicher auslagern, wenn dieser nicht ausreicht. Verbunden ist dies freilich mit Geschwindigkeitsverlust, da das Lesen und Schreiben von Dateien vergleichsweise zeitaufwendig ist.

String Das englische Wort *String* (im übertragenen Sinne *Kette, Reihe [von Einzelgliedern]*) wird in der Informatik als Terminus für *Zeichenkette, Zeichenfolge* gebraucht, also zur Benennung einer lückenlosen Sequenz von Zeichen eines Zeichensatzes.

Tag *Tags* dienen in →XML dazu, Textstücke oder auch leere Zeichenfolgen bestimmten Strukturen (Elementen) zuzuordnen und sie gegebenenfalls über Attribute mit zusätzlichen Informationen zu versehen.

TEI Die *Text Encoding Initiative (TEI)* zielt auf die Vereinheitlichung der Textcodierung in geisteswissenschaftlichen Projekten, um die langfristige Nutzbarkeit und die Austauschbarkeit der erstellten Dateien zu ermöglichen beziehungsweise zu erleichtern. Nach dem heutigen Stand wird dabei eine Textauszeichnung in →XML vorausgesetzt.[718] Die *TEI Guidelines*[719] enthalten insbesondere Vorschläge, mit welchen →*Tags* und Attributen bestimmte (objektive oder nach der jeweiligen Interpretation des Bearbeiters vorliegende) Merkmale eines Textstücks zu kennzeichnen sind. Die Strukturspezifikation der TEI (zum Beispiel in einer →DTD) enthält formale Regeln für den Einsatz der *Tags* und Attribute, so dass auf dieser Basis überprüft werden kann, ob eine XML-Datei dieser Datenstruktur entspricht. Durch den modularen Aufbau dieser Strukturregeln ist es möglich, für eigene Dokumente nur Teile der insgesamt sehr komplexen Spezifikation in einer XML-Schemasprache zu verwenden und gegebenenfalls auch eigene Erweiterungen oder Änderungen vorzunehmen.

Unicode *Unicode* ist eine (nicht abgeschlossener) Zeichencodierung, die letztlich die Zeichen aller Schriftsysteme umfassen soll. Den Zeichen sind jeweils bestimmte →*Codepoints* zugeordnet. Es gibt aber verschiedene Möglichkeiten, wie diese →*Codepoints* in einer

[717] Vgl. zur Unschärfe in der eingeführten Begrifflichkeit Gusfield 1997, S. 4.

[718] Ursprünglich setzte die TEI eine Auszeichnung nach den Regeln der *Standard Generalized Markup Language (SGML)* voraus, und auch für die Zukunft sind Änderungen prinzipiell denkbar – vgl. *http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html*.

[719] Einstiegsseite für die aktuelle Version (*P5*): *http://www.tei-c.org/Guidelines/P5/*.

Unicode-codierten Datei gespeichert werden können. Ein verbreitetes Datenformat ist →UTF-8.

UTF-8 *UTF* steht für *Unicode Transformation Format*. Es gibt davon mehrere, die jeweils unterschiedliche Regeln enthalten, wie die →*Codepoints* von →*Unicode* in Bytesequenzen gespeichert werden. Besonders verbreitet ist UTF-8, weil darin für die Zeichen des 7-Bit-ASCII-Zeichensatzes, also die Buchstaben des lateinischen Alphabets und einige weitere Zeichen, jeweils nur ein Byte benötigt wird, das der Codierung in ASCII entspricht. Für die übrigen Unicode-Zeichen werden 2–4 Bytes benötigt. Für XML-Dokumente gilt die Grundregel, dass die Zeichen in UTF-8 gespeichert sind, sofern zu Beginn nichts anderes angegeben ist.

XML Die *Extensible Markup Language (XML)* ist ein generisches Auszeichnungsschema zur Anreicherung von Textdateien mit Zusatzinformationen. Damit eine Textdatei eine im Sinne der XML-Spezifikation „wohlgeformte" („well-formed"[720]) Datei und damit ein XML-Dokument ist, müssen einige Grundregeln eingehalten sein: – Die Zeichen „<", „>" und „&" haben eine besondere Funktion. Wenn diese Zeichen selbst Bestandteil des eigentlichen Textes sein sollen, müssen sie codiert werden durch „<" (für *less than*), „>" (für *greater than*) beziehungsweise „&" (für *ampersand*). – Jedes Stück des eigentlichen Textes ist Bestandteil zumindest eines sogenannten Elements. Die Zugehörigkeit wird dadurch gekennzeichnet, dass der entsprechende Abschnitt von einem Anfangs- und einem Schluss-*Tag* umrahmt wird, die den jeweiligen Elementnamen enthalten. Ein Anfangs-*Tag* hat die Grundform „<BEZEICHNER>", ein Schluss-*Tag* die Grundform „</BEZEICHNER>" (wobei „BEZEICHNER" durch den jeweiligen Elementnamen zu ersetzen ist; das Anfangs-*Tag* kann außerdem vor „>" noch sogenannte Attribute enthalten). Die Bereiche verschiedener Elemente können ineinander geschachtelt sein, einander aber nicht überlappen. Jedes XML-Dokument muss ein einziges Wurzelelement haben; alle weiteren Elemente (sofern vorhanden) müssen direkt oder über Zwischenglieder vom Wurzelelement umschlossen sein. – Zu Beginn der Datei ist in einer sogenannten XML-Deklaration die Sprachversion und (sofern die Datei nicht in →UTF-8 gespeichert ist) der Zeichensatz anzugeben. Als Beispiel sei die Deklaration für eine Datei genannt, die die XML-Version 1.0 zugrunde legt und den *ISO-Latin-1*-Zeichensatz verwendet: *<?xml version="1.0" encoding="iso-8859–1"?>* Für die weiteren XML-Regeln, die zum Beispiel die Einfügung von Kommentaren und von Element-Attributen sowie die Festlegung von Regeln für die Element- und Attributstruktur ermöglichen, sei auf die XML-Spezifikation (*http://www.w3.org/TR/xml/* beziehungsweise *http://www.w3.org/TR/xml11*) verwiesen. Außerdem gibt es eine ganze Reihe von weiteren Spezifikationen, die im Zusammenhang mit der Verarbeitung von XML-Dateien stehen. Für die vorliegende Arbeit ist insbesondere →*XPath* von Bedeutung.

XML-Schemasprache Eine XML-Schemasprache ermöglicht die Erstellung von formalen Spezifikationen für →XML-Dokumente. Beispiele sind das für DTDs verwendete Notationssystem, *RELAX NG*[721] und *XML Schema*[722].

[720] *http://www.w3.org/TR/2006/REC-xml11-20060816/#sec-well-formed.*
[721] Vgl. *http://relaxng.org/.*
[722] Vgl. *https://www.w3.org/TR/xmlschema-0/*

XPath Die *XML Path Language (XPath)* ist eine formale Sprache, die es insbesondere ermöglicht, Teile eines →XML-Dokuments zu bezeichnen. Dabei wird die Baumstruktur der Elemente (und sonstigen Bestandteile) eines XML-Dokuments zugrunde gelegt. Von den zahlreichen Adressierungsmöglichkeiten sind hier nur zwei von Interesse: – 1.) Jedes Element kann durch einen Pfad bezeichnet werden, der vom mit „/“ bezeichneten Wurzelknoten (für das gesamte Dokument) über die ineinander geschachtelten Elemente zu ihm führt. Dabei werden die jeweiligen Elementnamen in einer bestimmten Weise aneinandergereiht (in der vereinfachten Notation mit „/“ als Trennzeichen). Da unter den in einem Eltern-Element unmittelbar enthaltenen Kind-Elementen dieselbe Elementbezeichnung mehrfach vorkommen kann, ist zur eindeutigen Bezeichnung gegebenenfalls anzugeben, um das wievielte von ihnen es sich handelt (in der vereinfachten Notation in der Form „[ZAHL]“). Ein Beispiel für einen XPath-Ausdruck nach diesem Schema (wobei Elementbezeichnungen der →TEI verwendet werden) ist: */TEI/text/body/div[3]/p[5]/seg[2].* – 2.) Ein Element kann über ein Attribut mit einem Wert bezeichnet werden. Dabei wird zunächst durch „//“ angegeben, dass die Position des Elements im Dokument keine Rolle spielt, dann wird der Elementname angegeben (oder „*“, wenn jedes Element in Frage kommt) und dann der Attributname und -wert in einer bestimmten Notation. Auch hierzu ein Beispiel entsprechend der →TEI: *//div[@n="VI 5 § 3"].* Inwieweit das Attribut geeignet ist, das Element zu identifizieren, hängt von seiner Verwendung ab (die durch die Spezifikation in einer →XML-Schemasprache reglementiert sein kann). Während die TEI für das n-Attribut zulässt, dass ein Wert mehrfach vorkommt, ist das in den Dokumenten des Projekts DRQEdit nicht vorgesehen.

Quellen und Literatur

Im Quellenverzeichnis sind die Texte aufgeführt, die bei der hier beschriebenen Untersuchung ausgewertet wurden, sowie in einzelnen Fällen (mit entsprechendem Vermerk) weitere Texte aus dem Korpus des Projekts DRQEdit oder andere Ausgaben der Texte. Die bibliographischen Angaben zum Quellenkorpus orientieren sich an den für neuere Werke etablierten Zitierkonventionen, sind aber mit den entsprechenden Nummern im *Incunabula Short Title Catalogue (ISTC)* beziehungsweise im *Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts (VD 16)* versehen, um in Zweifelsfällen eine Unterscheidung von anderen Ausgaben desselben Erscheinungsjahrs zu ermöglichen. Welche Exemplare für die Volltexterfassung zugrunde gelegt wurden, ist im Quellenverzeichnis des Projekts DRQEdit[723] dokumentiert. Die zu den Drucken des 16. Jahrhunderts angeführten Titel beruhen im Wesentlichen auf den Angaben im VD16, sind aber – neben sonstigen Anpassungen – in einer Reihe von Fällen stärker als dort gekürzt. Auf den genannten Katalogen basieren auch die Ansetzungsform der Autorennamen sowie die Informationen zu Druckort und Erscheinungsjahr.

Literatur wird in der Regel nach Verfasser beziehungsweise Herausgeber in Kapitälchen sowie Erscheinungsjahr zitiert. Wenn dies nicht möglich oder sinnvoll ist, insbesondere bei Publikationen, die unter einer Sigle bekannt sind, wird die etablierte Sigle beziehungsweise ein abgekürzter Titel verwendet. Eine Nennung von URL, URN oder DOI ist bei neueren Publikationen auch ein Hinweis darauf, dass die online verfügbare Version zugrunde gelegt wurde. In den betreffenden Fällen ist nicht immer klar ersichtlich, ob die Texte auch gedruckt wurden, was die genauen Titeldaten der betreffenden Gesamtpublikation sind und ob der online publizierte Text exakt mit dem gedruckten übereinstimmt. Bei älteren Werken sind im Interesse der Benutzerfreundlichkeit auch Adressen von nicht verwendeten, aber offenbar übereinstimmenden Retrodigitalisaten angegeben. Onlineangebote, die sich nicht (oder nicht klar erkennbar) als datierte Publikation eines Autors beschreiben lassen, werden hier nicht aufgeführt. Bei ihnen wird die URL (gegebenenfalls: die URL der Homepage) in den Anmerkungen des Haupttextes genannt. Die URLs wurden am 28. 12. 2017 überprüft; soweit die genannten Ressourcen zu diesem Zeitpunkt offenbar nicht mehr existierten oder jedenfalls nicht aufrufbar waren, ist dies vermerkt.

Quellen

AmbergGesatzB. 1554: Der Churfürstlichen Stat Amberg Gesatzbuch / widerumb verneut / vnd mit der new erlangten freyheyt gemehrt / Auch ... geendert ..., Amberg 1554 (VD16: A 2171).

AugsbDomKapGO. 1539: Gerichts ordnung Durch ein Ehrwürdig Capitul deß Thůmbstiffts zu Augspurg / in vnd an jren vndern Gerichten zuhalten / fürgenommen, o. O. 1539 (VD16: ZV 870).

[723] *http://drw-www.adw.uni-heidelberg.de/drqedit/.*

AugsbHochstiftOrdn. 1552: DEß Hochwirdigisten Fůrsten vnd Herrn / Herrn Otho / ... Bischouen zů Augspurg / vnd jhrer F. G. Stiffts Vndergerichts vnd Straffordnungen, Dillingen 1552 (VD16: A 4092).

AugsbHochstiftStrafO. 1552 = AugsbHochstiftOrdn. 1552, Bl. 17 r - 23 r.

AugsbHochstiftUGO. 1552 = AugsbHochstiftOrdn. 1552, Bl. 1 r - 16 v.

BadenTestamentO. 1511: Der MargGraffschafft Baden Statuten vnd Ordenungen in Testamenten / Erbfållen / vnd Vormündschafften, o. O. 1511 (VD16: B 91).

BambDomstiftGerRef. 1463 (1488) = BambDomstiftOrdn. 1488, [S. 3] - [S. 19].

BambDomstiftGerRefErkl. 1488 = BambDomstiftOrdn. 1488, [S. 20] - [S. 24].

BambDomstiftOrdn. 1488: Die Reformation des gerichtes der Dechaney des Thumstifts zu Bamberg durch den hochwirdigen ... herñ Georgen Bischoff zu Bamberg gemacht vñ geordẽt von latein in teutsch bracht, [Bamberg 1488] (ISTC: ir00036000).

BambHalsGO. 1507: Bambergische halßgerichts ordenũg, Bamberg 1507 (VD16: B 256).

BambHalsGO. 1580: Bambergische Peinliche Halßgerichtßordnung, Bamberg 1580 (VD16: B 266).

BambHofGO. 1497: Ein newe reformation des hochwirdigen fursten vnd herren Heinrichen Bischoue czu Bamberg, [Bamberg 1497] (ISTC: ih00011400).

BambLGRef. 1503: Des lantgerichts zu Bamberg Reformation, Bamberg 1503 (VD16: B 255).

BayrGO. 1520: Gerichtzordnung Jm̃ Fůrstñthumb Obern vnd Nidern Baÿrn / Anno 1520 aufgericht, München 1520 (VD16: B 989).

BayrLandhandf. 1516: Die Ordnung vber gemeiner Landtschafft in Bairn aufgerichte Hanndtüesst. Tausent funffhundert vnd jm sechzeheden jar. zu Jngolstat beslossen, Landshut 1516 (VD16: B 1003).

BayrLandrichterInstr. 1565: Jnstruction für alle Pfleger vnd Landtrichter / auch andere Obrigkhaiten / des Fůrstenthumbs Bayrn / Was sy zů abstellung vnd verhůetung des schedlichen mordprennens fůrnemblich anordnen bestellen vnd handlen sollen, München 1565 (VD16: B 1050).

BayrLFreihErkl. 1553: Erclärung der Landsfreihait in Obern vnnd Nidern Bairn widerumb verneut Jm Funfftzehenhundert Dreiundfunfftzigistem Jar, München 1553 (VD16: B 1028).

BayrLO. 1553: Bairische Lanndtß 15 ordnung 53, Ingolstadt 1553 (VD16: B 1034).

BayrLRRef. 1518: Reformacion der baÿrischñ Lanndrecht nach Cristj vnsers Hailmachers geburde Jm Funftzehenhundert vnnd Achtzehendm Jar Aufgericht, München 1518 (VD16: B 1007).

Belial(dt.) 1472 (Zainer): Jacobus de Theramo, [ohne Titel, Textanfang: Hie hebt sich an eyn gůtt núczlich bůch von der rechtlichen úberwúndung cristi wider sathan, [Augsburg] 1472 (ISTC: ij00074000) [nicht im ausgewerteten Korpus].

Bemel,TraktTestam. 1587: Georg Bemel, TRACTATVS AD FORMVLAM TESTAMENTI, Das ist: Grünndlicher Bericht ..., Frankfurt/Main 1587 (VD16: B 1665).

BerthRechtssumme(dt.) 1472: Johannes <Friburgensis>, [ohne Titel, Textanfang: Hie nach volget das Register über das bůch genãt Sũma Johãnis Nach ordnung des A b c 7c], Augsburg 1472 (ISTC: ij00317000) [nicht im ausgewerteten Korpus].

BrandenbAnsbGO. 1539: BRANDENBVRGISCHE GERICHTS ORDNVNG VND REFORMATION DES BVRGGRAFTHVMBS zů Nürmberg / Kayserlichen Lannd / vnd Fürstlichen Oberhofgerichts, o. O. 1539 (VD16: ZV 15787).

BrandenbAnsbHalsGO. 1516: Brandenburgische halßgerichtsordnung, Nürnberg 1516 (VD16: B 6937).

BrandenbErbfKonst. 1527 (1549): CONSTITVTION, Wilkȯr / vnd ordnung der Erbfelle / vnd anderer sachen / wie damit durch die gantzen Marck zu Brandenburg / vnd darzu gehȯrenden landen / hinfůr sol gehalten werden, Frankfurt / Oder 1549 (VD16: B 6911).

BrandenbKGO. 1540: Reformation Churfurstlicher gnaden zu Brandemburg Camergerichts zu Coln an der Sprew, Berlin 1540 (VD16: B 6915).

BraunschweigCalenbergHofGO. 1544: Reformation vnd satzung / vnser Elysabethen ... Hertzoginnen zu Braunschweig vnd Leuneburg etc. Witwen / der Ober vnd Hoffgericht / So wir in vnser Leibzucht Můnden etc. vnd vnsers ... Sons Hertzog Erichen Fůrstenthumen vnd Landen ... geordent haben, Hannover 1544 (VD16: B 7258).

BraunschweigLüneburgHofGO. 1564: Vnser von Gottes Gnaden Heinrichen vnd Wilhelmen der Jüngern / Gebrüder / Hertzogen zu Braunschweig vnd Lünenburg Hoffgerichts Ordnung Wie es vor demselbigen / auch sunst vor den Vntergerichten in unserem Fůrstenthumb Lůnenburg / mit Process vnd sunst / solle gehalten werden, Wittenberg 1564 (VD16: B 7294).

BraunschweigWolfenbüttelHofGO. 1556: Hoffgerichts ordnung des Durchleuchtigen ... Hern Heinrichs des Jůngern Hertzogs zu Braunschweig vnd Lůnenburg etc. Newlich geordnet vnd auffgericht, Wolfenbüttel 1556 (VD16: B 7292).

BraunschweigWolfenbüttelHofGO. 1571: Hoffgerichts Ordnung: Des Durchleuchtigen ... Herrn Juliussen / Hertzogs zu Braunschweig vnd Lůneburg / etc. Auffs new verbessert / gemehret ... Sampt angehengter Key. Confirmation / Auch Priuilegio, de non appellando ..., Wolfenbüttel 1571 (VD16: ZV 18814).

BreslauGO. 1591: Der Kayserlichen Stadt Breßlaw Vornewerte Gerichts Ordnung vnd Process, Breslau 1591 (VD16: ZV 18411).

BreslauStat. 1588: Der Kayserlichen Stadt Breßlaw Statuta vnd Ordnungen / auffs New vmbgefertiget / vormehret vnd gebessert, Breslau 1588 (VD16: B 8011).

CoburgHofGO. 1581: Der Durchlauchtigsten / Durchlauchtigen / Hochgebornen Fůrsten vnd Herrn / Herrn Ludowigs Pfaltzgrafen bey Rhein ... Herrn Augusten Hertzogen zu Sachssen ... auch Herrn Georg Friedrichs / Marggraffen zu Brandenburg ... in semptlicher Vormundschafft / jhrer ... Pflegesȯne ... Hertzogen zu Sachssen etc. ... bedachte Hoffgerichtsordnunge / welcher gestalt dieselbe in jhrer F. G. ort Landes zu Francken jherlichen zu Coburgk gehalten werden solle, Jena 1581 (VD16: ZV 22704).

CoburgHofGO. 1598: DEr Durchlauchtigen Hochgebornen Fůrsten vnnd Herrn / Herrn Johann Casimirn / vnd Herrn Johann Ernsten / Gebrůdere / Hertzogen zu Sachsen ... Hofgerichts Ordnung / welcher gestallt dasselbe ... zu Coburgk gehalten / vnnd darinnen allenthalben verfahren werden soll ..., Coburg 1598 (VD16: S 1113).

Damhouder,Patrocinium 1576 = Damhouder,Werke(Burckhard) 1576, S. 1 - 119.

Damhouder,Praxis 1565: Joost de Damhoudere, PRAXIS RERVM Criminalium. Grůndlicher Bericht vnd Anweisung / Welcher massen in Rechtfȧrtigung Peinlicher sachen / nach gemeynen beschribenen Rechten / vor vnd in Gerichten ordenlich zuhandeln. ... in hoch Teutsche Sprach mit vleiß verwandelt ... Durch Michaelem Beůther ... [ohne den angehängten Text], Frankfurt/Main 1565 (VD16: D 61).

Damhouder,Verganten 1576 = Damhouder,Werke(Burckhard) 1576, Blatt [Aa] r - S. 140.

Damhouder,Werke(Burckhard) 1576: Joost de Damhoudere, PATROCINIVM PVPILLORVM. Von Vormundtschafften ... in Teutsche Sprach gebracht Durch ... Johan Burckharden ..., Frankfurt/Main 1576 (VD16: D 54).

EiderstedtLR. 1572 (1573): Vnser van Gades gnaden / Adolffen Eruen tho Norwegen Hertogen tho Schleßwig ... beschreuen Landtrecht / welckes wy vnsen Vnderdanen in vnsen dreien Landen / Eiderstede / Euerschop vnde Vttholm / tho gewisser handthauinge vnde vorfolge der Justitien ordentlikes Gerichts vnde der Gerechticheit gnedichlick gegeuen vnde bestediget. ..., Hamburg 1573 (VD16: S 2969).

EiderstedtLR. 1591 = EiderstedtOrdn. 1591, Blatt B r - [Aa iiij] v.

EiderstedtOrdn. 1591: VNser von Gottes gnaden Johan Adolffen / Postulirten vnd Erwelten zu Ertz vnd Bischoffen der Stiffte Bremen vnd Lůbeck / Erben zu Norwegen / Hertzogen zu Schleßwig ... Beschrieben newes LandtRecht / welches wir vnsern getrewen Vnderthanen / in vnsern dreyen Landen / Eyderstede / Euerschop vnd Vtholm / zu gewisser handthabung vnd verfolge der Justicien / ordentliches Gerichts vnd der Gerechtigkeit auffs new gnedigst gegeben / verbessert Confirmirt vnd bestetigt haben. ... (Titel zum enthaltenen Text: Policey Ordnung), Schleswig 1591 (VD16: S 2973).

EiderstedtPolO. 1591 = EiderstedtOrdn. 1591, Blatt [Aa iiij] v - [Ii iij] r.

Ekhardi,MagdebR.(Pölmann) 1574 = Pölmann,Handbuch 1574, ab 2. Bogensignatur A.

Faber,Form. 1539: Dionysius Faber, Formulare procuratorum Proces vnde Rechtes ordeninge / Rechter arth vnde wise / der Ridderrechte yn Lifflande / So wol yn den Stifften /alse yn Harrien vnde Wirlande / vnde gemennichliken ym gebruke auer gantzem Lifflande. ..., Magdeburg 1539 (VD16: F 50).

FrankenLGRef. 1512 = SammelwerkFranken(Schubart) 1512, Blatt [A] r - [B vij] v.

FrankfurtRef. 1509: Reformacion d̃er S̃tat F̃ranckenfort am M̃eine d̃es h̃eilgê R̃omischê R̃ichs C̃ãmer ão ĩ509, Mainz 1509 (VD16: F 2307).

FrankfurtRef. 1578: Der Statt Franckenfurt erneuwerte Reformation, Frankfurt/Main 1578 (VD16: F 2298).

FreiburgStR. 1520: Nüwe Stattrechten vnd Statuten der loblichen Statt Fryburg im Pryßgow gelegen. ..., Basel 1520 (VD16: F 2540).

Fruck,TeutschForm. 1522: Ludwig Fruck, Teutsch Formularj wie man in gerichtssachen vnnd andern handlen / vnd geschefftẽ / brieffe / verschreybung / vnd instrument stellen mag / nicht allein dienstlich / besonder ser nützlich. Ludouicus Fruck, Wien 1522 (VD16: F 3146).

Fuchsperger,Inst. 1536: Ortolf Fuchsperger (Bearb.), Justinianischer Jnstituten warhaffte dolmetschung / darin der großmechtigst Kayser IVSTINIAN. den ersten grond geschribner recht hat fürgebildet: durch Orth. Fuchsper. von Dit, Augsburg 1536 (VD16: C 5236).

Gessler,Form. 1493: Gessler, Heinrich, WJe man einem yecklichẽ was wurden vnd stads der ist / schryben soll /new practicirt rethoric vñ brieff formulary des adels / stetten vnd lendern des hochtútschẽ yetz louffenden stylums vñ gebruchs /vormals durch die synreichẽ kunst bůchtrůcken jn gmein nit yßgegossen, Strassburg [1493] (ISTC: ig00279000) [nicht im ausgewerteten Korpus].

Gobler,GerProz. 1536: Justinus Göbler, GErichtlicher Proceß auß grund der Rechten vnd gemeyner ůbung / zum fleissigsten in drei theyl verfasset. Das Erst Theyl ... Die Termin des Gerichtlichen Proceß ... Das Annder Theyl / Von herkommen / art vnd gestalt aller Action / oder Klagen ... Das Dritt Theyl / Von allen vnd ieden Exception oder Außzügen ..., Frankfurt/Main 1536 (VD16: G 2296).

Gobler,GerProz. 1542: Justinus Göbler, DEr Gerichtlich Proceß / Auß geschribenen Rechten / vnd nach gemeynem im Heyligen Reich Teutscher Nation gebrauch vnnd vbung / in zwey theyl verfaßt / deren Erster teyl inhelt die ware vnd recht Practicen aller vnd ieder Gerichtlicher Terminen ... Der Ander teyl helt innen die Theorica ... Jetzt von newem / vnnd hieuor der gestalt im Truck nit mehr außgangen, Frankfurt/Main 1542 (VD16: G 2298) [nicht im ausgewerteten Korpus].

Gobler,GerProz. 1549: Justinus Göbler, Der Gerichtlich Proceß/ Auß geschribenen Rechten/ vnd nach Gemeynen im Heyligen Reich Teutscher Nation/ gebrauch vnd ůbung. Erstlich die Practic Gerichtlicher Terminen ... Zum andern die Theorica ... Zusampt dem Proceß in Peinlichen / Criminal vnd Malefitz Gerichten. Jetzt von newem/ vnd hieuor der gestalt im Truck nit mehr außgangen. , Frankfurt/Main 1549 (VD16: G 2299) [nicht im ausgewerteten Korpus].

Gobler,GerProz. 1562: Justinus Göbler, Der Gerichtlich Proceß/ Auß geschribenen Rechten / vnd nach Gemeynem / im Heyligen Reich Teutscher Nation / gebrauch / vnd ůbung. Erstlich die Practic Gerichtlicher Terminen ... Zum andern die Theorica ... Jetzt vffs newe ersehen vnd gebessert Durch Herrn Justinum Goblerum ..., Frankfurt/Main 1562 (VD16: G 2301) [nicht im DRQEdit-Korpus].

Gobler,Inst. 1552: Justinus Göbler (Bearb.), KEyserlicher vnd des H. Reichs Rechten / die Vier Bůcher der Jnstituten vnd Vnderweisung Keysers Justiniani. Mit einfůrung Natürlicher / beschribenen Geystlich vnnd Weltlichen Rechte vnnd Billicheyt. Teutscher Nation Constitutioñ vnd Gebrůauchen. ... Auffs new verteutscht vnd außgelegt / Durch D. Justin Gobler / von Sant Gewere, Frankfurt/Main 1552 (VD16: C 5242).

Gobler,Rsp. 1550: Justinus Göbler, Der Rechten Spiegel: Auß den beschribenen Geystlichen / Weltlichen / Natürlichem / vnd andern gebreuchlichen Rechten / Auch gemeynen im Heiligen Reich Teutscher Nation / Constitutionen / vnd ůbungen / zůgericht ... Durch Herrn Justin Goblern / von Sanct Gwere / der Rechten Doctorem / vnd Nassawischen Catzenelnbogischen Rath beschriben, Frankfurt/Main 1550 (VD16: G 2313).

Gobler,StatB. 1553: STatuten Bůch / Gesatz Ordnungen vnd Gebrůauch / Kaiserlicher Allgemainer / vnd etlicher Besonderer Land vnd Stett Rechten, Frankfurt/Main 1553 (VD16: D 704).

GoldBulle(Gobler) 1550 = Gobler,Rsp. 1550, Bl. 237 v - 251 v.

HadelnLR. 1583 (1584): VErordnung des Hadelerschen Landtgerichts vnnd Rechtens / zu fortsetzung vnd Erhaltung der Ordentlichen Iustitien, vnnd gleichformiges Rechtens verfasset vnd Publicirt. ..., Hamburg 1584 (VD16: B 7430).

HeilbronnStat. 1541: Statuten / Satzung / Reformation vnd Ordnung / Burgerlicher Pollicey des Heyligen Reychßstat Haylpronn, Nürnberg 1541 (VD16: H 1396).

HennebergLO. 1539: Johannes Gemel (Bearb.), Landts Ordnung der Fürstlichen Graffschafft Hennenberg. [von (Johann Gemeln der Rechten Doctori / ... in gegenwertige Ordnung gebracht)], Nürnberg 1539 (VD16: H 1924).

HessenRef. 1524: REFORMation Gesetze vnd ordnung vnser von Gots gnaden / Philipssen Lantgrauen zu Hessen / Graue zu Catzennelnbogen 7c. ANno domini millesimo quingentesimo. Vicesimo Quarto, Erfurt 1524 (VD16: H 2873).

HirschbergLGO. 1518: Ettlich nottürfftig freihaitbrief vñ ordnung über das Landtgericht Hirsperg Anno M D xviij, München 1518 (VD16: B 983).

Hugen,Rhetor. 1528: Alexander Huge, Rethorica vnnd Formularium Teütsch / der gleich nie gesehen ist / bey nach all schreyberey betreffend / von vilerley Episteln / vnder vnd überschrifften / allen Geistlichen vnd Weltlichen ... Ein gantz gerichtlicher proceß ... Darauß die jungen / beinach alle schreyberey leichtlich lernen / vnd die erfarnen ... wol vnderweisen mǒgen. ..., Tübingen 1528 (VD16: H 5808).

JenaHofGO. 1566: Der Durchlauchtigen Hochgebornen Fůrsten vnd Herrn / Herren Johans Fridrichen des Mittlern / vnd Herren Johans Wilhelmen / Gebrůdere / Hertzogen zu Sachsen ... Newe Hoffgerichts Ordnung / so Jerlich zu vier vnterschiedenen zeiten / in jrer F. G. Stad Jhena besatzt vnd vie es darinnen allenthalben gehalten werden solle. Anno M. D. LXVI. auffgericht, Jena 1566 (VD16: S 1103).

JülichEdikt 1554 (1555) = JülichOrdn. 1555, S. 157 - 169.

JülichEdikt 1554 (1556) = RavensbergOrdn. 1556, S. 44 - 60.

JülichGerichtsgefällO. 1555 = JülichOrdn. 1555, S. 118 - 124.

JülichLehenGO. 1555 = JülichOrdn. 1555, S. 141 - 156.

JülichMannhäuserO. 1555 = JülichOrdn. 1555, S. 125 - 140.

JülichOrdn. 1555: VAn Gotteß gnaden Vnser Wilhelms Hertzogen zu Gülich / Cleue vnd Berge ... Ordnung vnd Reformation des Gerichtlichen Proceß / sampt ercklerung etlicher Felle ... Wie es darmit hinfürter in vnsern Fürstenthumben vnnd Landen Gülich vnd Berg / gehalten ... werden soll ... Dergleichen wie es an vnsern Manheusern in Lehensachen zu halten / neben dem Edict so wir hiebeuor in bestimpten vnsern Landen außgehen lassen, Köln 1555 (VD16: K 1289).

JülichRef. 1555 = JülichOrdn. 1555, S. 3 - 117.

JütLowbock 1590 (1593): Blasius Ekenberger (Bearb.), Dat Rechte Judske Lowbock Anno 1590 auergesehn / Corrigeret / vnde in dem Densken vorbetert: tho Copenhagen in Druck vthgegahn. Nu e̊uerst vth dem Densken in de Holsteinische Sprake / van Worde tho Worde / Alse dat beiden Spraken am negesten hefft geschehn mo̊gen / Vp dat trůwlikeste gebracht vnde vmmegesettet, Schleswig 1593 (VD16: D 30).

KärntLGO. 1577 (1578): DEs Ertzhertzogthumbs Khärndten New aufgerichte Landtgerichtsordnung / Jm ain tausend fůnffhundert vnd Sibenvndsibentzigisten Jar, Graz 1578 (VD16: K 4).

KärntLRO. 1577 (1578): DEs Ertzhertzogthumbs Khärndten New aufgerichte Landtßrechtsordnung / Jm ain tausend fůnffhundert vnd Sibenvndsibentzigisten Jar, Graz 1578 (VD16: K 5).

Kistner,Form. 1584: Nikolaus Cisnerus, Formular Allerley Gewålden / Tutorien / Curatorien / Actorien / so im Hochlo̊blichen Keyserlichen Cammergericht eyngebracht: Jetzt aber ... ordentlich zusammen getragen ..., Mainz 1584 (VD16: C 3956).

Klagspiegel(Brant) 1516: Sebastian Brant (Hg.), Der Richterlich Clagspiegel. Ein nutzbarlicher begriff: wie man setzẽ vñ formierẽ sol nach ordenũg der rechtẽ ein yede clag / antwort / vñ vßzesprechene vrteylẽ / gezogẽ vß geistlichẽ vñ weltlichen rechtẽ ... Durch doctorem Sebastianum Brandt wider durchsichtiget vnnd zům teyl gebessert, Straßburg 1516 (VD16: B 7085).

Kolle,LübR. 1586: Joachim Kolle (Bearb.), Ein Rechtbuch Darinne die Artikele / so man Lůbisch Recht nennet vnd in den manuscriptis Exemplaribus gefunden Nicht alleine in eine bequeme vnd richtige Ordenung gebracht / Besondern auch das Sechsische / Keyserliche vnd Go̊ttliche Recht zugleich mit eingefůhret vnd angezogen. ..., Hamburg 1586 (VD16: L 3162).

KölnErzstiftExekProz. 1593: Executions Process Wie sich alle Ambtleuth Vo̊gt / Schultheiß / Scheffen / Pastorn, vnd Gerichtsbotten / auff anruffen des Geistlichen Richters bey den Executionibus in dem Ertzstifft Co̊lln verhalten sollen, Münster 1593 (VD16: ZV 19159).

KölnErzstiftPolLO. 1596: DEß Ertzstiffts Co̊lln Pollicey vnd LandsOrdnung. DVrch ... Herrn Ernsten Ertzbischoffen zu Co̊lln vnd Churfůrsten ... Pfaltzgraffen bey Rheyn / in Ober: vnd Nieder Bayern ... auffgerichtet, Münster 1596 (VD16: ZV 9094).

KölnErzstiftRef. 1538: Des Ertzstiffts Co̊ln Reformation. Dere weltlicher Gericht Rechts / vnd Pollicey. Durch ... H. Herman Ertzbischoffen zů Co̊ln ... vffgericht. Anno M. D. XXXVII, Köln 1538 (VD16: K 1740).

KölnOrdn. 1562: Abdruck vnd gemeiner begriff der Pollicey / Ordnungen / Plebisciten / vnnd Statuten der alten Lo̊blichen Freyen ReichsStadt Co̊llen etc, o. O. 1562 (VD16: ZV 9096).

KölnStat. 1437 (1562) = KölnOrdn. 1562, Bl. 1 r - 45 r.

KölnVerbundbrief 1396 (1562) = KölnOrdn. 1562, Bl. 45 v - 51 v.

König,Proz. 1541 = SammelwerkSächsR.(Wolrab) 1541, Blatt [A iiij] r - Bl. 307 r.

KrainLGO. 1535: Des Hertzogthumbs Crain \ vnd der angeraichten Herrschafftn vnd Grafschafften der Windischen March \ Meetling \ Ysterreich \ vnd Karst \ LanndtgerichtßOrdnung, Wien 1535 (VD16: K 2210).

KulmbachHofGO. 1543 = KulmbachHofuOHofGO. 1543, Blatt a iij r - f ij r.

KulmbachHofuOHofGO. 1543: Des Durchleuchtigen ... herrn Albrechts Marggrauen zu Brandenburg ... Hof vnd Ober Hofgerichts Ordnung aufm Gebirg, Nürnberg 1543 (VD16: B 6990).

KulmbachOHofGO. 1543 = KulmbachHofuOHofGO. 1543, Blatt f [iij] r - g iij v.

LaiischeAnzeigung 1531: Ain laijsche / anzaigung / So allen Landsässen / vnd denen / die ordenlich / oder beuolhen / oberkhait haben / als Hofmarch / vnd gerichtsherren ... auch der Stott / vnd schrannenrednern ... auch in gemain / allen jnwonern / des loblichen hauss / vnd Fürstenthumbs Bairn / zů dienst ... in druckh / geben worden. ..., München 1531 (VD16: L 123).

LeipzOHofGO. 1529: Sechsische oberhofgerichts Ordenung, Leipzig 1529 (VD16: S 943).

LeipzOHofGO. 1548 (1549): Ordenunge des Churfürstlichen Sechssischen Obernhofegerichts, Leipzig 1549 (VD16: S 823).

Lettscher,Notariat 1576: Samuel Lettscher, Notariat bůch der Kunst / so zuuor dergleichen nie nit in Teutscher sprach gesehen / noch in Truck gebracht worden ist ... Mit angehenckten Coloribus Rethoricalibus ..., Dillingen 1576 (VD16: L 1321).

LothringenAssiseProzO. 1599 = LothringenOrdn. 1599, [II] S. 3 - 27.

LothringenBalleiProzO. 1599 = LothringenOrdn. 1599, [II] S. [28] - 83.

LothringenLandsbr. 1599 = LothringenOrdn. 1599, [I].

LothringenNeueBr. 1599 = LothringenOrdn. 1599, [II] S. 98 - 108.

LothringenOrdn. 1599: Die gemeine Landtßbrauche der dreyen / Nemlich / Nancaischen / Vogischen / vnd Teutschen Ballisthůmben in Lotharingen. ... durch Johan Huart ... erstlich verdolmetschet / vnd folgendts durch etliche ... vbersehen vnd verbessert. (Titel zu den enthaltenenen Texten: Kurtzer Begriff / der Form vnd Weise / so in Auffrichtung der Processen ... der Nancaischen / Vogischen vnd Teutschen Ballißthůmben gehalten werden solle. Zusambt der Ordnung vnd Regulierung vber der Richtern / Procuratorn vnd andern Gerichts Personen / Vacationen vnd Belohnungen.), Frankfurt/Main 1599 (VD16: L 2852).

LothringenTaxO. 1599 = LothringenOrdn. 1599, [II] S. [84] - 93.

LübeckStat. 1586: Der Kayserlichen Freyen vnd des Heiligen ReichsStadt Lůbeck Statuta vnd Stadt Recht. Auffs Newe vbersehen / Corrigiret / vnd aus alter Sechsischer Sprach in Hochteudsch gebracht, Lübeck 1586 (VD16: L 3163).

LübR. 1509: Ludwig Dietz (Bearb.), [ohne Titel, Textbeginn: Eyne vorrede dusses bokes], Rostock 1509 (VD16: L 3161).

LünebStat. 1594: Dassel, Hartwig von (Hg.), Der Fürstlichen lőblichen vnd weitberhůmbten Stadt Lůneburgk Statuta vnd Stadtrecht. Aus einem manuscripto exemplari corrigirt, reuidirt, auch mit kurtzen Concordantijs vnd Glossen der Keyser / Sachsischen / vnd Lůbeckschen Rechten / zu mehrer vnterrichtung locupletirt, sampt einer Vorrede an den gutwilligen Leser, Uelzen 1594 (VD16: L 3187 [enthalten in D 204]).

MainzHofGO. 1516 (1521): MEintzisch hoffgerichts Ordnũg zu allen andern gerichten dienlich. 1521, Mainz 1521 (VD16: M 262).

MainzUGO. 1534: VNdergerichts ordnung des Ertzstiffts Meyntz: iñ welcher gantz fleissig angezeygt / wie ... iñ recht gehandelt vnd procedirt werden soll ..., Mainz 1534 (VD16: M 273).

MecklHofGO. 1568: Reformation vnd Hoffgerichts Ordnung vnser ... Johans Albrechten vnd Vlrichen gebrůdern / Hertzogen zu Meckelnburg ... Auffs newe vbersehen vnd verbessert. Anno M. D. Lxviij, Rostock 1568 (VD16: M 1837).

MecklKirchenGO. 1570: Der Durchleuchtigen Hochgebornen Fůrsten vnd Herren / Herrn Johans Albrechts vnd Herren Vlrichs gebrůdern / Hertzogen zu Meckelnburgk ... Kirchengerichts oder Consistorijordnung ..., Rostock 1570 (VD16: M 1833).

MecklLGO. 1558: REformation vnd Landtgerichts Ordnung Vnserer ... Johans Albrechten / vnd Vlrichen gebrůdern / Hertzogen zu Meckelnburgk ..., Rostock 1558 (VD16: M 1839).

Meurer,Billigkeit 1561: Noe Meurer, Von dem waren oder gerechten Rechten der Teütschẽ Gerechtigkeit / AEQVITATE oder Billigkeit / wie vnd wann die von einem jeden gerechten vnd nicht zůstrengem Richter nach gestalt vnd gelegenheit fürfallender vngleicher fell oder sachen / zůhalten / vnd darnach zůvrtheilen seye., Frankfurt/Main 1561 (VD16: M 5015 [enthalten in M 5017]).

Meurer,Liberey(Rücker) 1597: Noe Meurer / Nikolaus Rücker (Bearb.), Liberey Keyserlicher / Auch Teutscher Nation Landt vnd Statt Recht ... Erstlich Durch Weilandt den Hochgelehrten Herrn Noe Meurer ... verfasset. ... Endtlich durch H. Nicolaum Růcker / der Rechten Doctor / auff ein newes vbersehen ... vnnd mit etlichen newen Statuten vnd Ordnungen ... vermehret / abermals an den Tag gegeben, Frankfurt/Main 1597 (VD16: M 5011).

MindelheimGO. 1536: Confirmation vnd Gerichtsordnũg der Herrschafft Mindelhaim, Augsburg 1542 (VD16: M 5417).

MünsterGemO. 1571: Vnnsers Johans Von Gotts gnaden Bischoffen zu Můnster ... verfaste / vnd durch vnsere Můnsterische Stiffts Stende angenommene / Auch folgents durch die Roм̃. Kay. May. vnsern Allergnedigsten Hern / Bestettigte Můnsterische Gemeine Ordnungen, Münster 1571 (VD16: M 6624).

MünsterHofGO. 1571: Vnnsers Johans Von Gotts gnaden Bischoffen zu Můnster ... verfaste / vnd durch vnsere Můnsterische Stiffts Stende angenommene / Auch folgents durch die Roм̃, Kay. May. vnsern Allergnedigsten Hern / Bestettigte Můnsterische Hoffgerichts Ordnung, Münster 1571 (VD16: M 6625).

MünsterLGO. 1571: Vnnsers Johans Von Gotts gnaden Bischoffen zu Můnster ... verfaste / vnd durch vnsere Můnsterische Stiffts Stende angenommene / Auch folgents durch die Roм̃. Kay. May. vnsern Allergnedigsten Hern / Bestettigte Můnsterische Landtgerichts Ordnung, Münster 1571 (VD16: M 6626).

Murner,Inst. 1519: Thomas Murner (Bearb.), Jnstituten ein warer vrsprung vnnd fundament des Keyserlichen rechtens / von dem hochgelerten herren Thomã Murner ... verdütschet ..., Basel 1519 (VD16: C 5233).

Murner,KaisStatR. 1521: Thomas Murner (Bearb.), Der keiserlichen stat rechten ein ingãg vnd wares fundamẽt. Meister vnd rädten tütscher nation von Doctor Thomas Murner gegabet vnd zů gefallen verteütschet, Straßburg 1521 (VD16: C 5235).

NiederlausitzGO. 1538: Ordnung vnd bestellung der Gericht des Marggraffthumbs Niderlausitz. M. D. XXXVIII, Frankfurt / Oder 1538 (VD16: N 1638).

NÖstExekO. 1572: AJner Ersamen Landschafft des Ertzhertzogthumbs Osterreich vnter der Enns / ExecutionOrdnung, Wien 1572 (VD16: N 1660).

NÖstLGO. 1514 (1528): Hieriñ seind die Artickel der Lãde gericht des Fürstenthũb Osterreich durch die Rőmisch Keyserlich Maiestat 7c. auffgericht, Augsburg 1528 (VD16: N 1640).

NÖstLRO. 1540: Gerichts Ordnung des Lañdsrechten des hochlőblichen Ertzhertzogthumbs Osterreich vnder der Enns, Wien 1540 (VD16: N 1647).

NÖstLRO. 1557: GerichtsProceß vnd ordnung des Landßrechtens des Hochlőblichen Ertzhertzogthumbs Osterreich vnnder der Enns, Wien 1557 (VD16: N 1648).

Notariatbuch 1534: Notariatbůch / Wes einem Notarienn odder Schreiber / aller seiner Praccic / in ieden Sachen / Contracten / vnd Verbrieffungen / zuwissen / zubetrachten / zuuersehen / vnd fürzunemen sei ... . Cantzleibůch / Allerhand Missiuen vnd Schrifften Formlich zustellen, Frankfurt/Main 1534 (VD16: N 1867) [nicht im ausgewerteten Korpus].

NotariatKunst(Nawer) 1502: Andreas Nawer (Bearb.), Kunst deß Notariat vnd wie sich der Notarius in seinem Ampt halten vnd regieren soll. Jst verdeütscht. Durch den ... herren Andressen nawer. Arcium Magister. der tzeit Pfarer zu Lorch ..., Nürnberg 1502 (VD16: A 3824).

NürnbGO. 1549: VErneute vnd gepesserte Gerichts Ordnung zu Nürnberg, Nürnberg 1549 (VD16: N 1975).

NürnbRef. 1479 (1484): [Newe Reformacion der Stat Nurẽberg Nach crist gepurt Tausent vierhundert Vnd in dẽ newnvndsibentzigstẽ Iare furgenomẽ] (Titel aus Überschrift des Inhaltsverzeichnisses konstruiert), Nürnberg 1484 (ISTC: ir00037000).

NürnbRef. 1503: Reformacion der Kayserlichen Stat Nuremberg:, Nürnberg 1503 (VD16: N 2026).

NürnbRef. 1564: Der Stat Nurmberg verneute Reformation, Nürnberg 1564 (VD16: N 2029).

OÖstLGO. 1559: Römischer Kaÿserlicher Maÿestat. 7c. Lanndtgerichts Ordnung des Ertzhertzogthumbs Osterreich des Lanndts ob der Enns, Wien 1559 (VD16: O 73).

OÖstLRO. 1535: ORdnung des Landsrechtens des Ertzhertzogtumb Osterreich ob der Enns, Wien 1535 (VD16: O 75).

OrdoJudiciarius(dt.) 1472: [Ohne Titel. Handschriftliche Kopfzeile auf der 1. Seite: processus Juris. Textanfang: Jn dem namẽ d´ heÿligen vñ vnteilperñ triuål̊tikeÿt Amen Von ordnung ze reden / vñ besund´ zů an gedingtem freüntlichem rechten], [Augsburg um 1472] (ISTC: io00088800) [nicht im ausgewerteten Korpus].

PassauGO. 1536 (1539): Gerichts ordnůng / durch den Hochwirdigen ... herrñ Ernnsten Administratorn des Stiffts Passaw ... Hertzogen in Obern vnnd Nidern Bayrñ 7c. in seyner Fürstlichen genaden Stiffts / Stat Passaw / Auffgericht. M. D. XXXVI, Landshut 1539 (VD16: P 865).

PeinlGO. 1532 (1533): DEs allerdurchleuchtigsten großmechtigstẽ vnüberwindtlichsten Keyser Karls des fünfften: vnnd des heyligen Rů̊mischen Reichs peinlich gerichts ordnung / auff den Reichßtå̊gen zů Augspurgk vnd Regenspurgk / iñ jaren dreissig / vñ zwey vnd dreisssig gehalten / auffgericht vnd beschlossen, Mainz 1533 (VD16: D 1069).

Perneder,Inst. 1544: Andreas Perneder, Institutiones. Auszug vñ anzaigung etlicher geschriben Kaiserlichen vnnd deß heiligen Reichs rechte / wie die gegenwertiger zeiten in vbung gehalten werden: in den Titeln vnderschidlich nach ordnung der vier Bůcher Kaiserlicher Jnstitution gestelt / mit einfůrung Lateinischer allegatiõ daneben auch etlicher Lande vnd Oberkaiten besonderer gewonhaiten vnnd Statuten ... Mit ainer Vorrede des Hochgelerten herrn Wolffgang Hunger ..., Ingolstadt 1544 (VD16: P 1493).

Perneder,Lehnr. 1544 = Perneder,Werke 1544 A, [I] Blatt [A iiij] r - Bl. 41 r.

Perneder,Malef. 1544 = Perneder,Werke 1544 A, [II] Blatt [A] r - Bl. 24 r.

Perneder,Proz. 1544: Andreas Perneder, Gerichtlicher Process / in welchem die gemainen Weltlichen vnd Gaistlichen recht ... allegirt / ... auch ... ... verdolmetschet seind: Jtem des Hailigẽ Reichs vil newe breuchliche ordnung / auch etlicher Land vnd Stett besonder Statut ... Mit ainer Vorrede des Hochgelerten herrn Wolffgang Hunger ..., Ingolstadt 1544 (VD16: P 1476).

Perneder,RegelnSocin. 1544 = Perneder,Werke 1544 B, Bl. 27 r - 59 r.

Perneder,SummaRoland. 1544 = Perneder,Werke 1544 B, Bl. 1 r - [26] v.

Perneder,Werke 1544 A: Andreas Perneder, Der Lehenrecht kurtze vnd aygentliche Verteütschung nit allain auß den Kayserlichen satzungen vnd derselben Texten sonder auch vilen Hochberů̊mpten Doctorn ... gezogen ... Jtem ain Gerichtliche Practica aller Malefitz oder Peinlichen sachen / etc. Durch den ... Herrn Andreas Perneder / des Fürstlichen Hoff zů München Rath vnnd Secretarien. Mit ainer Vorrede des Hochgelerten herrn Wolffgang Hunger ..., Ingolstadt 1544 (VD16: P 1510).

Perneder,Werke 1544 B: SV̄MA ROLANDINA. Das ist ein kurtzer bericht / von allerhand Contracten vnnd Testamenten ... Jtem ein Tractat der Regeln / oder kurtzen schlußreden gemayner Recht ... Bartholomei Soccini. Beides durch ... Herrn Andreas Perneder / des Fürstlichen Hoff zů München Rath vnd Obristen Secretarien / zierlich verteütscht 7c. Mit ainer Vorrede des Hochgelerten herrn Wolffgang Hunger ..., Ingolstadt 1544 (VD16: R 2936).

PfalzHofGO. 1573: HoueGerichts Ordnung Des Durchleuchtigsten ... ... Herrn Friderichen / Pfalntzgrauen bey Rhein ... vnd Churfůrsten ... Wie am Churfůrstlichen Pfalntzgrůischem HoueGericht fůrbaßhin ... procediert / auch die ergangne Vrthail exequiert vnd volstreckt werden sollen, Heidelberg 1573 (VD16: P 2163).

PfalzLO. 1582: ChürFůrstl. Pfaltz Landts Ordnung. ..., Heidelberg 1582 (VD16: P 2205).

PfalzLR. 1582: ChurFůrstlicher Pfaltz LandtRecht, Heidelberg 1582 (VD16: P 2207).

PfalzZweibrGO. 1536: Gerichts Ordnung der Fůrmůnder Hertzog Wolffgangs Pfaltzgrauen 7c., Simmern 1536 (VD16: P 2270).

Pflanzmann,Lehnr. 1493: Jodocus Pflantzman, Das bůch der lehenrecht, Augsburg 1493 (ISTC: ij00603000).

Pölmann,Handbuch 1574: Albert Pölmann, Handtbuch Dariñen in der kůrtze zu befinden / was sich fast teglich bey Gerichte zutregt / Daraus man sich zu erlernen vnd zu spieglen habe / Was die Rechte dauon sagen ..., o. O. 1574 (VD16: P 3824).

Pölmann,UntergerProz. 1577: Albert Pölmann, Der gantze Proceß des ordentlichen Gerichts in Bůrgerlichen Sachen / wie es bey den Vntergerichten dieses Hertzogthumbs Preussen gehalten wird. Durch Albertum Pôlman Publicum Notarium auffs new ausgangen, Königsberg 1577 (VD16: ZV 24180).

Pölmann,Urteil 1577: Albert Pölmann, Die lauffende Vrtel So man teglich bey Gerichte braucht. Durch Albertum Polman Notarium Publicum. Auffs new ausgangen vnd zum teil vermehret, Königsberg 1577 (VD16: ZV 24178).

PommernStettinHofGO. 1566 (1569): Vnser von Gotts gnaden Barnims des Eltern / Johans Friderichs / Bugslaffs / Ernst Ludwigs / Barnims des Jůngern vnd Casimirs / Geuettern vnd Gebrůdere Hertzogen zu Stettin Pommern ... Gerichts Ordnung / wie es inn vnsern Fůrstlichen Hoffgerichten des Stettinischen vnd Wolgastischen orts zuhalten. ..., Stettin 1569 (VD16: P 4131).

PreußHofGO. 1578: Hoffgerichts Ordnung des Hertzogthumbs Preussen: Von dem Durchleuchtigen ... Herrn Georgen Friderichen / Marggraffen zu Brandenburg ... corrigiret / gemehrt vnd gebessert / Anno 1578, Königsberg 1578 (VD16: P 4800).

PreußHofGO. 1583: Hoffgerichts Ordnung des Hertzogthumbs Preussen: Von dem Durchlauchtigen ... Herrn Georgen Friderichen Marggrafen zu Brandenburg ... ... auffs newe corrigiret / gemehret vnd gebessert / Anno 1583, Königsberg 1583 (VD16: P 4801).

RAbsch. 1500 Augsburg = SammelwerkReichsR.(Hölzel) 1500, Blatt B iij r - [F iij] r.

RAbsch. 1521 Worms = SammelwerkReichsR.(Schöffer) 1521, Blatt [aa] r - [bb iiij] r.

RAbsch. 1541 Regensburg: ABschiedt deß Reichßtags zů Regenspurg gehalten ANNO M. D. XLI, Mainz 1541 (VD16: R 785).

RAbsch. 1542 Nürnberg: ABschiedt Deß Reichßtags zů Nürmberg auffgerichtet: im Jar als man zalt nach Christi geburt M. D. XLII. Den XXVI. tag des Monats Augusti geschehen, Mainz 1542 (VD16: R 788).

RAbsch. 1542 Speyer: ABschiedt des Reychßtags zů Speir auffgerichtet / im Jar M. D. XLII, o. O. 1542 (VD16: R 789).

RAbsch. 1548 Augsburg = SammelwerkReichsR.(Schöffer) 1548, [Teil I].

RAbsch. 1555 Augsburg = SammelwerkReichsR.(Behem) 1555, 1. Blattzählung.

RAbsch. 1559 Augsburg: Abschiedt Der Rǒmischen Keyserlichen Maiestat / vnd gemeyner Stende / Auff dem Reichßtag zu Augspurg / Anno Domini M. D. LIX. auffgericht, Mainz 1559 (VD16: R 805).

RAbsch. 1566 Augsburg: Abschiedt Der Rǒmischen Keyserlichen Maiestatt / vnnd gemeiner Stǎndt / auff dem Reichßtag zu Augspurg / Anno Domini M. D. LXVI. auffgericht, Mainz 1566 (VD16: R 809).

RAbsch. 1567 Regensburg: Abschiedt Der Rǒmischen Keyserlichen Maiestatt / vnnd gemeiner Stendt / auff dem Reichstag zu Regenspurg / Anno Domini M. D. LXVII. auffgericht, Mainz 1567 (VD16: R 811).

RAbsch. 1570 Speyer (1571) = SammelwerkReichsR.(Behem) 1571, [unbezeichnet 1] - Bl. 59 r.

RAbsch. 1576 Regensburg: Abschiedt der Rǒmischen Kayserlichen Maiestat / vnd gemeyner Stǎnde auff dem Reichstag zu Regenspurg / Anno Domini M. D. LXXVI. auffgericht, Mainz 1576 (VD16: R 817).

RAbsch. 1594 Regensburg: Abschiedt Der R . Kaÿ. Mt: Vnd gemeiner Stǎndt / auff dem Reichßtag zu Regenspurg / Anno Domini M. D. XCIIII. auffgericht, Mainz 1594 (VD16: R 824).

RavensbergGO. 1556 = RavensbergOrdn. 1556, Blatt [A] r - S. 43.

RavensbergOrdn. 1556: VAn Gotteß gnaden / vnser Wilhelms Hertzogen zu Gulich / Cleue vnd Berge / Grauen zu der Marck vnd Rauenßberg / Herrn zu Rauenstein / 7c. Ordnung des gerichtlichen Proceß / wie es damit hinfurter inn vnser Graffschafft Rauenßberg gehaltenn werden soll / im iar tausent fünfhondert vnnd sechsvndfünftzig außgangen, Düsseldorf 1556 (VD16: K 1288).

RDeputAbsch. 1571 Frankfurt: Abschiedt Der Rǒmischen Kayserlichen Maiestat / auch Churfǔrsten / deputirter Fǔrsten vnd Stende / für sich vnd in namen gemeiner des heiligen Reichs Stende auff dem Deputationtag zu Franckfort Anno Domini M. D. LXXI. auffgericht, Mainz 1571 (VD16: R 700).

RefFreiGer. 1571: Der Freien vnd heimlichen Gerichten Reformation / dauon im dritten Titull / des dritten theils vnser Johans von Gottes gnaden Bischoffen zu Mǔnster ... Landtgerichts / Ordnung relation vnd meldung geschicht vnnd in ermelten Gerichten vnsers Stiffts Mǔnster hinfuro gehaltenn soll werdenn, Münster 1571 (VD16: M 6622).

Reiterbestallung 1570 (1571) = SammelwerkReichsR.(Behem) 1571, Bl. [60] r - 103 r.

Riederer,Rhetorik 1493: Riedrer, Friedrich, Spiegel der waren Rhetoric. vß .M. Tulio .C. vnd andern getutscht : Mit Jrn glidern clǔger reden Sandbriefen / vnd formen menicher contract / seltzam. Regulierts Tútschs vnd nutzbar exempliert / mit fǔgen Vff gǒttlich vnd keiserlich schrifft vnd rechte gegrúndt : nuwlich (vnd vormaln Jn gemein nÿe gesehen) ÿetz loblich vß gangen, Freiburg im Breisgau 1493 (ISTC: ir00197000) [nicht im ausgewerteten Korpus].

RKGO. 1495: Ordenung der romis. ko. ma. Camergericht mit allen seinen punckten vñ artickeln wie das dan vff der versamblung des heiligen Reichs dag zu wormß jm jar .M. cccc.xcv. ... beslossen ist., Mainz 1495 (ISTC: im00396000).

RKGO. 1521 = SammelwerkReichsR.(Schöffer) 1521, Blatt [AA] r - [DD vj] r.

RKGO. 1548: DEr Rǒmischen Key. Mai. vnd gemeyner Stend deß Heyligen Reichs angenommene vnd bewilligte Cammergerichts Ordnung ... auß allen alten Cammergerichts Ordnungen vnd Abschieden / jetzt vff dem Reichßtag zǔ Aupspurg / ... M. D. XLVIII. von newem zǔsammen gezogen ..., Mainz 1548 (VD16: D 991).

RKGO. 1555 = SammelwerkReichsR.(Behem) 1555, 2. Blattzählung mit ungezählten Seiten davor.

RLandfr. 1521 = SammelwerkReichsR.(Schöffer) 1521, Blatt [A] r - [C iiij] r.

RLandfr. 1548: Rǒmischer Keyserlicher Maiestat vnd deß heyligen Reychs Landtfriden / auff dem Reychßtag zǔ Augspurg declariert / erneüweret / auffgericht / vnnd beschlossen Anno Domini M. D. XLVIII. ..., Mainz 1548 (VD16: D 1013).

RLandfrErkl. 1522: ROmischer Kayserlicher Maiestat ordnungen fürsehungen vñ erclerungen / wie allenthalben im hailigen Reich ... wider die manigfeltigen vergweltiger / beschediger / vnd des hayligen Reichs landtfridens verprecher / darzu desselben declarirt Echter ... gehandelt werden sol ..., Augsburg 1522 (VD16: D 1067).

RNotariatsO. 1512: Ordnung von kays'licher Maiestat zu vnd'richtũg der offen Notariẽ wie die jr Ampter vben sollen ausgangẽ: mitsampt eynem penlichẽ mandat das die nymãds nachtrucken: oder ob solichs darwider gescheche: dieselben: nymands: vffkauffen noch verkauffen noch feyl haben solle ..., Mainz 1512 (VD16: D 828).

ROrdn. 1512: ROmischer Keyserlicher Maiestat vñ gemeiner Stende des Reichs vffsatzung vnd ordnung vff dem Reichstag zu Collen. Anno. XVc. Vnd. XII. vffgericht, Straßburg 1512 (VD16: R 753).

RostockGO. 1574: Gerichts Ordnung eines Erbarn Radts der Stadt Rostock. Publiciret Anno M. D. LXXIIII. den 24. Aprilis, Rostock 1574 (VD16: R 3180).

RostockGO. 1586: Eines Erbarn Rhats der Stadt Rostock Newe Gerichtsordnung. Publicirt ANNO M. D. LXXXVI, Rostock 1586 (VD16: R 3181).

RottweilHofGO. (1523): Ordenũg vnd sundere gesatz des heilgẽ römschẽ reichs hofgericht zů rotweil. Auch wie weilet künig Cunrat ein hertzog zů schwaben solchs einr stat rotweil vm̃ irs sond'n v'dienẽs gnediklich gebẽ hat, Straßburg 1523 (VD16: D 764).

RottweilHofGO. 1573: Ernewerte Ordnung. Der R . Kaÿ. Mt. Kaiserlichen Hoffgerichts zü Rottweil, Mainz 1573 (VD16: D 1265).

RPolO. 1548: DEr Rőmischen Keyserlichen Maiestat Ordnung vnd Reformation / gůter Pollicey / ... vff dem Reichßtag zů Augspurg Anno Domini M. D. XLVIII. vffgericht. ..., Mainz 1548 (VD16: D 1059).

RPolO. 1577 (1578): Der Römischen Keyserlichen Maiestat reformirte vnd gebesserte Policey Ordnung / ... auff Anno M. D. LXXVII. zu Franckfort gehaltenem Reichs Deputation tag verfast vnd auffgericht, Mainz 1578 (VD16: D 1295).

RRegimentsO. 1500 = SammelwerkReichsR.(Hölzel) 1500, Blatt A ij r - B ij v.

RRegimentsO. 1521 = SammelwerkReichsR.(Schöffer) 1521, Blatt a ij r - [b iiij] v.

Salwechter,GerProz. 1543: Jakob Salwechter, Gerichtlicher Process EJn vast kurtze Gerichts ordnung vber aus nůtzlich / in Dőrffern vnd Stetten / an den vnder gerichten zugebrauchẽ ..., Frankfurt/Main 1543 (VD16: S 1513).

SammelwerkFranken(Schubart) 1512: Newe Reformation des Lanndtgerichts des Hertzogthumbs zu Franncken. (Titel zum enthaltenen Text: Zusatzung vnd erleuterung auff die vorige Reformation ..., Würzburg 1512 (VD16: W 4539).

SammelwerkReichsR.(Behem) 1555: Abschiedt Der Rőmischen Kőniglichen Maiestat / vnd gemeiner Stendt / auff dem Reichßtag zu Augspurg / Anno Domini M. D. LV. auffgericht. Sampt Der Keyserlichen Maiestat Cam̃ergerichts Ordnung wie die auff diesem Reichßtag / durch die Kőnigliche Maiestat / vnd gemeine Stendt / widerumb ersehen ernewert / vnd an vilen orten geendert, Mainz 1555 (VD16: R 801).

SammelwerkReichsR.(Behem) 1571: Abschiedt der Rőmischen Kayserlichen Maiestat / vnd gemeiner Stånde auff dem Reichstag zu Speyr / Anno Domini M. D. LXX. auffgericht. (Der Römischen Kayserlichen Maiestat / vnnd deß heyligen Reichs reutter bestallung ...), Mainz 1571 (VD16: R 813).

SammelwerkReichsR.(Hölzel) 1500: Hernach volget dye verschreibung : so des Reichs Regiments hilff vnd ordnung halben : auff dem Reichs tag zu Augspurg beschlossen vnnd auffgericht ist. (Titel zum enthaltenen Text: Abschid des Reichs tags zu Augspurg. Anno domini Tausent funffhundert.), [Nürnberg 1500] (ISTC: im00395000).

SammelwerkReichsR.(Schöffer) 1521: ROmischer kayserlicher Maiestat Regiment Camergericht lantfride vnd Abschied. vff dem Reichstag zu wormbs Anno M vc XXj. beschlossen vnd auffgericht, Mainz 1521 (VD16: R 759).

SammelwerkReichsR.(Schöffer) 1548: ABschiedt Der Rom. Keys. Maiest. vnd gemeyner Stend / vff dem Reichßtag zů Augspurg vffgericht / Anno Domini M. D. XLVIII. Resolution vnd Erklerung der Rom. Key. Maie. Wie es der Religion halben / biß nach endung deß Concilij gehalten werden soll ... inn Lateinischer vnd Teütscher sprach. Key. Maiest. Reformation / den Geystlichen Standt betreffendt. Landtfriden ... Cammergerichts Ordnung ... sampt der Güldin Bull / inn Latein ... mit etlichen andern Constitutionibus, Vff hieuor gehalten Reichßtagen beschlossen. Reformation vnd Ordnung gůter Pollicey ..., Mainz 1548 (VD16: R 796)[Teile II-IV und VII nach der Zählung im VD 16 nicht im ausgewerteten Korpus].

SammelwerkSächsR.(Wolrab) 1541: Kilian König, Processus vnd Practica der gerichtsleuffte / nach dem gebrauch Sechsischer Landart / aus den gemeinen Bapstlichen / Keiserlichen vnd Sechsischen Rechten / Durch D. Chilianum Konig etwan zusamen gezogen / jtzund zum andern mal auffs new corrigirt / vñ mit vil nůtzichen guten Additionen ... (Titel zum enthaltenen Text: Von dem Baume der angebornen Mageschafft / wie man nach Sachsenrecht Erbe nimpt vnd gibt / sampt den Regeln D. Thammonis von Boxdorff ...), Leipzig 1541 (VD16: K 1847).

Saur,Fasc. I (1589): Abraham Saur, FASCICVLVS IVDICIARII ORDINIS SINGVLARIS: Das ist: Ein schoner Außbund: Etlicher Chur- vnd Fursten Gerichts ober vnd vnder / auch Grafen vnd Herrn LandOrdnung / deßgleichen vorneffer ReichsStatten ernewerten Reformationen vnd Processen in Burgerlichen Rechtsachen ... in ein Corpus zusammen gebracht vnd abgetheilt in acht Fascicul / vnnd in zween vnterschiedliche Theile getheilet. ... Durch Angeben deß ... Rechtserfahrnen Herrn / M. Abrahami Saurij, Furstlichen Hessischen Hoffgerichts zu Marpurgk / verordneten Aduocaten vnd Procuratorn ... Das erste Theil, Frankfurt/Main 1588 /1589 (VD16: S 1904) [nicht im ausgewerteten Korpus].

Saur,Fasc. II (1589): Abraham Saur, FASCICVLI IVDICIARII ORDINIS SINGVLARlS Das Ander Theil. Etlicher Chur- vnd Fursten Gerichts ober vnd vnder / auch Grafen vnd Herrn deßgleichen vornemer Reichs Statten Reformationen / den Gerichtlichen Proceß in Burgerlichen Sachen begreiffende / zierlich / vnd zu befurderung der geliebten IVSTITIEN, in drey Fascicul abgetheilet vnd zusammen in ein Corpus gebracht. Durch angeben deß ... Rechtserfahrnen Herrn M. Abrahami Saurij, Fürstlichen Hessischen Hofgerichts zu Marpurgk verordneten Advocaten vnd Procuratorn ..., Mainz (Verlag: Frankfurt/Main) 1589 (VD16: S 1905) [nicht im ausgewerteten Korpus].

SausenbergLO. 1582: Landtsordnung Der Landtgraueschafft Sausemberg vnd Herrschafft Rotlen, Basel 1582 (VD16: ZV 989).

SchlesLO. 1577: DEr Romischen Kayserlichen ... Mayest: 7c. Confirmation etlicher durch die Herrn Fursten vnd Stende in Ober vnd Nieder Schlesien / auff gemainen Fursten vnd Landtagen / dem Lande zu nutz vnd gutten auffgerichte Policey vnd Ordnung, o. O. 1577 (VD16: ZV 19434).

SchleswigLGO. 1573: Vnser von Gottes gnaden Friederichen des andern zu Denmarcken ... Konig / Vnd ... Johansen des Eltern / vnd Adolffen Erben zu Norwegen / aller Hertzogen zu Schleßweig ... Landtgerichts Ordnung / zu befurderunge der ordentlichen Justitien vnd Rechtens in vnsern Furstenthumben Schleßweig Holstein vnd Stormarn verfasset ..., Hamburg 1573 (VD16: S 2977).

Schubeus,Erbschaft 1597: Aegidius Schubeus, Kurtzer DOch Grundtlicher Bericht von Erbschafft / So einer ohne Testament verstirbt / Aus den allgemeinẽ Keyserlichen / Sechsischen / Culmischen vnd Lubischen Rechten vnd Statuten gezogen / vornemblich auff die Ansehe Stedte ... gerichtet ..., Stettin 1597 (VD16: ZV 14209).

SchwabenLGO. 1562: Der Rom: Kaÿ: Maÿestat 7c. Reformation Jrer May: Landtgerichts Jn Obern Vnnd Nidern Schwaben, Dillingen 1562 (VD16: ZV 24519).

SchwäbHallRef. 1573: DEs Hey.Rom. Reichs Statt Schwabischen Hall / Reformation / Ernewerung vnnd Erclarung / alter wolhergebrachter Stattrechten / Gebrauch / Satzung vnd Ordnungen ..., Tübingen 1573 (VD16: S 4562).

Schwartzkopf,DiffIur. 1586: Ludwig Fachs / Georg Schwartzkopff (Bearb.), DIFFERENTIAE IVRIS CIVILIS ET SAXONICI. Das ist Vnderscheide der Keiserlichen vnd Sechsischen Rechte ... in die Deutsche Sprach vorsetzet ..., Helmstedt 1586 (VD16: ZV 5760).

SiebbLR. 1583: Der Sachssen inn Siebenbuergen: STATVTA: Oder eygen Landtrecht. Durch Matthiam Fronium vbersehen/ gemehret vnd ... in Druck gebracht, Kronstadt (Siebenbürgen) 1583.

SponheimHGHofGO. 1586 (1587): HOffgerichts Ordnung der Hindern Graffschafft Spanheim. ..., Frankfurt/Main 1587 (VD16: P 2282).

SponheimHGUGO. 1544: DEr durchleüchtigen Hochgebornẽ beider Fürsten Grauen zu Spanheim / vndergerichts Ordnung / Jn dero Fürstlichen gnaden Hindern Graueschafft Spanheim, Speyer 1544 (VD16: P 2262).

SponheimHGUGO. 1578: Vndergerichts Ordnung der hindern Graueschafft Spanheym. Dabej etliche Statuta vnd Satzungen in Successionen oder Erbfållen / Einkindschafften / Abdrieb oder Losungen / Kauffen vnd Verkauffen / ... Auch das man sich in Malefitzsachen des Reichs Peinlichen Gerichts Ordnung nach verhalten soll, o. O. 1578 (VD16: P 2281).

SponheimVGGO. 1530: DEr Durchleuchtigstenn Durchleuchtigen / Hochgebornen Fůrsten vnd herren / herrnn Ludwigs ... herrn Johansen / beden Phaltz grauẽ bei Reine / Hertzogen in Beyern / Graue zu Spanheym 7c vnd herrnn Philipsen Marggrauen zu Baden 7c / gemein ordnũg jrer gnaden Houegerichts zu Creůtzenach. Auch der Vndtergericht ... der Fo̊rdern Graueschafft Spanheim / gen Creůtzenach geho̊rig. Wie daselbst durch die partheiẽ ... gemeynem Rechten gemeß / procedirt / vnd durch die Richter vnd Scho̊ffenn gehandelt werden sol ..., Simmern 1530 (VD16: ZV 12386).

SponheimVGHofGO. 1530 = SponheimVGGO. 1530, Blatt A iij v - [C v] v.

SponheimVGUGO. 1530 = SponheimVGGO. 1530, Blatt D r - [F vj] r.

SteirBergRB. 1543 (1583): Ro̊mischer Keyserlicher auch zu Hungern vnd Behaim / Ko̊niglicher Mayestat / 7c. Ertzhertzog zu Osterreich / 7c. Confirmation vnd bestettung des Fůrstenthumbs Steyr PerckrechtsBůchel, Augsburg 1583 (VD16: S 8768).

SteirLRRef. 1533: Des lo̊blichen Fürstenthum Steyr beståttũg der Newen Reformation des Lanndsrechtens daselbst, Wien 1533 (VD16: S 8764).

SteirLRRef. 1574 (1575): Ainer Ersamen Landschafft des Lo̊blichen Fürstenthumbs Steyr / New verfaste Reformation des Landts vnd Hofrechts daselbst / Jm M. D. LXXIIII. Jar auffgericht, Augsburg 1575 (VD16: S 8772).

SteirPGO. 1574 (1575): Des Loblichen Fürstenthumbs Steyer Landt vnd Peindlich Gerichts Ordnung Jm M. D. LXXIIII. Jar / verpessert / erleüttert / verglichen vnd auffgericht ..., Augsburg 1575 (VD16: S 8769).

StraßburgGO. 1597 (1598) = StraßburgOrdn. 1598, Bl. 1 r - 16 v.

StraßburgOrdn. 1598: Ernewerte Ordnungen Eines Ehrsamen Rhats der Statt Straßburg / Von Gerichten vnd Gerichtlichen Processen. Mit anhangender verbesserter Ordnung der Procuratoren, Straßburg 1598 (VD16: S 9419).

StraßburgProkuratorO. 1598 = StraßburgOrdn. 1598, Bl. 17 r - 23 r.

Stumphart,Proz. 1541: Friedrich Stumphart, TEutscher Process / weltlichs Burgerlichs Rechtens / mit allen notturfftigen Formen der klagen / antwurten / vnd aller anderer furtråge ..., Tübingen 1541 (VD16: S 9878).

TirolLO. 1526: Der Fürstlichen Grafschaft Tirol Landsordnung, Augsburg 1526 (VD16: T 1355).

TirolLO. 1532: Lanndtßordnung / der Fürstlichen Grafschafft Tirol, Augsburg 1532 (VD16: T 1356).

TirolLO. 1573 (1574) = TirolOrdn. 1574, LO. Blatt B r - LO. Blatt [Hh ij] r.

TirolOrdn. 1574: New Reformierte Landsordnung der Fürstlichen Grafschafft Tirol. (Fürstlicher Durchleuchtigkayt Ertzhertzog Ferdinanden zů Osterreich / Hertzogen zů Burgundi 7c. Grafen zů Tirol 7c. Ordnung vnd Reformation gůter Policey / in jrer Durchleuchtigkait Fürstlichen Grafschafft Tirol.), Augsburg 1574 (VD16: T 1361).

TirolPolO. 1573 (1574) = TirolOrdn. 1574, PolO. Blatt [A] r - PolO. Bl. 29 r.

TrierUGO. 1537: VNdergerichts ordnung des Ertzstiffts Thrier / durch den Hochwirdigsten ... Herrn Johansen Ertzbischouen zů Thrier ... gegeben Jm Jare / M. D. XXXVII, Mainz 1537 (VD16: T 1938).

UlmOrdn. 1579: DEr Statt Vlm Gesatz vnnd Ordnungen / wie Es inn der Statt / vnd derselben Herrschafft vnd Oberkeit. I. BEy den vnuerdingten vnd verdingten Heuraten / der Anfőlligen / Heurat vnd anderer Gůter halber. II. JNn verwaltung aller Pflegschafften. III. MJt vffrichtung der Testament / Donation / vnd anderer vermechtnussen. IIII. BEy den Vnderpfanden / befreyten vnd vnbefreyten Schuldtforderungen. V. VNnd dem Proceß am Gandt oder Frongericht / gehalten werden solle, Ulm 1579 (VD16: U 57).

WaterR. 1555: Dat Water Recht Vnde Schypp Ordeninge. Wo sick de holden schőlen / So beyde der Ost vnde West See / gebruken: Allen Schyppern / Koplůden / Handelern vnde Bosslůden / tho nutte vnde wolgeuallen / yn den Druck gegeuen. ..., Magdeburg 1555 (VD16: ZV 21267).

Weidmann,Lehnr. 1530: Lorenz Weidmann (Bearb.), DJe Lehenrecht verteütscht: auch iñ eyn newe vnd richtige ordnung der titel gesatzt: vnd zůsamen bracht. Mit erklerung vnd außlegung etlicher Lateinischer vnd Welscher wort welch nit fůglich iñs teutsch haben verăndert mőgen werden, Mainz 1530 (VD16: D 727).

WimpfenRef. 1544: REformation vñ Ordnung / Altenherkomens vnd Rechtens / Auch etlicher Newgesetztẽ Statuten der Statt Wympffen, Nürnberg 1544 (VD16: W 3327).

WittenbergHofGO. 1550: Ordenung des Churfůrstlichen Sechsischen Hoffgerichts zu Wittenberg, Wittenberg 1550 (VD16: ZV 13581).

WormsRef. 1498 (1499): Der Statt Wormbs Reformation, [Speyer] 1499 (ISTC: ir00040000).

WürtHofGO. 1557: Des Fürstenthumbs Würtemberg hieuor außgangne / vnnd jetzo von newem gebesserte vnd gemehrete Houegerichtsordnung / wie es künfftiglich in den Hăndeln / an dasselbig erwachsend / gehalten werden solle, Tübingen 1557 (VD16: ZV 20127).

WürtHofGO. 1587: Des Fůrstenthumbs Wůrtemberg hieuor außgangne / vnd jetzo widerumb von newem gebesserte vnd gemehrte Hoffgerichts Ordnung ..., Tübingen 1587 (VD16: W 4529).

WürtLO. 1515: Vlrich von gotes gnaden hertzog zů Wirtemperg vnd zu Tegke / graue zů Mümppelgart 7c. UNsern grus zu vor lieben getreuẽ ..., Tübingen 1515 (VD16: W 4455).

WürtLO. 1536: Des Fürstenthumbs Wirtemberg newe Landsordnung, Tübingen 1536 (VD16: W 4453).

WürtLO. 1552 = DEs Fürstenthumbs Wirtemberg newe Landtsordnung / gebessert vnd gemehret / sampt darzůgedruckten der armen Casten / auch Holtz vnnd Vorst ordnungen, Tübingen 1552 (VD16: W 4509), 1. Blattzählung.

WürtLR. 1555: New landtrecht des Fürstenthumbs Würtemberg / jn vier Theil verfaßt. ..., Tübingen 1555 (VD16: W 4513).

WürtLR. 1567: Des Fürstenthumbs Würtemberg gemein Landtrecht / in vier Theil verfaßt. ..., Tübingen 1567 (VD16: W 4514).

WürzbRefGeistlGer. 1512 = SammelwerkFranken(Schubart) 1512, Blatt ij r - Schluss.

Zasius,Lehnr.(Lauterbeck) 1553: Ulrich Zasius / Georg Lauterbeck (Bearb.), Die Summa des gantzen Keyserlichen Lehenrechtens durch Doctor Vlrich Zasium / mit sonderm vleiß zůsamen gezogen / vnd jetzt newlich ins deutsch gebracht ... Durch Georgium Lauterbeken / Syndicum zůr Naumburg, Basel 1553 (VD16: Z 164).

Zwengel,Form. 1568: Johann Peter Zwengel, New Groß Formular vnd vollkommlich Cantzlei Bůch / von den besten vnd außerlesenen Formularien aller deren Schrifften / so in ... fůrnemen Cantzleyen ... brăuchlich seindt. Sampt allem andern zu den Cantzleyen ... dienstlichen ... Vnderricht ..., Frankfurt/Main 1568 (VD16: Z 715).

Literatur

Abouelhoda/Kurtz/Ohlebusch 2004: Mohamed Ibrahim Abouelhoda / Stefan Kurtz / Enno Ohlebusch, Replacing suffix trees with enhanced suffix arrays, in: Journal of Discrete Algorithms 2 (2004) Nr. 1, S. 53-86.
DOI: *10.1016/S1570-8667(03)00065-0*.

Aczel 2008: Richard Aczel, Art. „Intertextualität und Intertextualitätstheorien", in: Nünning (Hg.) 2008, S. 330-332.

Argamon/Levitan 2005: Shlomo Argamon / Shlomo Levitan, Measuring the Usefulness of Function Words for Authorship Attribution, in: The Association for Computers and the Humanities / The Association for Literary and Linguistic Computing, ACH / ALLC 2005. The International Conference on Humanities Computing and Digital Scholarship. The 17th Joint International Conference. University of Victoria June 15 – June 18, 2005, Conference Abstracts (2nd Edition), 2005, S. 4-7.
URL: *http://web.uvic.ca/hrd/achallc2005/ach_allc_2005_abstracts_2nd_edition.pdf*.

Bader 1954/1984: Karl Siegfried Bader, Zur rechtshistorischen Quellenlehre, in: Zeitschrift für Schweizerisches Recht, N. F. 73 (1954), S. 261-278 (ND in: Karl Siegfried Bader, Ausgewählte Schriften zur Rechts- und Landesgeschichte. Bd. 1. Schriften zur Rechtsgeschichte, ausgew. und hrsg. v. Clausdieter Schott, Sigmaringen 1984, S. 71-88).

Basile u. a. 2009: Chiara Basile / Dario Benedetto / Emanuele Caglioti / Giampaolo Cristadoro / Mirko Degli Esposti, A plagiarism detection procedure in three steps: selection, matches and „squares", in: Stein u. a. (Hg.) 2009, S. 19-23.

Bedenbender 2013: Almuth Bedenbender, Fassungen des Kalumnieneides in frühneuhochdeutschen Rechtstexten. Eine Untersuchung auf der Basis von Quellen in DRQEdit, in: Deutsch (Hg.) 2013, S. 315-340.

Bein 1998: Thomas Bein, „Mit fremden Pegasusen pflügen". Untersuchungen zu Authentizitätsproblemen in mittelhochdeutscher Lyrik und Lyrikphilologie, Berlin 1998 (Philologische Studien und Quellen 150).

Bein 1999: Thomas Bein, Zum ›Autor‹ im mittelalterlichen Literaturbetrieb und im Diskurs der germanistischen Mediävistik, in: Fotis Jannidis / Gerhard Lauer / Matias Martinez / Simone Winko (Hg.), Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs, Tübingen 1999 (Studien und Texte zur Sozialgeschichte der Literatur 71), S. 303-320.

Berndt/Tonger-Erk 2013: Frauke Berndt / Lily Tonger-Erk, Intertextualität. Eine Einführung. Mit einer Auswahlbibliographie von Sebastian Meixner, Berlin 2013 (Grundlagen der Germanistik 53).

Böckenhauer/Bongartz 2003: Hans-Joachim Böckenhauer / Dirk Bongartz, Algorithmische Grundlagen der Bioinformatik. Modelle, Methoden und Komplexität, Stuttgart/Leipzig/Wiesbaden 2003.

Boockmann u. a. (Hg.) 1998: Hartmut Boockmann / Ludger Grenzmann / Bernd Moeller / Martin Staehelin (Hg.), Recht und Verfassung im Übergang vom Mittelalter zur Neuzeit. I. Teil. Bericht über Kolloquien der Kommission zur Erforschung der Kultur des Spätmittelalters 1994 bis 1995, Göttingen 1998 (Abhandlungen der Akademie der Wissenschaften in Göttingen. Philologisch-historische Klasse. 3. Folge 228).

Boukhaled/Sellami/Ganascia 2015: Mohamed-Amine Boukhaled / Zied Sellami / Jean-Gabriel Ganascia, Phoebus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires. 22ème Conférence sur le Traitement Automatique des Langues Naturelles, 2015, Caen, France. 2015. <hal-01198411>.
URL: *http://hal.upmc.fr/hal-01198411*.

Bourdaillet 2009: Julien Bourdaillet, Alignement monolingue avec recherche de déplacements pour la critique génétique, in: Traitement Automatique des Langues 50 (2009), Nr. 1, S. 61-85.
URL: *http://www.atala.org/sites/default/files/TAL-2009-50-1-03-Bourdaillet.pdf*.

Bourdaillet/Ganascia 2006: Julien Bourdaillet / Jean-Gabriel Ganascia, MEDITE: A Unilingual Textual Aligner, in: Tapio Salakoski u. a. (Hg.), Advances in Natural Language Processing. 5th International Conference on NLP, FinTAL 2006. Turku, Finland, August 23-25, 2006. Proceedings, Berlin/Heidelberg 2006 (LNCS 4139), S. 458-469.

Bourdaillet/Ganascia 2007: Julien Bourdaillet / Jean-Gabriel Ganascia, Alignment of noisy unstructured data, in: Proceedings of the IJCAI Workshop on Analytics for Noisy Unstructured Text Data (AND 2007) of the 20th International Joint Conference on Artificial Intelligence (IJCAI), 2007, S. 139-146.
URL: *http://research.ihost.com/and2007/cd/Proceedings_files/p139.pdf*.

Bourne/Ford 1961: Charles P. Bourne / Donald F. Ford, A Study of Methods for Systematically Abbreviating English Words and Names, in: JACM 8 (1961), S. 538-552.
DOI: *10.1145/321088.321094*.

Broich 1985: Ulrich Broich, Formen der Markierung von Intertextualität, in: Broich/Pfister (Hg.) 1985, S. 31-47.

Broich 2000: Ulrich Broich, Art. „Intertextualität“, in: RLW 2 (2000), S. 175-179.

Broich/Pfister (Hg.) 1985: Ulrich Broich / Manfred Pfister (Hg.), Intertextualität. Formen, Funktionen, anglistische Fallstudien, Tübingen 1985.

Büchler 2013: Marco Büchler, Informationstechnische Aspekte des *Historical Text Re-use*, Diss. Leipzig 2013.
URL: *http://www.informatik.uni-leipzig.de/~graebe/Texte/Buechler-13-Diss.pdf*.

Büchler u. a. 2010: Marco Büchler / Annette Geßner / Thomas Eckart / Gerhard Heyer, Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts, in: Journal of the Chicago Colloquium on Digital Humanities and Computer Science 1, Nr. 2 (2010).
URL: *https://letterpress.uchicago.edu/index.php/jdhcs/article/view/60/71*.

Büchler u. a. 2014: Marco Büchler / Greta Franzini / Emily Franzini / Maria Moritz, Scaling Historical Text Re-use, in: Jimmy Lin u. a. (Hg.), 2014 IEEE International Conference on Big Data. 27 – 30 October 2014 Washington DC, USA. Proceedings, 2014.
DOI: *10.1109/BigData.2014.7004449*.

Bumke 1996: Joachim Bumke, Der unfeste Text. Überlegungen zur Überlieferungsgeschichte und Textkritik der höfischen Epik im 13. Jahrhundert, in: Jan-Dirk Müller (Hg.), ›Aufführung‹ und ›Schrift‹ in Mittelalter und Früher Neuzeit, Stuttgart/Weimar 1996 (Germanistische Symposien. Berichtsbände 17), S. 118-129.

Bussmann 1990: Hadumod Bußmann, Lexikon der Sprachwissenschaft, 2., völlig neu bearb. Aufl., Stuttgart 1990 (Kröners Taschenausgabe 452).

Byloff 1907: Fritz Byloff, Die Land- und peinliche Gerichtsordnung Erzherzog Karls II. für Steiermark vom 24. Dezember 1574. Ihre Geschichte und ihre Quellen, Graz 1907 (Forschungen zur Verfassungs- und Verwaltungsgeschichte der Steiermark 6, 3).

CICiv.(Krüger/Mommsen): Corpus iuris civilis, 3 Bde.: 1.) Institutiones. Recognovit Paulus Krueger. Digesta. Recognovit Theodorus Mommsen. Retractavit Paulus Krueger, 22. Aufl., Dublin/Zürich 1973; 2.) Codex Iustinianus. Recognovit et retractavit Paulus Krueger, 14. Aufl., Dublin/Zürich 1967; 3.) Novellae. Recognovit Rudolfus Schoell. Opus Schoellii morte interceptum absolvit Guilelmus Kroll, 10. Aufl., Dublin/Zürich 1972.

Clough u. a. 2002: Paul Clough / Robert Gaizauskas / Scott S. L. Piao / Yorick Wilks, METER: MEasuring TExt Reuse, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, S. 152-159.
URL: *http://www.aclweb.org/anthology/P02-1020.pdf*.

Coffee u. a. 2013: Neil Coffee / Jean-Pierre Koenig / Shakthi Poornima / Christopher W. Forstall / Roelant Ossewaarde / Sarah L. Jacobson, The Tesserae Project: intertextual analysis of Latin poetry,

in: LLC 28 (2013), Nr. 2, S. 221-228.
DOI: *10.1093/llc/fqs033*.

Coing 1973: Helmut Coing, Erster Teil. Wissenschaft. Erster Abschnitt. Die juristische Fakultät und ihr Lehrprogramm, in: Ders. (Hg.), Handbuch der Quellen und Literatur der neueren europäischen Privatrechtsgeschichte. Bd. 1: Mittelalter (1100-1500). Die gelehrten Rechte und die Gesetzgebung, München 1973, S. 39-128.

Coing 1977: Helmut Coing, Erster Teil. Wissenschaft. Erster Abschnitt. Die juristische Fakultät und ihr Lehrprogramm, in: Ders. (Hg.), Handbuch der Quellen und Literatur der neueren europäischen Privatrechtsgeschichte. Bd. 2: Neuere Zeit (1500-1800). Das Zeitalter des gemeinen Rechts. Teilbd. 1: Wissenschaft, München 1977, S. 3-102.

Crochemore/Rytter 1994/2009: M. Crochemore / W. Rytter, Text Algorithms (2009 erstellte, anders paginierte Fassung des gleichnamigen Buchs von 1994).
URL: *http://monge.univ-mlv.fr/~mac/REC/text-algorithms.pdf*.

Dekker/Middell 2011: Ronald H. Dekker / Gregor Middell, Computer-Supported Collation with CollateX. Managing Textual Variance in an Environment with Varying Requirements, in: Bente Maegaard (Hg.), Supporting Digital Humanities. Copenhagen 17 – 18 November 2011. Conference Proceedings.
URL: *http://crdo.up.univ-aix.fr/SLDRdata/doc/show/copenhagen/SDH-2011/submissions/sdh2011_submission_54.pdf* (am 28. 12. 2017 nicht aufrufbar).

Deutsch 2004: Andreas Deutsch, Der Klagspiegel und sein Autor Conrad Heyden. Ein Rechtsbuch des 15. Jahrhunderts als Wegbereiter der Rezeption, Köln 2004 (Diss. Heidelberg 2003).

Deutsch 2008: Andreas Deutsch, Die „Rethorica und Formulare teütsch" des Pforzheimer Stadtschreibers Alexander Hugen – ein juristischer Bestseller des 16. Jahrhunderts, in: Neue Beiträge zur Pforzheimer Stadtgeschichte 2 (2008), S. 31-75.

Deutsch 2009: Andreas Deutsch, Art. „Gobler, Justin (1503-1567)", in: HRG^2 Bd. 2, Lfg. 10 (2009), Sp. 438-440.

Deutsch 2010: Andreas Deutsch, *Klagspiegel* und *Laienspiegel* – Sebastian Brants Beitrag zum Ruhm zweier Rechtsbücher, in: Klaus Bergdolt u. a. (Hg.), Sebastian Brant und die Kommunikationskultur um 1500, Wiesbaden 2010 (Wolfenbütteler Abhandlungen zur Renaissanceforschung 26), S. 75-98.

Deutsch 2012: Andreas Deutsch, Art. „Klagspiegel", in: HRG^2 Bd. 2, Lfg. 16 (2012), Sp. 1864-1869.

Deutsch (Hg.) 2011: Andreas Deutsch (Hg.), Ulrich Tenglers Laienspiegel. Ein Rechtsbuch zwischen Humanismus und Hexenwahn, Heidelberg 2011 (Akademiekonferenzen 11).

Deutsch (Hg.) 2013: Andreas Deutsch (Hg.), Historische Rechtssprache des Deutschen, Heidelberg 2013 (Akademiekonferenzen 15 / Schriftenreihe des Deutschen Rechtswörterbuchs).

DH2012: Jan Christoph Meister (Hg.), Digital Humanities 2012. Conference Abstracts. University of Hamburg, July 16-22, 2012, Hamburg 2012.
URL: *http://hup.sub.uni-hamburg.de/HamburgUP/DH2012_Book_of_Abstracts*.

Dilcher 2014: Gerhard Dilcher, Art. „Libri Feudorum", in: HRG^2 Bd. 3, Lfg. 20 (2014), Sp. 970-976.

Duden 9^7: Duden. Richtiges und gutes Deutsch. Das Wörterbuch der sprachlichen Zweifelsfälle. Hg. und überarb. von der Dudenredaktion unter Mitwirkung von Peter Eisenberg und Jan Georg Schneider, 7., vollständig überarb. Aufl., Mannheim/Zürich 2011 (Duden 9).

DWB: Jakob Grimm / Wilhelm Grimm, Deutsches Wörterbuch, Leipzig 1 (1854) – 16 (1954).
URL: *http://woerterbuchnetz.de/DWB/*.

Eisenhardt 1995: Ulrich Eisenhardt, Deutsche Rechtsgeschichte, 2. Aufl., München 1995 (Grundrisse des Rechts).

Eisenstein 1997: Elizabeth L. Eisenstein, Die Druckerpresse. Kulturrevolutionen im frühen modernen Europa, Wien / New York 1997 (Übersetzung von: The Printing Revolution in Early Modern Europe, 1983) (Ästhetik und Naturwissenschaften : Medienkultur).

Elm (Hg.) 2000: Kaspar Elm (Hg.), Literarische Formen des Mittelalters. Florilegien. Kompilationen. Kollektionen, Wiesbaden 2000 (Wolfenbütteler Mittelalter-Studien 15).

Erler 1956: Adalbert Erler, Thomas Murner als Jurist, Frankfurt am Main 1956 (Frankfurter wissenschaftliche Beiträge. Rechts- und wirtschaftsgeschichtliche Reihe 13).

Erler 1981: Adalbert Erler, Art. „Murner, Thomas“, in: HRG Bd. 3, Lfg. 20 (1981), Sp. 793-795.

Ernst-Gerlach 2013: Andrea Ernst-Gerlach, Retrievalmethoden für historische Korpora mit nicht standardisierten Schreibweisen, Diss. Duisburg-Essen 2013.
URN: *urn:nbn:de:hbz:464-20130723-144610-7*.

Fernandes/Freitas 2013/14: slaMEM: efficient retrieval of maximal exact matches using a sampled LCP array / Francisco Fernandes / Ana T. Freitas, in: Bioinformatics Volume 30, 4 (2014), S. 464-471.
DOI: *10.1093/bioinformatics/btt706* („Published: December 11, 2013“).

Fitch 1969: Walter M. Fitch, Locating Gaps in Amino Acid Sequences to Optimize the Homology Between Two Proteins, in: Biochemical Genetics 3 (1969), S. 99-108.

Forstall u. a. 2014: Christopher Forstall / Neil Coffee / Thomas Buck / Katherine Roache / Sarah Jacobson, Modeling the scholars: Detecting intertextuality through enhanced word-level n-gram matching, in: Digital Scholarship in the Humanities. Advance access, Version 2.12.2014.
DOI: *10.1093/llc/fqu014*.

Friedl 2004: Jeffrey E. F. Friedl, Reguläre Ausdrücke, Übers. v. Andreas Karrer. 2. Aufl., 1., korrig. Nachdruck, Beijing u. a. 2004.

FWB: Frühneuhochdeutsches Wörterbuch, hg. v. Robert R. Anderson (Bd. 1) / Ulrich Goebel / Anja Lobenstein-Reichmann (ab Bd. 5) / Oskar Reichmann / Institut für Deutsche Sprache (Bd. 4 u. 7), Berlin [u. a.] 1989ff.

Ganascia 2007: Jean-Gabriel Ganascia, EDITE – MEDITE: un passage des versions aux variantes?, in: David Trotter (Hg.), Actes du XXIV^{e} Congrès International de Linguistique et de Philologie Romanes. Aberystwyth 2004, Bd. 1, Tübingen 2007, S. 357ff.
URL: *http://www-poleia.lip6.fr/~ganascia/Medite_Project?action=AttachFile&do=get&target=CILPR+2005*.

Ganascia 2011: Jean-Gabriel Ganascia, MEDITE – A Unilingual Text Aligner for Humanities. Applications to Textual Genetics and to the Edition of Text Variants. [Beitrag zur Tagung „Supporting Digital Humanities“, Kopenhagen 17./18.11. 2011].
URL: *http://www-poleia.lip6.fr/~ganascia/Medite_Project?action=AttachFile&do=view&target=SDH2011.pdf*.

Ganascia/Glaudes/Del Lungo 2014: Jean-Gabriel Ganascia / Peirre Glaudes / Andrea Del Lungo, Automatic detection of reuses and citations in literary texts, in: LLC 29 (2014), Nr. 3, S. 412-421.
DOI: *10.1093/llc/fqu020*.

Geldner 1978: Ferdinand Geldner, Inkunabelkunde. Eine Einführung in die Welt des frühesten Buchdrucks, Wiesbaden 1978 (Elemente des Buch- und Bibliothekswesens 5).

Genette 1993: Gérard Genette, Palimpseste. Die Literatur auf zweiter Stufe. Aus dem Französischen von Wolfram Bayer und Dieter Hornig, Frankfurt am Main 1993 (Originalausgabe: Palimpsestes. La littérature au second degré, Paris 1982. Übersetzt nach der ergänzten 2. Auflage).

Gerber 1846: C[arl] F[riedrich] Gerber, Das wissenschaftliche Princip des gemeinen deutschen Privatrechts. Eine germanistische Abhandlung, Jena 1846.
URL: *https://books.google.de/books?id=VmpDAAAAcAAJ*.

Gessner 2010: Annette Geßner, Das automatische Auffinden der indirekten Überlieferung des Platonischen Timaios und die Bedeutung des Tools „CitationGraph“ für die Forschung, in: Schubert/Heyer (Hg.) 2010, S. 26-41.

Gibbs/McIntyre 1970: Adrian J. Gibbs / George A. McIntyre, The Diagram, a Method for Comparing Sequences : Its Use with Amino Acid and Nucleotide Sequences, in: European Journal of Biochemistry 16 (1970), Nr. 1, S. 1-11.
DOI: *10.1111/j.1432-1033.1970.tb01046.x*.

Gierl 2001: Martin Gierl, Kompilation und die Produktion von Wissen im 18. Jahrhundert, in: Zedelmaier/Mulsow (Hg.) 2001, S. 63-94.

Giesecke 1991: Michael Giesecke, Der Buchdruck in der frühen Neuzeit. Eine historische Studie über die Durchsetzung neuer Informations- und Kommunikationstechnologien, Frankfurt am Main 1991.

Gieseke 1957: Ludwig Gieseke, Die geschichtliche Entwicklung des deutschen Urheberrechts, Göttingen 1957 (Göttinger rechtswissenschaftliche Studien 22).

Gieseke 1995: Ludwig Gieseke, Vom Privileg zum Urheberrecht. Die Entwicklung des Urheberrechts in Deutschland bis 1845, Göttingen 1995.

Goltschnigg/Grollegg-Edler/Gruber (Hg.) 2013: Dietmar Goltschnigg / Charlotte Grollegg-Edler / Patrizia Gruber (Hg.), Plagiat, Fälschung, Urheberrecht im interdisziplinären Blickfeld, Berlin 2013.

Gottron 2010: Thomas Gottron, External Plagiarism Detection Based on Standard IR Technology and Fast Recognition of Common Subsequences. Lab Report for PAN at CLEF 2010.
URL: *http://www.uni-weimar.de/medien/webis/events/pan-10/pan10-papers-final/pan10-plagiarism-detection/gottron10-notebook.pdf*.

Grafton 1997: Anthony Grafton, The footnote. A Curious History, Rev. ed., Cambridge/Mass. 1997 1997.

Greenberg 2003: Ronald I. Greenberg, Bounds on the Number of Longest Common Subsequences, revidiert (v2), 6.8.2003.
URL: *http://arxiv.org/abs/cs/0301030v2*.

Grubmüller 1997: Klaus Grubmüller, Art. „Florilegium“, in: RLW Bd. 1 (1997), S. 605-607.

Gusfield 1997: Dan Gusfield, Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology, Cambridge u. a. 1997 (12. Nachdruck 2009).

Haferland 2011: Harald Haferland, Wer oder was trägt einen Namen? Zur Anonymität in der Vormoderne und in der deutschen Literatur des Mittelalters, in: Stephan Pabst (Hg.), Anonymität und Autorschaft. Zur Literatur- und Rechtsgeschichte der Namenlosigkeit, Berlin/Boston 2011 (Studien und Texte zur Sozialgeschichte der Literatur 126), S. 49-72.

Harel/Feldman 2006: David Harel / Yishai Feldman, Algorithmik. Die Kunst des Rechnens, Übers. v. Micaela Krieger-Hauwede, Berlin/Heidelberg 2006.

Hartweg/Wegera 2005: Frédéric Hartweg / Klaus-Peter Wegera, Frühneuhochdeutsch. Eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit, 2., neu bearb. Aufl. Tübingen 2005.

Hasebrink 2000: Burkhard Hasebrink, Zersetzung? Eine Neubewertung der Eckhartkompilation in *Spamers Mosaiktraktaten*, in: Elm (Hg.) 2000, S. 73-90.

Haustein/Stackmann 1998: Jens Haustein / Karl Stackmann, Sangspruchstrophen in Tönen Frauenlobs. Aus den Vorarbeiten für die Ergänzung der Göttinger Ausgabe, in: Joachim Heinzle / L. Peter Johnson / Gisela Vollmann-Profe (Hg.), Neue Wege der Mittelalter-Philologie. Landshuter Kolloquium 1996, Berlin 1998 (Wolfram-Studien 15), S. 74-103.

Heckel 1978: Paul Heckel, A technique for isolating differences between files, in: Communications

of the ACM 21 (1978), Nr. 4, S. 264-268.
DOI: *10.1145/359460.359467*.

Horton/Olsen/Roe 2010: Russell Horton / Mark Olsen / Glenn Roe, Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections, in: Digital studies / Le champ numérique 2, Nr. 1 (2010).
URL: *https://www.digitalstudies.org/articles/10.16995/dscn.258/*.

HRG: Adalbert Erler u. a. (Hg.), Handwörterbuch zur deutschen Rechtsgeschichte : HRG, 5 Bde., Berlin 1971-1998.

HRG^2: Albrecht Cordes u. a. (Hg.), Handwörterbuch zur deutschen Rechtsgeschichte : HRG, 2., völlig überarb. und erw. Aufl., Berlin 2004 (1. Lieferung) ff.

Hunt/McIlroy 1976/2012: J. W. Hunt / M. D. McIlroy, An Algorithm for Differential File Comparison, 1976 (Bell Laboratories Computing Science Technical Report 41), OCR-Version 2012.
URL: *http://www.cs.dartmouth.edu/~doug/diff.pdf*.

Hunt/Szymanski 1977: James W. Hunt / Thomas G. Szymanski, A fast algorithm for computing longest common subsequences, in: Communications of the ACM 20 (1977), Nr. 5, S. 350-353.
DOI: *10.1145/359581.359603*.

ISTC: Incunabula Short Title Catalogue.
URL: *http://www.bl.uk/catalogues/istc/index.html*.

JACM: Journal of the ACM (JACM).

Jäger 1828: Carl Jäger, Geschichte der Stadt Heilbronn und ihres ehemaligen Gebietes. Ein Beitrag zur Geschichte des schwäbischen Städtewesens. Nach handschriftlichen Quellen bearbeitet, Bd. 2, Heilbronn 1828.
URN: *urn:nbn:de:bvb:12-bsb10019705-0*.

Johanek 1992: Peter Johanek, Art. ‚Schwabenspiegel', in: VL^2 8 (1992), Sp. 896-907.

Jurish 2011: Bryan Jurish, Finite-State Canonicalization Techniques for Historical German, Diss. Berlin 2011.
URN: *urn:nbn:de:kobv:517-opus-55789*.

Kaeding (Hg.) 1898: Friedrich Wilhelm Kaeding (Hg.), Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuß der deutschen Stenographiesysteme, Steglitz 1898.
URL: *http://archive.org/details/hufigkeitswrter00kaedgoog*.

Kaib 1987: Hildegard Kaib, Zu den juristischen Schriften Thomas Murners, in: Thomas Murner. Elsässischer Theologe und Humanist (1475 – 1537). Eine Ausstellung der Badischen Landesbibliothek, Karlsruhe und der Bibliothèque Nationale et Univ., Strasbourg. Ausstellungskatalog, Karlsruhe 1987, S. 93-112.

Kallweit 2000: Hilmar Kallweit, Art. „Kompilation", in: RLW 2 (2000), S. 317-321.

Kantorowicz 1933/1987: Hermann U[lrich] Kantorowicz, Die Allegationen im späten Mittelalter, in: Eltjo J. H. Schrage (Hg.), Das römische Recht im Mittelalter, Darmstadt 1987 (Wege der Forschung 635), S. 71-88 (ursprünglich erschienen in: Archiv für Urkundenforschung 13 (1933), S. 15-29).

Kärkkäinen/Ukkonen 1996: Sparse Suffix Trees / Juha Kärkkäinen / Esko Ukkonen, in: Jin-Yi Cai / Chak Kuen Wong (Hg.), Computing and Combinatorics. COCOON 1996, Heidelberg 1996 (LNCS 1090), S. 219-230.
DOI: *10.1007/3-540-61332-3_155*.

Kaufmann 1986: E[kkehard] Kaufmann, Art. „Rechtsquellen", in: HRG Bd. 4, Lfg. 26 (1986), Sp. 335-337.

Kewes (Hg.) 2003: Paulina Kewes (Hg.), Plagiarism in Early Modern England, Basingstoke/New York 2003.

Khan u. a. 2009: Zia Khan / Joshua S. Bloom / Leonid Kruglyak / Mona Singh, A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays, in: Bioinformatics 25 (2009), S. 1609-1616.
DOI: *10.1093/bioinformatics/btp275*.

Khiste/Ilie 2014/15: Nilesh Khiste / Lucian Ilie, E-MEM: efficient computation of maximal exact matches for very large genomes, in: Bioinformatics 31, 4 (2015), S. 509-514.
DOI: *10.1093/bioinformatics/btu687* („Published: 17 October 2014“).

Kiessling 1930: Gerhard Kießling, Die Anfänge des Titelblattes in der Blütezeit des deutschen Holzschnitts (1470-1530), Leipzig [1930] (Monographien des Buchgewerbes 14).

Kisch 1962: Guido Kisch, Die Anfänge der Juristischen Fakultät der Universität Basel. 1459-1529, Basel 1962 (Studien zur Geschichte der Wissenschaften in Basel 15).

Klein/Fix (Hg.) 1997: Josef Klein / Ulla Fix (Hg.), Textbeziehungen. Linguistische und literaturwissenschaftliche Beiträge zur Intertextualität, Tübingen 1997.

Knuth 1998: Donald E. Knuth, The art of computer programming. Bd. 3: Sorting and Searching, 2. Aufl. Boston u. a. 1998.

Kochendörfer 1974: Günter Kochendörfer, Maschinelle Rekonstruktion mehrfach überlieferter Texte, in: Protokoll des 3. Kolloquiums über die Anwendung der Elektronischen Datenverarbeitung in den Geisteswissenschaften an der Universität Tübingen vom 18. Mai 1974.
URL: *http://www.tustep.uni-tuebingen.de/prot/prot3.html#kochendoerfer*.

Kramer 2010: Christian Kramer, Deutschsprachige Rechtsliteratur der Frühen Neuzeit – eine bio-bibliographische Datenbank, in: Andreas Deutsch (Hg.), Das Deutsche Rechtswörterbuch – Perspektiven, Heidelberg 2010 (Akademiekonferenzen 8), S. 235-243.

Kristeva 1967: Julia Kristeva, Bakhtine, le mot, le dialogue et le roman, in: Critique. Revue générale des publications françaises et étrangères 23 (1967), S. 438-465.

Kroeschell 1998: Karl Kroeschell, Von der Gewohnheit zum Recht. Der Sachsenspiegel im späten Mittelalter, in: Boockmann u. a. (Hg.) 1998, S. 68-92.

Krohn 2008: Jan Peter Krohn, Die oberösterreichische Landtafel von 1616/1629 und die Rezeption des römisch-kanonischen Rechts – eine erste Bilanz, in: Mitteilungen des Oberösterreichischen Landesarchivs 21 (2008), S. 425-616 (Diss. Konstanz 2007).

Kühnel 1976: Jürgen Kühnel, Der "offene Text". Beitrag zur Ueberlieferungsgeschichte volkssprachiger Texte des Mittelalters. (Kurzfassung), in: Leonard Forster / Hans-Gert Roloff (Hg.), Akten des V. Germanisten-Kongresses Cambridge 1975. Heft 2, Frankfurt/M. 1976 (Jahrbuch für Internationale Germanistik. A: Kongreßberichte / 2), S. 311-321.

Kümmel 2011: Alexander Kümmel, Text Re-use-Extraktion auf Basis eines Sequence-Alignment Problems. Masterarbeit Weimar, 2011.
URL: *http://www.uni-weimar.de/medien/webis/teaching/theses/kuemmel_2011.pdf*.

Kunze/Lemnitzer 2007: Claudia Kunze / Lothar Lemnitzer, Computerlexikographie. Eine Einführung, Tübingen 2007.
URL: *http://www.digicontent.narr.de.ubproxy.ub.uni-heidelberg.de/16315/* (aufrufbar nach Login von *http://katalog.ub.uni-heidelberg.de/titel/66869788* aus).

Kurtz 2016: Stefan Kurtz, The Vmatch large scale sequence analysis software. A Manual, January 11, 2016.
URL: *http://www.vmatch.de/virtman.pdf*.

Kurtz u. a. 2004: Stefan Kurtz / Adam Phillippy / Arthur L. Delcher / Michael Smoot / Martin Shumway / Corina Antonescu / Steven L. Salzberg, Versatile and open software for comparing large genomes, in: Genome Biology 5:R12 (2004).
URL: *http://genomebiology.com/2004/5/2/R12*.

Landès/Hénaut/Risler 1993: Claudine Landès / Alain Hénaut / Jean-Loup Risler, Dot-plot comparisons by multivariate analysis (DOCMA): a tool for classifying protein sequences, in: Computer applications in the biosciences 9 (1993), Nr. 2, S. 191-196.
DOI: *10.1093/bioinformatics/9.2.191*.

Laufs 1984: Adolf Laufs, Rechtsentwicklungen in Deutschland, 3., erg. Aufl., Berlin/New York 1984 (de Gruyter Lehrbuch).

Lee 2007: John Lee, A Computational Model of Text Reuse in Ancient Literary Texts, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. June 23 – 30, 2007, Prague, Czech Republic, S. 472-479.
URL: *http://aclweb.org/anthology/P/P07/P07-1060.pdf*.

Leipold/Ritter/Solms 2014: Aletta Leipold / Jörg Ritter / Hans-Joachim Solms, Neue Wege zu Textzeugenvergleich und Edition am Beispiel der Wundarznei des Heinrich von Pfalzpaint, in: Jahrbuch für Germanistische Sprachgeschichte 5 (2014), S. 335-358.

Lexer: Matthias Lexer, Mittelhochdeutsches Handwörterbuch. Zugleich als Supplement und alphabetischer Index zum Mittelhochdeutschen Wörterbuche von Benecke-Müller-Zarncke. 3 Bde., Leipzig 1872-1878.
URL: *http://woerterbuchnetz.de/Lexer/*.

Lipp 2007: Martin Lipp, Art. „Einkindschaft“, in: HRG^2 Bd. 1, Lfg. 6 (2007), Sp. 1296-1298.

LLC: Literary and Linguistic Computing.
URL: *http://llc.oxfordjournals.org/*.

LNCS: Lecture Notes in Computer Science.

Lück 1997: Heiner Lück, Die kursächsische Gerichtsverfassung 1423 – 1550, Köln [u. a.] 1997 (Forschungen zur deutschen Rechtsgeschichte 17).

Lyon/Barrett/Malcolm 2004: Caroline Lyon / Ruth Barrett / James Malcolm, A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector, in: Proceedings of the first Plagiarism Conference (2004).
URL: *http://uhra.herts.ac.uk/handle/2299/2114*.

Manber/Myers 1993: Udi Manber / Gene Myers, Suffix Arrays: A New Method for On-Line String Searches, in: SIAM Journal on Computing 22 (1993), S. 935-948.
DOI: *10.1137/0222058*.

Marquordt 1938: Gerhard Marquordt, Vier rheinische Prozeßordnungen aus dem 16. Jahrhundert. Ein Beitrag zum Prozeßrecht der Rezeptionszeit, Bonn 1938 (Rheinisches Archiv 33; Diss. Göttingen 1938).
URL: *http://repertorium.at/sl/marquordt_1938.html*.

McGill 2012: Scott McGill, Plagiarism in Latin Literature, Cambridge u. a. 2012.

Medek u. a. 2015: André né Gießler Medek / Marcus Pöckelmann / Thomas Bremer / Hans-Joachim Solms / Paul Molitor / Jörg Ritter, Differenzanalyse komplexer Textvarianten. Diskussion und Werkzeuge, in: Datenbank Spektrum 2015 (online).
DOI: *10.1007/s13222-014-0173-y*.

Meierhofer 2010: Christian Meierhofer, Alles neu unter der Sonne. Das Sammelschrifttum der Frühen Neuzeit und die Entstehung der Nachricht, Würzburg 2010 (Epistemata. Würzburger wissenschaftliche Schriften. Reihe Literaturwissenschaft 702; Diss. Würzburg 2008/09).

MGH.DD H II: Societas aperiendis fontibus rerum germanicarum medii aevi = Gesellschaft für ältere deutsche Geschichtskunde (Hg.), Diplomatum regum et imperatorum Germaniae Tomus III. Heinrici II. et Arduini diplomata = Die Urkunden der deutschen Könige und Kaiser. Bd. 3. Die Urkunden Heinrichs II. und Arduins, Hannover 1900-1903 (Monumenta Germaniae Historica inde ab anno Christi quingentesimo usque ad annum millesimum et quingentesimum).

Micol u. a. 2010: Daniel Micol / Óscar Ferrández / Fernando Llopis / Rafael Muñoz, A Textual-Based Similarity Approach for Efficient and Scalable External Plagiarism Analysis. Lab Report for PAN at CLEF 2010.
URL: *http://www.uni-weimar.de/medien/webis/events/pan-10/pan10-papers-final/pan10-plagiarism-detection/micol10-notebook.pdf*.

Minnis 1979: Alastair J. Minnis, Late-Medieval Discussions of *compilatio* and the Rôle of the *compilator*, in: Beiträge zur Geschichte der deutschen Sprache und Literatur 101 (1979), S. 385-421.

Moretti 2000: Franco Moretti, Conjectures on World Literature, in: New Left Review, Second Series 1 (2000), S. 54-68.
URL: *http://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature*.

Moser 1929: Virgil Moser, Frühneuhochdeutsche Grammatik. I. Band: Lautlehre. 1. Hälfte: Orthographie, Betonung, Stammsilbenvokale, Heidelberg 1929.

Motloch 1907: [Theodor] Motloch, Art. „Länder: A. Landesordnungen (geschichtlich) und Landhandfesten: I. Österreichische Ländergruppe“, in: Ernst Mischler / Josef Ulbrich (Hg.), Österreichisches Staatswörterbuch. Handbuch des gesamten österreichischen öffentlichen Rechts, 2., wesentl. umgearb. Aufl., Bd. 3, Wien 1907, S. 331-357.
URL: *http://repertorium.at/sl/motloch_1907.html*.

Munk Olsen 1982: Birger Munk Olsen, Les florilèges d'auteurs classiques, in: Les genres littéraires dans les sources théologiques et philosophieques médiévales. Définition, critique et exploitation. Actes du Colloque international de Louvain-la-Neuve 25-27 mai 1981, Louvain-la-Neuve 1982 (Université catholique de Louvain, Publications de l'institut d'études médiévales, 2^{e} série 5), S. 151-164.

Muther 1860: Theodor Muther, Die Gewissensvertretung im gemeinen Deutschen Recht, mit Berücksichtigung von Particulargesetzgebungen, besonders der Sächsischen und Preußischen, Erlangen 1860.
URL: *https://books.google.de/books?id=7wkQAAAAYAAJ*.

Muther 1876: Theodor Muther, Zur Geschichte der Rechtswissenschaft und der Universitäten in Deutschland. Gesammelte Aufsätze, Jena 1876.
URL: *https://books.google.de/books?id=B8wFAAAAQAAJ*.

Nawab/Stevenson/Clough 2012: Rao Muhammad Adeel Nawab / Mark Stevenson / Paul Clough, Detecting Text Reuse with Modified and Weighted N-grams, in: First Joint Conference on Lexical and Computational Semantics (*SEM), Montréal, Canada, June 7-8, 2012, S. 54-58.
URL: *http://www.aclweb.org/anthology/S12-1008*.

Needleman/Wunsch 1970: Saul B. Needleman / Christian D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, in: Journal of Molecular Biology 48 (1970), S. 443-453.
DOI: *10.1016/0022-2836(70)90057-4*.

Neumann 2001: Florian Neumann, Jeremias Drexels *Aurifodina* und die *Ars excerpendi* bei den Jesuiten, in: Zedelmaier/Mulsow (Hg.) 2001, S. 51-61.

Nitsche 2013: Gunter Nitsche, Plagiat und Urheberrecht, in: Goltschnigg/Grollegg-Edler/Gruber (Hg.) 2013, S. 77-88.

NSRA: Johann Jacob Schmauß / Heinrich Christian von Senckenberg (Hg.), Neue und vollständigere Sammlung der Reichs-Abschiede, Welche von den Zeiten Kayser Conrads des II. bis jetzo, auf den Teutschen Reichs-Tägen abgefasset worden, 4 Teile, Frankfurt a. M. 1747.
URL: *http://hdl.handle.net/hdl:11858/00-001M-0000-002A-02F6-7*.

Nünning (Hg.) 2008: Ansgar Nünning (Hg.), Metzler Lexikon Literatur- und Kulturtheorie. Ansätze – Personen – Grundbegriffe, 4., akt. u. erw. Aufl., Stuttgart/Weimar 2008.

OAHeilbronn^2 I: Beschreibung des Oberamts Heilbronn. Hrsg. von dem K. Statistischen Landesamt, [2. Aufl.], Teil 1, Stuttgart 1901.

Ohlebusch/Gog/Kügel 2010: Enno Ohlebusch / Simon Gog / Adrian Kügel, Computing Matching Statistics and Maximal Exact Matches on Compressed Full-Text Indexes, in: Edgar Chávez / Stefano Lonardi (Hg.): String Processing and Information Retrieval. 17th International Symposium, SPIRE 2010. Los Cabos, Mexico, October 11-13, 2010. Proceedings, Berlin/Heidelberg 2010 (LNCS 6393), S. 347-358.

Orwant/Hietaniemi/Macdonald 2000: Jon Orwant / Jarkko Hietaniemi / John Macdonald, Algorithmen mit Perl, Dt. Übers. v. Andreas Karrer, Beijing u. a. 2000.

Otte 1964: Albert Otte, Die Mainzer Hofgerichtsordnung von 1516, 1521 und die Gesetzgebung auf dem Gebiet der Zivilgerichtsbarkeit im 16. Jahrhundert. Geschichte, Quellen und Wirkung des Gesetzes für die Zentraljustizbehörde eines geistlichen Kurfürstentums, Diss. Mainz 1964.

Ottmann/Widmayer 1996: Thomas Ottmann / Peter Widmayer, Algorithmen und Datenstrukturen, 3., überarb. Aufl., Heidelberg/Berlin/Oxford 1996.

Paulsen 1919: Friedrich Paulsen, Geschichte des gelehrten Unterrichts auf den deutschen Schulen und Universitäten vom Ausgang des Mittelalters bis zur Gegenwart. Mit besonderer Rücksicht auf den klassischen Unterricht. Bd. 1, 3., erw. Aufl., hg. und in einem Anhang fortgesetzt von Rudolf Lehmann, Leipzig 1919.

Pauser 2004: Josef Pauser, Landesfürstliche Gesetzgebung (Policey-, Malefiz- und Landesordnungen), in: Josef Pauser / Martin Scheutz / Thomas Winkelbauer (Hg.), Quellenkunde der Habsburgermonarchie (16. – 18. Jahrhundert). Ein exemplarisches Handbuch, Wien 2004 (Mitteilungen des Instituts für Österreichische Geschichtsforschung. Ergänzungs-Band 44), S. 216-256.

Peters (Hg.) 2001: Ursula Peters (Hg.), Text und Kultur. Mittelalterliche Literatur 1150-1450, Stuttgart/Weimar 2001 (Germanistische Studien. Berichtsbände 23).

Pfister 1985: Manfred Pfister, Konzepte der Intertextualität, in: Broich/Pfister (Hg.) 1985, S. 1-30.

Plachta 2006: Bodo Plachta, Editionswissenschaft. Eine Einführung in Methode und Praxis der Edition neuerer Texte, 2., erg. u. akt. Aufl. Stuttgart 2006.

Plett 1990: Heinrich F. Plett, Oralität und Literalität in Rhetorik und Poetik der englischen Renaissance, in: Wolfgang Raible (Hg.), Erscheinungsformen kultureller Prozesse. Jahrbuch 1988 des Sonderforschungsbereichs "Übergänge und Spannungsfelder zwischen Mündlichkeit und Schriftlichkeit", Tübingen 1990 (ScriptOralia 13), S. 167-195.

Plett 1994: Heinrich F. Plett, Renaissance-Poetik. Zwischen Imitation und Innovation, in: Heinrich F. Plett (Hg.), Renaissance-Poetik = Renaissance Poetics, Berlin/New York 1994, S. 1-20.

Posner 2007: Richard A. Posner, The Little Book of Plagiarism, New York 2007.

Potthast 2011: Martin Potthast, Technologies for Reusing Text from the Web, Diss. Bauhaus-Universität Weimar 2011.
URL: *http://www.uni-weimar.de/medien/webis/publications/papers/potthast_2011b.pdf.*

Potthast u. a. 2009: Martin Potthast / Benno Stein / Andreas Eiselt / Alberto Barrón-Cedeño / Paolo Rosso, Overview of the 1st International Competition on Plagiarism Detection, in: Stein u. a. (Hg.) 2009, S. 1-9.

Potthast u. a. 2010: Martin Potthast / Alberto Barrón-Cedeño / Andreas Eiselt / Benno Stein / Paolo Rosso, Overview of the 2nd International Competition on Plagiarism Detection, in: Martin Braschler / Donna Harman (Hg.), Notebook Papers of CLEF 10 Labs and Workshops, September 2010.
URL: *http://www.uni-weimar.de/medien/webis/publications/papers/stein_2010t.pdf.*

Potthast u. a. 2011: Martin Potthast / Andreas Eiselt / Alberto Barrón-Cedeño / Benno Stein / Paolo Rosso, Overview of the 3rd International Competition on Plagiarism Detection, in: Vivien Petras /

Paul Clough (Hg.), Notebook Papers of CLEF 11 Labs and Workshops, September 2011.
URL: *http://www.uni-weimar.de/medien/webis/publications/papers/stein_2011t.pdf*.

Potthast u. a. 2012: Martin Potthast / Tim Gollub / Matthias Hagen / Jan Graßegger / Johannes Kiesel / Maximilian Michel / Arnd Oberländer / Martin Tippmann / Alberto Barrón-Cedeño / Parth Gupta / Paolo Rosso / Benno Stein, Overview of the 4th International Competition on Plagiarism Detection, in: Pamela Forner / Jussi Karlgren / Christa Womser-Hacker (Hg.), CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy, 2012.
URL: *https://www.uni-weimar.de/medien/webis/publications/papers/stein_2012t.pdf*.

Potthast u. a. 2013: Martin Potthast / Matthias Hagen / Tim Gollub / Martin Tippmann / Johannes Kiesel / Paolo Rosso / Efstathios Stamatatos / Benno Stein, Overview of the 5th International Competition on Plagiarism Detection, in: Pamela Forner / Roberto Navigli / Dan Tufis (Hg.), CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain, 2013.
URL: *http://www.uni-weimar.de/medien/webis/publications/papers/stein_2013h.pdf*.

Potthast u. a. 2014: Martin Potthast / Matthias Hagen / Anna Beyer / Matthias Busse / Martin Tippmann / Paolo Rosso / Benno Stein, Overview of the 6th International Competition on Plagiarism Detection, in: In Linda Cappellato / Nicola Ferro / Martin Halvey / Wessel Kraaij (Hg.), Working Notes Papers of the CLEF 2014 Evaluation Labs, CEUR Workshop Proceedings, September 2014.
URL: *http://www.uni-weimar.de/medien/webis/publications/papers/stein_2014k.pdf*.

Rautenberg 2004: Ursula Rautenberg, Das Titelblatt. Die Entstehung eines typographischen Dispositivs im frühen Buchdruck, 2004 (Alles Buch. Studien der Erlanger Buchwissenschaft 10).
URL: *http://www.alles-buch.uni-erlangen.de/forschung/publikationen/Rautenberg.pdf*.

Reeg 1977: Gottfried Reeg, Maschineller Vergleich von Textzeugen zur Vorbereitung einer kritischen Edition, in: Protokoll des 12. Kolloquiums über die Anwendung der Elektronischen Datenverarbeitung in den Geisteswissenschaften an der Universität Tübingen vom 2. Juli 1977.
URL: *http://www.tustep.uni-tuebingen.de/prot/prot12.html#reeg*.

Reichmann 1989: Oskar Reichmann, Lexikographische Einleitung, in: FWB Bd. 1. Einführung. *a – äpfelkern*. Bearb. v. Oskar Reichmann, Berlin/New York 1989, S. 10-164.

Reichmann/Wegera 1993: Oskar Reichmann / Klaus-Peter Wegera, Schreibung und Lautung, in: Oskar Reichmann / Klaus-Peter Wegera (Hg.), Frühneuhochdeutsche Grammatik. Von Robert Peter Ebert, Oskar Reichmann, Hans-Joachim Solms und Klaus-Peter Wegera, Tübingen 1993 (Sammlung kurzer Grammatiken germanischer Dialekte. A. Hauptreihe, Nr. 12), S. 13-163.

Reyscher (Hg.) 1831: Sammlung der württembergischen Gerichts-Gesetze. 1. Enthaltend die erste Reihe der Gerichts-Gesetze vom Jahre 1298 bis zum Jahre 1608. Von Chr[istian] H[einrich] Riecke, Stuttgart 1831 (Reyscher, A[ugust] L[udwig] (Hg.), Vollständige, historisch und kritisch bearbeitete Sammlung der württembergischen Gesetze 4).
URN: *urn:nbn:de:bvb:12-bsb10552283-6*.

RLW: Reallexikon der deutschen Literaturwissenschaft. Neubearbeitung des Reallexikons der deutschen Literaturgeschichte, hg. v. Klaus Weimar / Harald Fricke / Georg Braungart / Jan-Dirk Müller u. a., 3 Bde., Berlin u. a. 1997-2003 (Nachdruck 2007).
URL: *http://www.degruyter.com/doi/book/10.1515/9783110914672*.

Robinson 1989: P. M. W. Robinson, The Collation and Textual Criticism of Icelandic Manuscripts (1): Collation, in: LLC 4 (1989), Nr. 2, S. 99-105.

Robinson 2009: Peter Robinson, Towards a Scholarly Editing System for the Next Decades, in: Gérard Huet / Amba Kulkarni / Peter Scharf (Hg.), Sanskrit Computational Linguistics. First and Second International Symposia. Rocquencourt, France, October 29-31, 2007. Providence, RI, USA, May 15-17, 2008. Revised Selected and Invited Papers, Berlin/Heidelberg 2009 (Lecture Notes in Artifical Intelligence 5402), S. 346-357.

Roe u. a. 2012: Glenn H. Roe / The ARTFL Project, Intertextuality and Influence in the Age of Enlightenment: Sequence Alignment Applications for Humanities Research, in: DH2012, S. 345-347.

Roelcke 1998: Thorsten Roelcke, Die Periodisierung der deutschen Sprachgeschichte, in: Werner Besch u. a. (Hg.), Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung, 2., vollständig neu bearb. u. erweit. Aufl., 1. Teilbd., Berlin/New York 1998 (Handbücher zur Sprach- und Kommunikationswissenschaft Bd. 2.1, 2. Aufl.), S. 798-815.

Rosso u. a. 2016: Paolo Rosso / Francisco Rangel / Martin Potthast / Efstathios Stamatatos / Michael Tschuggnall / Benno Stein, Overview of PAN'16. New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation, in: Norbert Fuhr u. a. (Hg.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016. Proceedings, Berlin/Heidelberg/New York 2016 (LNCS 9822), S. 332-350.
URL: *http://www.uni-weimar.de/medien/webis/publications/papers/stein_2016i.pdf*.

RTA.JR: Historische Kommission bei der Bayerischen Akademie der Wissenschaften (Hg.), Deutsche Reichstagsakten, Jüngere Reihe. Deutsche Reichstagsakten unter Kaiser Karl V., Bd. 1 (1893) ff.

Russell 1918: Robert C. Russell, Index. United States Patent Office, Patent Nr. 1.261.167, 2.4.1918 (eingereicht am 25.10.1917).
URL: *http://worldwide.espacenet.com/publicationDetails/originalDocument?CC=US&NR=1261167A&KC=A&FT=D&ND=&date=19180402&DB=&locale=en_EP*.

Russell 1922: Robert C. Russell, Index. United States Patent Office, Patent Nr. 1.435.663, 14.11.1922 (eingereicht am 28.11.1921).
URL: *http://worldwide.espacenet.com/publicationDetails/originalDocument?CC=US&NR=1435663A&KC=A&FT=D&ND=&date=19221114&DB=&locale=en_EP*.

Sahle 2013: Patrick Sahle, Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe; Teil 2: Befunde, Theorie und Methodik; Teil 3: Textbegriffe und Recodierung, Norderstedt 2013 (Diss. Köln 2009; Schriften des Instituts für Dokumentologie und Editorik 7-9).
URL: *http://kups.ub.uni-koeln.de/id/eprint/5351* (Teil 1)
URL: *http://kups.ub.uni-koeln.de/id/eprint/5352* (Teil 2)
URL: *http://kups.ub.uni-koeln.de/id/eprint/5353* (Teil 3).

Sankoff 1972: D. Sankoff, Matching Sequences Under Deletion/Insertion Constraints, in: Proceedings of the National Academy of Sciences of the United States of America 69 (1972), Nr. 1, S. 4-6.
URL: *http://www.pnas.org/content/69/1/4.full.pdf*.

Schälkle/Ott 2017: Kuno Schälkle / Wilhelm Ott, TUSTEP. Tübinger System von Textverarbeitungs-Programmen. Version 2017. Handbuch und Referenz, Tübingen 2017.
URL: *http://www.tustep.uni-tuebingen.de/pdf/handbuch.pdf*.

Schiewer 2005: Hans-Jochen Schiewer, Fassung, Bearbeitung, Version und Edition, in: Martin J. Schubert (Hg.), Deutsche Texte des Mittelalters zwischen Handschriftennähe und Rekonstruktion. Berliner Fachtagung 1.-3. April 2004, Tübingen 2005 (Beihefte zu *editio* 23), S. 35-50.

Schmid 1831: Karl Ernst Schmid, Carolus Ernestus Schmid Theol. Et Iuris Utriusque Doctor Ordinis Iure Consultorum In Universitate Literaria Ienensi H. T. Decanus Solemnia Inauguralia Iuris Utriusque Candidati Aug. Henr. Aemil. Danz D. III Mensis Aug. MDCCCXXXI Publice Habenda Indicit. Disseritur de ordinationis provincialis Hennebergicae origine fontibus et auctoritate, Jena 1831.
URL: *https://books.google.de/books?id=uv1aAAAAQAAJ*.

Schmidt 1965: Eberhard Schmidt, Einführung in die Geschichte der deutschen Strafrechtspflege, 3., völlig durchgearb. u. veränd. Aufl., Göttingen 1965.

Schmidt 2000: Paul Gerhard Schmidt, "Neue Klassiker" in lateinischen Dichterflorilegien des Spätmittelalters, in: Elm (Hg.) 2000, S. 19-23.

Schnell 1998: Rüdiger Schnell, ‚Autor' und ‚Werk' im deutschen Mittelalter. Forschungskritik und Forschungsperspektiven, in: Joachim Heinzle / L. Peter Johnson / Gisela Vollmann-Profe (Hg.), Neue

Wege der Mittelalter-Philologie. Landshuter Kolloquium 1996, Berlin 1998 (Wolfram-Studien 15), S. 12-73.

Schnell 2001: Rüdiger Schnell, Vom Sänger zum Autor. Konsequenzen der Schriftlichkeit des deutschen Minnesangs, in: Peters (Hg.) 2001, S. 96-149.

Schrenk/Weckbach 1993: Christhard Schrenk / Hubert Weckbach, Die Vergangenheit für die Zukunft bewahren. Das Stadtarchiv Heilbronn. Geschichte – Aufgaben – Bestände, 1993 (Veröffentlichungen des Archivs der Stadt Heilbronn 33; Online-Publikationen des Stadtarchivs Heilbronn 17).
URL: *https://stadtarchiv.heilbronn.de/fileadmin/daten/stadtarchiv/online-publikationen/17-schrenk-weckbach_vergangenheit-fuer-die-zukunft.pdf*.

Schubert 2010: Charlotte Schubert, Zitationsprofile, Suchstrategien und Forschungsrichtungen, in: Schubert/Heyer (Hg.) 2010, S. 42-55.

Schubert 2015: Charlotte Schubert, Editorial: Close Reading und Distant Reading. Methoden der Altertumswissenschaften in der Gegenwart, in: Digital Classics Online 1 (2015) Nr. 1 S. 1-6.
DOI: *10.11588/dco.2015.1.20483*.

Schubert/Heyer (Hg.) 2010: Ch[arlotte] Schubert / G[erhard] Heyer (Hg.), Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung I, Leipzig 2010 (Working Papers CONTESTED ORDER 1).
DOI: *10.11588/ea.2010.0*.

Schubert/Klank (Hg.) 2012: Ch[arlotte] Schubert / M[arkus] Klank (Hg.), Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung III, Leipzig 2012 (Working Papers CONTESTED ORDER 7).
URL: *http://www.uni-leipzig.de/~order/content/images/stories/wp_7_schubert.pdf* (am 28. 12. 2017 nicht mehr aufrufbar).
DOI: *10.11588/ea.2012.2* (wohl inhaltlich übereinstimmend mit der verwendeten Fassung, aber mit anderer Reihenbezeichnung, anderen Angaben zu den Herausgebern und anderer Gestaltung von Titel, Impressum und Quellenangabe am Ende).

Schulz/Leese/Held 2008: Jan Schulz / Florian Leese / Christoph Held, Inroduction[!] to dot-plots, 2008.
URL: *http://www.code10.info/index.php?option=com_content&view=article&id=64:inroduction-to-dot-plots&catid=52:cat_coding_algorithms_dot-plots&Itemid=76*.

Schumann 2007: Eva Schumann, Beiträge studierter Juristen und anderer Rechtsexperten zur Rezeption des gelehrten Rechts. (Vortrag in der Plenarsitzung am 12. Oktober 2007), in: Jahrbuch der Akademie der Wissenschaften zu Göttingen 2007, Berlin 2008, S. 443-461.
URL: *http://hdl.handle.net/11858/00-001S-0000-0007-375C-6*.

Schumann 2013: Eva Schumann, Rechts- und Sprachtransfer am Beispiel der volkssprachigen Praktikerliteratur, in: Deutsch (Hg.) 2013, S. 123-174.

Schwartz 1898: Johann Christoph Schwartz, Vierhundert Jahre deutscher Zivilprozeß-Gesetzgebung. Darstellungen und Studien zur deutschen Rechtsgeschichte, Berlin 1898 (ND Aalen 1986).
URL: *http://digi.ub.uni-heidelberg.de/diglit/schwartz1898* (Digitalisat des Originaldrucks).

Schwerin 1950: Claudius Freiherr von Schwerin, Grundzüge der deutschen Rechtsgeschichte, 4. Aufl., besorgt von Hans Thieme, Berlin 1950.

Shannon 1948: Claude E. Shannon, A Mathematical Theory of Communication, in: The Bell System Technical Journal 27 (1948), S. 379-423 u. 623-656.
DOI: *10.1002/j.1538-7305.1948.tb01338.x*
DOI: *10.1002/j.1538-7305.1948.tb00917.x*.

Simon 1898: Jacob Simon, Die Henneberger Landesordnung vom 1. Januar 1539, in: Schriften des Vereins für Sachsen-Meiningische Geschichte und Landeskunde 31 (1898), S. 29-45.
URN: *urn:nbn:de:urmel-abb26693-58a1-4c9e-a25c-af489f7b3b421-00004542-301*.

Smith/Cordell/Dillon 2013: David A. Smith / Ryan Cordell / Elizabeth Maddock Dillon, Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers, in: Xiaohua Hu u. a. (Hg.), 2013 IEEE International Conference on Big Data : Silicon Valley, California, USA, 6 – 9 October 2013, Piscataway 2013, S. 86-94.
URL: *http://www.ccs.neu.edu/home/dasmith/infect-bighum-2013.pdf*.

Smith/Cordell/Mullen 2015: David Smith / Ryan Cordell / Abby Mullen, Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers [Preprint-Version].
URL: *http://viraltexts.org/2015/05/22/computational-methods-for-uncovering-reprinted-texts-in-antebellum-newspapers/*.

Snowsill 2012: Tristan Snowsill, Data mining in text streams using suffix trees, Diss. Bristol 2012.
URL: *http://www.tristansnowsill.co.uk/phd/thesis.pdf* (am 28. 12. 2017 nicht mehr aufrufbar).

Spangenberg 1825: E. P. Spangenberg, Gobler und seine Uebersetzung der Carolina, in: Neues Archiv des Criminalrechts 7 (1825), S. 429-458.
URL: *https://books.google.de/books?id=HrVCAAAAcAAJ*.

Stackmann 1998: Karl Stackmann, Wiederverwerteter Frauenlob. Nichts Ungewöhnliches – und was man daraus lernen kann, in: Joachim Heinzle / L. Peter Johnson / Gisela Vollmann-Profe (Hg.), Neue Wege der Mittelalter-Philologie. Landshuter Kolloquium 1996, Berlin 1998 (Wolfram-Studien 15), S. 104-113.

Stamatatos 2009: Efstathios Stamatatos, Intrinsic Plagiarism Detection Using Character *n*-gram Profiles, in: Stein u. a. (Hg.) 2009, S. 38-46.
URL: *http://ceur-ws.org/Vol-502/paper8.pdf*.

Stamatatos u. a. 2015: Efstathios Stamatatos / Martin Potthast / Francisco Rangel / Paolo Rosso / Benno Stein, Overview of the PAN/CLEF 2015 Evaluation Lab, in: Josiane Mothe u. a. (Hg.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 6th International Conference of the CLEF Association, CLEF'15 Toulouse, France, September 8-11, 2015. Proceedings, Berlin/Heidelberg/New York 2015 (LNCS 9283), S. 518-538.
URL: *http://www.uni-weimar.de/medien/webis/publications/papers/stein_2015m.pdf*.

Stein u. a. (Hg.) 2009: Benno Stein / Paolo Rosso / Efstathios Stamatatos / Moshe Koppel / Eneko Agirre (Hg.), 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse. 25th Annual Conference of the Spanish Society for Natural Language Processing, SEPLN 2009.
URL: *http://ceur-ws.org/Vol-502/pan09-proceedings.pdf*.

Stintzing 1867: Roderich Stintzing, Geschichte der populären Literatur des römisch-kanonischen Rechts in Deutschland am Ende des fünfzehnten und im Anfang des sechszehnten Jahrhunderts, Leipzig 1867.
URN: *urn:nbn:de:bvb:12-bsb10801178-7*.

Stintzing 1880: R[oderich] Stintzing, Geschichte der deutschen Rechtswissenschaft, Abt. 1, München 1880 (Geschichte der Wissenschaften in Deutschland : Neuere Zeit 18, 1).
URN: *urn:nbn:de:bvb:12-bsb11169799-3*.

Stobbe 1864: Otto Stobbe, Geschichte der deutschen Rechtsquellen, Bd. 2, Braunschweig 1864 (Geschichte des deutschen Rechts 1, 2).

Stock 2007: Wolfgang G. Stock, Information Retrieval. Informationen suchen und finden, München/Wien 2007.

Sturm 1996: Gudrun Sturm, Art. „Zasius, Udalricus (Ulrich Zäsi)", in: HRG Bd. 5, Lfg. 39 (1996), Sp. 1612-1614.

Swinson/Reyna 2013: Todd Swinson / Carlos Reyna, Authorship Attribution Using Stopword Graphs, 2013.
URL: *http://www.cs.utexas.edu/~swinson/files/Swinson_AuthorshipAttribution.pdf* (am 28. 12. 2017 nicht mehr aufrufbar).

Tegtmeyer 1997: Henning Tegtmeyer, Der Begriff der Intertextualität und seine Fassungen – Eine Kritik der Intertextualitätskonzepte Julia Kristevas und Susanne Holthuis', in: Klein/Fix (Hg.) 1997, S. 49-81.

Theisohn 2009: Philipp Theisohn, Plagiat. Eine unoriginelle Literaturgeschichte, Stuttgart 2009 (Kröners Taschenausgabe 351).

Theuerkauf 1968: Gerhard Theuerkauf, Lex, speculum, compendium iuris. Rechtsaufzeichnung und Rechtsbewußtsein in Norddeutschland vom 8. bis zum 16. Jahrhundert, Köln [u. a.] 1968 (Forschungen zur deutschen Rechtsgeschichte 6).

Titot 1865: Heinrich Titot, Beschreibung des Oberamts Heilbronn, Stuttgart 1865.
URL: *http://de.wikisource.org/wiki/Beschreibung_des_Oberamts_Heilbronn*.

Troje 1970: H[ans] E[rich] Troje, Art. „Gobler, Justin", in: HRG Bd. 1, Lfg. 7 (1970), Sp. 1726-1729.

UBHeilbronn: Urkundenbuch der Stadt Heilbronn. 4 Bde. Bd. 1 bearb. von Eugen Knupfer; Bd. 2-4 bearb. von Moriz von Rauch, Stuttgart 1904-1922 (Württembergische Geschichtsquellen 5/15/19/20).

Unger 1889: Albert Unger, Handbuch des im Herzogthume Sachsen Meinigen geltenden partikularen Privatrechts, Band 1, Hildburghausen 1889.
URL: *http://dlib-pr.mpier.mpg.de/m/kleioc/0010/exec/books/%22211483%22*.

UrhG: Gesetz über Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz), Ausfertigungsdatum: 09.09.1965. Zuletzt geändert durch Art. 1 G v. 20.12.2016 I 3037.
URL: *http://www.gesetze-im-internet.de/urhg/BJNR012730965.html*.

VD16: Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts (VD 16) [Datenbank-Version].
URL: *http://www.vd16.de/*.

Vihinen 1988: Mauno Vihinen, An algorithm for simultaneous comparison of several sequences, in: Computer applications in the biosciences 4 (1988), Nr. 1, S. 89-92.
DOI: *10.1093/bioinformatics/4.1.89*.

Vingron/Argos 1991: Martin Vingron / Patrick Argos, Motif recognition and alignment for many sequences by comparison of dot-matrices, in: Journal of Molecular Biology 218 (1991), S. 33-43.

VL^2: Kurt Ruh u. a. (Hg.), Die deutsche Literatur des Mittelalters. Verfasserlexikon, 2., völlig neu bearb. Aufl., 11 Bde., Berlin 1978-2004.

VL.DH: Franz Josef Worstbrock (Hg.), Deutscher Humanismus 1480 – 1520. Verfasserlexikon, 3 Bde., Berlin 2005-2015.

Vogl 2007: Heidemarie Vogl, Der „Spiegel der Seele". Eine spätmittelalterliche mystisch-theologische Kompilation, Stuttgart 2007 (Meister-Eckhart-Jahrbuch. Beihefte 2).

Vyverman u. a. 2013a: Michaël Vyverman / Bernard De Baets / Veerle Fack / Peter Dawyndt, essaMEM: finding maximal exact matches using enhanced sparse suffix arrays, in: Bioinformatics 29 (2013), S. 802-804.
DOI: *10.1093/bioinformatics/btt042*.

Vyverman u. a. 2013b: Michaël Vyverman / Bernard De Baets / Veerle Fack / Peter Dawyndt, Supplementary Material for „essaMEM: Finding Maximal Exact Matches Using Enhanced Sparse Suffix Arrays" (aufrufbar von der Onlineversion von Vyverman u. a. 2013a aus über den Link „Supplementary data").

Wächter 1836: [Carl Georg von] Wächter, Ueber die Deutsche criminalistische Literatur des XVI^{ten} Jahrhunderts an sich und in ihrem Verhältnisse zur Carolina., in: Archiv des Criminalrechts, N. F., 1836, S. 115-153.
URN: *urn:nbn:de:bvb:12-bsb10393652-9*.

Wächter 1839: Carl Georg [von] Wächter, Handbuch des im Königreiche Württemberg geltenden Privatrechts. Bd. 1: Geschichte, Quellen und Literatur des württembergischen Privatrechts, Abt. 1,

Stuttgart 1839.
URN: *urn:nbn:de:bvb:12-bsb10552893-7*.

Wagner/Fischer 1974: Robert A. Wagner / Michael J. Fischer, The String-to-String Correction Problem, in: JACM 21 (1974), Nr. 1, S. 168-173.
DOI: *10.1145/321796.321811*.

Wall/Christiansen/Orwant 2001/2003: Larry Wall / Tom Christiansen / Jon Orwant, Programmieren mit Perl. Dt. Übers. v. Peter Klicman, 2. Aufl. (dt. Ausgabe der 3. Aufl.), Beijing u. a. 2001, 2., korr. Nachdruck 2003.

Weiner 1973: Peter Weiner, Linear pattern matching algorithms, in: 14th Annual Symposium on Switching & Automata Theory. October 15-17, 1973, S. 1-11.
DOI: *10.1109/SWAT.1973.13*.

Wilpert 1979: Gero von Wilpert, Sachwörterbuch der Literatur, 6., verb. u. erw. Aufl., Stuttgart 1979 (Kröners Taschenausgabe 231).

Wilz 2005: Martin Wilz, Aspekte der Kodierung phonetischer Ähnlichkeiten in deutschen Eigennamen, Magisterarbeit Köln [2005].
URL: *http://phonetik.phil-fak.uni-koeln.de/fileadmin/Phonetik_Files/Allgemeine_Dateien/Martin_Wilz.pdf*.

Worstbrock 2013: F[ranz] J[osef] Worstbrock, Art. „Murner, Thomas“, in: VL.DH 2 (2013), Sp. 299-368.

Zedelmaier 1992: Helmut Zedelmaier, Bibliotheca universalis und Bibliotheca selecta. Das Problem der Ordnung des gelehrten Wissens in der frühen Neuzeit, Köln / Weimar / Wien / Böhlau 1992 (Beihefte zum Archiv für Kulturgeschichte 33; Diss. LMU München 1989).

Zedelmaier 2000: Helmut Zedelmaier, De ratione excerpendi: Daniel Georg Morhof und das Exzerpieren, in: Françoise Waquet (Hg.), Mapping the World of Learning: The *Polyhistor* of Daniel Georg Morhof, Wiesbaden 2000 (Wolfenbütteler Forschungen 91), S. 75-92.

Zedelmaier 2001: Helmut Zedelmaier, Lesetechniken. Die Praktiken der Lektüre in der Frühen Neuzeit, in: Zedelmaier/Mulsow (Hg.) 2001, S. 11-30.

Zedelmaier/Mulsow (Hg.) 2001: Helmut Zedelmaier / Martin Mulsow (Hg.), Die Praktiken der Gelehrsamkeit in der Frühen Neuzeit, Tübingen 2001 (Frühe Neuzeit 64).

Zoepfl (Hg.) 1883: Heinrich Zoepfl (Hg.), Die Peinliche Gerichtsordnung Kaiser Karl's V. nebst der Bamberger und der Brandenburger Halsgerichtsordnung. Sämmtlich nach den ältesten Drucken und mit den Projecten der peinlichen Gerichtsordnung Kaiser Karl's V. von den Jahren 1521 und 1529. Beide zum erstenmale vollständig nach Handschriften herausgegeben, 3., synopt. Ausg., Leipzig 1883.