Informationstechnische Aspekte des *Historical Text Re-use*

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR rerum naturalium
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt

von Dipl.-Inf. Marco Büchler
geboren am 24. Juli 1978 in Eilenburg

Leipzig, den 3. Februar 2013

An experiment is a question which science poses to Nature,

and a measurement is the recording of Nature's answer.

Max Planck, (1858-1947)

Allgemeine Zusammenfassung

Das Wiederverwenden von Informationen und das Zitieren von Texten begleitet unser tägliches Leben. Das können Nachrichten von Agenturen, wie der *dpa*, Plagiarismusvorwürfe in der Politik, die Replikation von digitalen Inhalten bzw. Artikeln auf mehreren Zeitungsseiten oder *geflügelte Wörter* als Teil der Allgemeinsprache sein. All diese Szenarien umfassen den Forschungsbereich des gesprochenen und geschriebenen *Text Re-use*, welcher als Oberbegriff für alle Methoden verstanden werden muss, ein derartiges Vervielfachen von Text systematisch aufzudecken.

Die Schwierigkeit dieses Forschungsbereiches liegt in seiner Komplexität und Vielfältigkeit. Insbesondere im historischen Bezugsrahmen werden Inhalte oftmals nicht identisch, sondern je nach Intention des Autors leicht oder auch stärker verändert wiedergegeben. Ein gutes Beispiel hierfür ist in der jüngeren Geschichte die Kubakrise 1962. Auch wenn in Texten das gleiche Ereignis beschrieben wird, so ist der Einfluss auf die Wiedergabe von Informationen stark vom Standpunkt des Autors abhängig, ob die Ansichten der USA oder der damaligen UdSSR unterstützt werden, so dass bis auf Entitäten, wie Länder, Zeitangaben und Personen, keine nennenswerten Überlappungen festzustellen sind. Auch wenn an dieser Stelle lediglich ein Beispiel genannt ist, so folgen in der Arbeit zahlreiche weitere Belege gleicher Art. In der Summe zeigen sie eine sehr starke Diversität in der Benutzung von Inhalten durch *Text Re-use* auf.

Insbesondere im historischen Kontext der letzten Jahrhunderte oder gar Jahrtausende ist diese Vielfalt umso größer, da neben autorspezifischen Aspekten auch Sprachevolution, verschiedene Dialekte, semantischer Wandel von Konzepten aber auch verschiedene Schreibweisen, bedingt durch weniger reglementierte oder einer in den Anfängen der Vertextlichung nicht einheitlichen Rechtschreibung, den Schreibprozess begleiten. Entsprechende Anpassungen in späteren Editionen, wie im 19. Jahrhundert in *Webster's Revision* der englischsprachigen *King James Version* der Bibel aus dem 16. Jahrhundert, welche unter anderem sämtliches archaische Englisch entfernt, oder auch nur Modifizierungen von Zitaten, sind jedoch keineswegs zufällig. Diese Veränderungen und unterschiedlichen Darstellungen gleicher Ereignisse sind fundamentaler Bestandteil in den Geisteswissenschaften. Während ein Historiker durch möglichst viele schriftliche Belegstellen eines Ereignisses eine vollständigere Rekonstruktion von Geschehnissen machen kann, helfen Schreibvarianten und Ersetzungen den Philologen, historische Texte, wie die von Platon, zu rekonstruieren, da keines der Werke heutzutage als noch erhalten angesehen werden kann. Neben solchen Aspekten der Textkritik helfen insbesondere systematische Veränderungen eines Autors bei der Wiedergabe von Informationen, auch den verändernden Autor anhand seines Kopierverhaltens wiederzuerkennen. Entgegen modernen Anwendungen des *Text Re-use*, wie dem Plagiarismus, kann der *Historical Text Re-use* daher als ein nützliches Instrument verstanden werden, welcher nicht nur Evidenzen von Transferwegen, sondern vielmehr auch einen fundamentalen Teil des sprachlich-kulturellen Erbes der Menschheit darstellt. Aus dieser Vielfalt des *Historical Text Re-use* ergeben sich für die Informatik im Rahmen der *eHumanities* vielschichtige Herausforderungen, die Gegenstand dieser Arbeit sind. Die *eHumanities* werden hierbei als das Zusammenwachsen der beiden Disziplinen *Informatik* und den Geisteswissenschaften verstanden.

Erstens, wie kann das Ergebnis einer *Text Re-use Analysis* evaluiert werden? Das Kernproblem, welches sich durch die *Vielfalt* des *Text Re-use* ergibt, ist die fehlende Möglichkeit einer Evaluierung von Ergebnissen. Einerseits liegen keine historischen Daten für *Text Re-use* digital vor. Andererseits müsste eine entsprechende *Evaluierungsbasis* auch die Vielfalt abdecken, was nahezu unmöglich ist.

Zweitens, bis zu welchem Grad der Veränderung kann ein *Text Re-use* automatisch noch erkannt werden? Je mehr ein Original verändert bzw. paraphrasiert wird, desto schwieriger wird es, insbesondere zur eingangs dargestellten Diversität alle möglichen Ziele eines *Text Re-use* innerhalb einer *Digital Library* aufzudecken.

Drittens, da entsprechende Veränderungen eines wiederverwendenden Autors meistens nicht zufällig sind, stellt sich die Frage: Wie können diese Veränderungen systematisch bestimmt und extrahiert werden?

Diese drei Fragestellungen können als hinreichende Motivation verstanden werden, den *Historical Text Re-use* in Shannon's *Noisy Channel Theorem* einzubetten. In diesem Kontext kann ein Original- bzw. zitierter Autor als *Source* und ein wiederverwendender Autor als *Target* verstanden werden. Der *Noisy Channel* stellt vielmehr ein unbekanntes Modell von Modifikationen, den äußeren Einflüssen, dar. Im Rahmen dieser Arbeit wird, gemessen an den drei genannten Herausforderungen, die Einbettung für zwei systematische Betrachtungsweisen verwendet.

Auf der einen Seite wird der *Noisy Channel* als Modell eingesetzt, um historisch paradigmatische Relationen zu bestimmen. Das ist insbesondere unter Berücksichtigung der großen Zeitfenster von historischen Texten jedes Genres von Interesse, da sich semantische Beziehungen von Konzepten im Laufe der Zeit verändern. So werden nur beispielhaft *Bier* und *Wein* in der Antike synonym benutzt, während sie heutzutage in einer kohyponymen Relation zueinander stehen. Diese Extraktionsmethode wird daher auch als *Noisy Channel Mining* eingeführt.

Auf der anderen Seite wird das *Noisy Channel Model* dazu eingesetzt, um ein zufälliges und rein künstliches Störsignal zum *Noisy Channel* hinzuzufügen, so dass eine *Randomised Digital Library* entsteht. Im Rahmen der Arbeit werden insgesamt fünf Randomisierungstechniken, die künstlichen Störsignale, vorgestellt, welche unterschiedliche Schwierigkeitsgrade einer rein quantitativen Evaluierung mit sich bringen. Für diese quantitative Evaluierung, die *Noisy Channel Evaluation*, wird der neuartige *Score* der *Mining Ability* eingeführt. Die *Mining Ability* setzt hierbei das Ergebnis einer *Text Re-use Analysis* auf einer *Digital Library* mit dem Resultat einer durch ein künstliches Störsignal veränderten *Randomised Digital Library* ins Verhältnis, wodurch nicht nur Parameter optimiert werden können, sondern auch verschiedene Sprachmodelle vollautomatisch und bzgl. des Ergebnisses ganzheitlich evaluiert werden. Als *Digital Libraries* werden für diese Form der Analysen sowohl die altgriechische *Perseus Digital Library* als auch sieben verschiedene englischsprachige Bibelversionen ausgewählt. Mit jeder der sieben Bibelversionen wird ein Interesse oder auch der kuturell-sprachliche Hintergrund des erstellenden Editors reflektiert. Während die *King James Version* aus dem 16. Jahrhundert als die älteste englischsprachige Bibel mit zahlreichen archaischen Wortformen verstanden werden darf, haben andere Bibeln die Absicht, den Inhalt mit möglichst einfachen Worten wiederzugeben oder sich wenn auch mit englischen Wörtern so zumindest an der hebräischen Satzsyntax zu orientieren, wodurch die eingangs dargestellte *Diversity* hinreichend gegeben ist.

Um den *Text Re-use* mit all seiner Vielfalt automatisch bestimmen zu können, wird die *Black Box* des *Text Mining* auf eine *7-Level-Architektur*, bestehend aus *Segmentation*, *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring* und *Postprocessing*, aufgeteilt. Dadurch können nicht nur den Fachwissenschaften Einsichten in Zwischenergebnisse gegeben werden, sondern vielmehr sind aus Sicht des *Software Engineering* über eine Million Kombinationen von Algorithmen der einzelnen Level möglich. Diese Architektur wurde explizit als technische Antwort der fachwissenschaftlich relevanten *Diversity* historischer Texte im Rahmen dieser Arbeit entwickelt.

Wissenschaftliche Zusammenfassung

Gegenstand der Arbeit

Was ist *Text Re-use*? *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. Während der Fokus dieser Arbeit auf dem Erstellen eines *Hypertextes* durch eine *Text Re-use Analysis* liegt, sind die *PageRanking*-Technik oder auch bibliometrische Analysen weiterführende Anwendungen. Im Kontext derartiger Einsatzmöglichkeiten kann auf historischen Dokumenten, die dieser Arbeit zugrunde liegen, durch eine automatische Analyse eine noch nie zuvor erstellte Breite von Zitierabhängigkeiten erstellt werden, welche heutzutage Aufschluss darüber geben, was in früheren Zeiten als wichtig erachtet worden ist, auch wenn es in der Gegenwart für Sprachen, wie dem Altgriechischen oder dem Latein, keine Muttersprachler mehr gibt.

Stand der Forschung

In der Plagiarismuserkennung, einer modernen Anwendung von *Text Re-use*, werden meist einfache *Ngramm*-Ansätze eingesetzt. Diese Form einer Abtastung eines Textes bietet in erster Linie den Vorteil, dass die benötigte Rechenzeit relativ klein bleibt. Ferner genügt dieser Ansatz, um ein einfaches *Copy & Paste* zu erkennen.

Außerhalb des Plagiarismus stellt sich der Forschungsstand so dar, dass nahezu beliebig Daten und Algorithmen kombiniert werden. Die Ergebnisse geben datenspezifische Charakteristika wieder und sind somit oft nicht auf andere Daten reproduzierbar. Der Forschungsstand reflektiert somit mehr Insellösungen als eine ganzheitliche Sicht auf das Thema.

Ganzheitliche Sicht auf *Text Re-use*

In Kapitel 2 wird die derzeit vollständigste Systematisierung des *Text Re-use* vorgenommen. Dies umfasst zwei wesentliche Aspekte:

- Es werden insgesamt 45 verschiedene Typisierungen von Textstellen, nachfolgend auch *Meme* im Sinne eines Gedanken oder Gedankensplitters genannt, eingeführt, welche in der Regel wiederverwendet werden. Entsprechende typisierte *Meme* reichen nur beispielhaft von *Sprichwort*, über *Schlachtruf* und *Vers* bis hin zur *Legende*.

- Es wird eine Systematik zu verschiedenen *Re-use Styles* definiert, welche beschreibt, wie ein entsprechendes *Meme* wiederverwendet wird. Das kann zum Beispiel ein wortwörtliches Zitat aber auch eine Paraphrase oder Allusion sein.

Das Ziel dieser ganzheitlichen Sicht besteht darin, grundlegende Eigenschaften der *Meme* sowie der *Re-use Styles* zu definieren. Während ein *Meme*, wie z. B. eine *Redewendung*, eher kurz und syntaktisch fest verwendet wird, ist es beim größeren *Meme Legende* üblich, dieses mündlich und damit wesentlich freier wiederzugeben.

Während die Typisierung der verschiedenen *Meme* die Frage aufwirft, warum bestimmte Textinhalte wiederverwendet werden, gibt die zweite Systematik des *Re-use Styles* Aufschluss darüber, wie jeder persönlich andere Inhalte wiedergibt.

Sowohl die Typisierung der verschiedenen *Meme* mit ihren unterschiedlichen Charakteristika als auch die Systematik der *Re-use Styles* reflektieren eine *Data Diversity*, welche eine Herausforderung sowohl für die *Text Re-use Analysis* aber auch für deren *Evaluation* aus ganzheitlicher Sicht bedeutet, da es keinen *Gold Standard* gibt, welcher sowohl alle möglichen *Meme* als auch die verschiedenen *Re-use Styles* adäquat repräsentiert.

Forschungsfragen

Aus ganzheitlicher Sicht ergeben sich somit für diese Arbeit die folgenden Forschungsfragen:

- Im Kontext der verschiedenen *Re-use Styles* muss die Frage danach gestellt werden, bis zu welchem Grad der Veränderung ein *Text Re-use* automatisch noch erkannt werden kann.
- Wie kann eine *Text Re-use Analysis* so gestaltet werden, dass sie auch für unterschiedliche *Meme* mit verschiedenen Charakteristika gleich gut funktioniert?
- Wie können Veränderungen eines wiederverwendenden Autors systematisch bestimmt und extrahiert werden?
- Wie kann das Ergebnis einer *Text Re-use Analysis* in einer *Digital Library*[1] in Anbetracht der *Data Diversity* ganzheitlich evaluiert werden?

Untersuchte *Digital Libraries*

Im Gegensatz zum *Text Re-use* auf modernen Texten, wie in der Plagiarismusforschung, wird eine Analyse auf historischen Dokumenten dadurch erschwert, dass keine vereinheitlichte Schreibweise angenommen werden kann, da neben autorspezifischen Aspekten auch Sprachevolution, Dialekte, semantischer Wandel von Konzepten aber auch verschiedene Varianten, verursacht durch eine weniger reglementierte und in den Anfängen der Verschriftlichung von Texten nicht existenten Rechtschreibung, den Schreibprozess begleiten.

Diese Arbeit vereint somit im Rahmen des *Historical Text Re-use* sowohl die *Data Diversity*, bedingt durch verschiedene *Meme* und *Re-use Styles*, mit der großen sprachlichen Vielfalt von historischen Dokumenten. Als Datenbasis liegen dieser Arbeit drei verschiedene *Digital Libraries* zugrunde:

- *Perseus Digital Library*: Diese Datenbasis wird eingesetzt, um sehr kurzen *Text Re-use* auf altgriechischen Werken mit einem hohen Maß an sprachlicher Vielfalt zu analysieren.
- *Bibelversionen*: Es werden insgesamt sieben verschiedene englischsprachige Bibelversionen verwendet. Ausgehend von der *King James Version* aus dem 16. Jh., welche noch sehr alte und archaische Wortformen enthält, werden die Verse auch mit anderen Versionen der Bibel verglichen, welche bspw. der hebräischen Satzsyntax folgen oder den Inhalt eines Verses mit möglichst einfacher Sprache wiedergeben.
- *Sammlungen von Redewendungen*: Weiterhin erfolgt die Analyse zweier kleiner deutschsprachiger Sammlungen von Redewendungen, welche einen Ursprung im Mittelalter bzw. einen Bezug zur Bibel haben.

Untersuchungsmethodik und Lösungsansatz

Da die *Data Diversity* aus informationstechnischer Sicht nicht mit einem einzelnen Algorithmus bzw. einer kleinen Menge von Ansätzen abgedeckt werden kann, wird in Kapitel 3 die *7-Level-Architektur* des *Historical Text Re-use* vorgestellt. Diese Architektur kann als ein modulares Konzept verstanden werden, um die *Text Re-use Analysis* auf die verschiedenen

[1] Als *Digital Library* wird eine digitale Kollektion von Texten verstanden.

Bedürfnisse, bedingt durch spezielle Eigenschaften von *Meme*, unterschiedlichen *Re-use Styles* aber auch verschiedenen Sprachvarianten, entsprechend anzupassen. Die einzelnen Level entsprechen den sieben Unteraufgaben *Segmentation*, *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring* und *Postprocessing*. In Kapitel 3 werden zu jedem Level in einem separaten Abschnitt entsprechende Implementierungen sowohl ausführlich vorgestellt als auch systematisiert. Zur Abgabe dieser Dissertation stehen in der *TRACER*-Implementierung, welche die *7-Level-Architektur* umsetzt, insgesamt über eine Million Kombinationsmöglichkeiten der verschiedenen Ansätze der einzelnen Level zur Verfügung.

Sowohl die drei genannten Forschungsfragen als auch die aufgezeigte *Data Diversity* des *Historical Text Re-use* werden im Rahmen der Dissertation als hinreichende Motivation verstanden, den *Historical Text Re-use* in Shannon's *Noisy Channel Theorem* einzubetten. In diesem Kontext kann ein Original- bzw. zitierter Autor als *Source* und ein wiederverwendender Autor als *Target* verstanden werden. Der *Noisy Channel* stellt ein unbekanntes Modell von Modifikationen, den äußeren Einflüssen, dar.

In Kapitel 4 wird das *Noisy Channel Model* dazu eingesetzt, ein zufälliges und rein künstliches Störsignal zum *Noisy Channel* hinzuzufügen, so dass eine *Randomised Digital Library* entsteht. Es werden insgesamt fünf Klassen von Randomisierungstechniken, die künstlichen Störsignale, im Sinne eines Turingtests vorgestellt, welche unterschiedliche Schwierigkeitsgrade einer rein quantitativen Evaluierung mit sich bringen. Für diese quantitative Evaluierung, die *Noisy Channel Evaluation*, wird der neuartige *Score* der *Mining Ability* eingeführt. Die *Mining Ability* setzt hierbei das Ergebnis einer *Text Re-use Analysis* auf einer *Digital Library* mit dem Resultat einer durch ein künstliches Störsignal veränderten *Randomised Digital Library* ins Verhältnis, wodurch nicht nur Parameter optimiert sondern auch verschiedene Sprachmodelle vollautomatisch und bzgl. des Ergebnisses ganzheitlich sowie ohne *Gold Standard* evaluiert werden können.

In Kapitel 5 wird der *Noisy Channel* als Modell eingesetzt, um historisch paradigmatische Relationen systematisch zu bestimmen. Das ist insbesondere unter Berücksichtigung der großen Zeitfenster von geisteswissenschaftlichen Texten von Interesse, da sich semantische Beziehungen von Konzepten im Laufe der Zeit verändert haben.

Ergebnisse

Die Ergebnisse dieser Arbeit sind sehr vielschichtig und umfassen neben Ergebnissen von Evaluierungen, auch Erfahrungen innerhalb der *eHumanities* sowie der entsprechenden Grundlagenarbeit. Im Detail können die Ergebnisse wie folgt zusammengefasst werden:

Es wird im einführenden Kapitel der Dissertation das Paradigma *ACID for the eHumanities* vorgestellt. *ACID* ist hierbei eine Abkürzung für *Acceptance*, *Complexity*, *Interoperability* und *Diversity*. Diese vier Säulen werden als Aspekte vorgestellt, denen sich die Informatik in der Zusammenarbeit mit den Geisteswissenschaften stellen muss. Der Fokus der Arbeit liegt auf der *Diversity* aber auch Aspekte der *Acceptance* und *Complexity* werden ausführlich verdeutlicht.

In Kapitel 4 wird neben der Einführung der *Noisy Channel Evaluation* auch aufgezeigt, welche statistischen Probleme probabilistische Sprachmodelle begleiten. Während probabilistische Sprachmodelle das *Gesetz der großen Zahlen* und somit eine hinreichend große Auftretenswahrscheinlichkeit voraussetzen, folgen verschiedene Charakteristika natürlicher Sprache einem *Power Law*, wie dem Zipfschen Gesetz, so dass für den *Long Tail* dieser Verteilung eine geringe Frequenz zugrunde liegt, woraus letztlich ein statistisches Problem resultiert. Im Detail kann so gezeigt werden, dass der eingeführte *Score* der *Mining Ability* bei zunehmender Größe einer *Digital Library* nach Erreichen eines Maximums wieder sinkt. Das resultiert daraus, dass mit zunehmender Größe der *Digital Library* vermehrt

aus Rauschen als Neuem "gelernt" wird. Auch wenn Kapitel 4 das auf den *Text Re-use* einschränkt, so sind die Ergebnisse einfach auf andere probabilistische Sprachmodelle adaptierbar. Insbesondere wird der Widerspruch des *Gesetzes der großen Zahlen*, welches den auf Wahrscheinlichkeiten aufsetzenden Sprachmodellen implizit zugrunde liegt, und den oftmals sehr seltenen Ereignissen beim Umgang mit natürlichsprachlichen Texten deutlich.

In Kapitel 5 wird weiterhin gezeigt, dass es kein *Text Re-use Model* gibt, welches in jedem Szenario optimale Ergebnisse liefert. Basierend auf sieben Bibelversionen mit unterschiedlichen Bezügen untereinander, wird verdeutlicht, dass sich nicht nur die Algorithmen der *7-Level-Architektur* unterscheiden können, sondern auch entsprechende Schwellwerte.

Im Rahmen der Arbeit werden zwei rein quantitative Evaluierungsgrößen, die *Text Re-use Compression* sowie die *Noisy Channel Evaluation*, eingeführt. In Kapitel 5 wird gezeigt, dass es eine signifikante Korrelation zu existierenden Evaluierungsgrößen gibt, welche jedoch einen *Gold Standard* oder zumindest eine Evaluierungsgrundlage benötigen. Einerseits gibt es eine nach Pearson sehr starke Korrelation zwischen dem *Recall* und der *Text Re-use Compression*. Andererseits wird auch gezeigt, dass das *F-Measure* sowie die im Rahmen dieser Arbeit eingeführte *Noisy Channel Evaluation* sehr vergleichbare Evaluierungsergebnisse erzeugen. Das wird im Rahmen einer *System Evaluation* in Kapitel 5 anhand der sieben Bibelversionen in insgesamt 504 verschiedenen Experimenten dargestellt.

Beitrag zur Forschung

Neben den aufgezeigten Ergebnissen stellt diese Arbeit Grundlagenforschung sowohl in der Systematisierung des *Text Re-use* aber auch bei der Evaluierung von Ergebnissen dar. Wie eingangs zum Forschungsstand umrissen wurde, verlieren sich derzeit viele Arbeiten in der nahezu beliebigen Kombination aus Daten und Algorithmen. Mit dieser Arbeit wird ein Evaluierungsszenario vorgestellt, welches es ermöglicht, auch ohne *Gold Standard* das Ergebnis zu bewerten. Somit wird das Resultat nicht mehr durch unterschiedliche Überlappungsgrade zwischen *Digital Library* und *Gold Standard* verfälscht.

Des Weiteren geht mit dieser Arbeit ein Paradigmenwechsel einher. Während in der Automatischen Sprachverarbeitung *Text Re-use* bisher aus einer "*1-Algorithmus-Sicht*" betrachtet wird, zeigen die Ergebnisse aus Kapitel 5 auf, dass zukünftig stärker der paarweise Vergleich zweier Werke im Forschungsvordergrund stehen sollte. Das geht damit einher, dass jeder Mensch einen eigenen *Re-use Style* besitzt, so dass durch das paarweise Vergleichen die menschlichen Individualitäten im Fokus der *Text Re-use Analysis* stehen. Deshalb wird vorgeschlagen, die Einzelergebnisse der werkweisen Vergleiche anschließend zu einem *Hybrid Text Re-use Graph* zusammenzusetzen. Mit der *Noisy Channel Evaluation* sowie der *Text Re-use Compression* stehen nun weiterführend auch vollautomatische Evaluierungstechniken zur Verfügung, so dass eine wesentlich präzisere *Text Re-use Analysis* möglich ist.

Perspektive

Entgegen modernen Anwendungen des *Text Re-use*, wie dem Plagiarismus, kann der *Historical Text Re-use* als ein nützliches Instrument verstanden werden, welches nicht nur Evidenzen von Transferwegen, sondern vielmehr auch einen fundamentalen Teil des sprachlich-kulturellen Erbes der Menschheit darstellt. Aus der Vielfalt des *Historical Text Re-use* ergeben sich für die Informatik im Rahmen der *eHumanities* vielschichtige Herausforderungen, die Gegenstand dieser Arbeit sind. Im Detail bedeutet das einen Paradigmenwechsel vom Pragmatismus im Vergleich von Sprachmodellen hin zur bestmöglichen Vollständigkeit.

Danksagung

Eine Dissertation ist wie ein Marathonlauf. Bei einer guten Leistung steht der Läufer im Vordergrund, auch wenn zeitgleich mindestens ein Trainer oder gar ein ganzer Trainerstab hinter diesem Erfolg steht sowie Teile des Privatlebens durch intensives Training von Verzicht geprägt sind. Mit diesem Abschnitt möchte ich meinen herzlichsten Dank den Menschen namentlich aussprechen, die im Hintergrund zu dieser Arbeit beigetragen haben.

Zuallererst gilt mein Dank meiner Frau **Anett Büchler**. Auch wenn ich die Dissertation einreiche, so ist dies ein kleines Stückchen Familienarbeit. Ich möchte mich natürlich nicht nur für das zahlreiche Korrekturlesen bedanken, sondern vor allem auch für die sehr gute Unterstützung insbesondere in den letzten Monaten. Immer wieder kleine, wenn mir auch im ersten Moment nicht immer in den "Kram" passende, Kurzurlaube waren in den letzten sechs Monaten dieser Arbeit Pole der Regeneration und des Kraftschöpfens. In jedem Fall vielen Dank für die Toleranz, mentale Unterstützung und Nachsicht, wenn ich Wochenenden teilweise komplett an der Dissertation "durchgemacht" habe. Ich hoffe, dass ich dies irgendwie wieder zurückgeben kann. In jedem Fall gelobe ich Besserung, auch wenn ich Dir das vielleicht besser nicht versprechen sollte. Danke!

Prof. **Gerhard Heyer** gilt mein Dank für die vergangenen 6,5 Jahre an seinem Lehrstuhl. In jener Zeit habe ich mich von einem Student zu einem Forscher entwickelt. Wichtig ist mir auch, explizit auf die entsprechend gewährten Freiheiten bei Prof. Heyer hinzuweisen. Während in der Forschung oftmals beobachtet werden kann, dass Professoren sich mit den Leistungen der Mitarbeiter schmücken, war es mir durch die von Prof. Heyer gewährten Freiheiten möglich, mich frei zu entwickeln und ein stabiles Fundament für eine Karriere in den *eHumanities* zu legen. Insbesondere meine bisherige Karriere und Reputation in diesem Bereich verdanke ich nicht zuletzt ihm. Seit 2008 konnten wir als Team auf politischer sowie inhaltlicher Ebene für die *eHumanities* in Leipzig enorme Fortschritte und internationale Sichtbarkeit erreichen.

Prof. **Gregory Crane** hat mich während eines sechsmonatigen Forschungsaufenthaltes in Boston als Geisteswissenschaftler für die Belange der *Humanities* am meisten geprägt. Weiterhin werde ich mich wahrscheinlich noch nach der Pensionierung an das Vor und Zurück für das *TPDL*-Papier erinnern, was als ein offener Konflikt über die Notwendigkeit von Algorithmen in den Geisteswissenschaften verstanden werden kann. Als Informatiker wollte ich Ansätze aus der Plagiarismusforschung, einfach auf antike Texten angewandt, einbringen, die wiederum zwar präzise waren, aber bei Weitem nicht der Ausbeute entsprach, die wir uns von automatischen Methoden versprochen hatten. Acht Monate und zahlreiche Experimente später konnten wir schließlich für beide Seiten sinnvolle Ergebnisse publizieren. Letztlich entsprechen diese acht Monate aktive Zusammenarbeit auf den ersten Blick zwar nur den viereinhalb Seiten in Abschnitt 5.2, jedoch sind die Erfahrungen daraus zu 100% auf das gesamte Ergebnis-Kapitel extrapolierbar. Mein Verständnis für viele Belange und Sichtweisen der *Humanities* verdanke ich Prof. Crane, der mir immer wieder geduldig Dinge erklärt hat. Vielen Dank für diesen Prozess und die damit verbundene Entwicklung in den letzten vier Jahren.

Ich möchte mich bei Dr. **Helge Kahler** (Bundesministerium für Bildung und Forschung) und Dr. **Hans Nerlich** (Deutsches Zentrum für Luft- und Raumfahrt) für einen sehr gut finanziell ausgestatteten Auslandsaufenthalt an der Tufts University in Boston, USA, bedanken. Aus heutiger Sicht fällt diese Zeit nicht nur unter Erfahrungen, wie "einmal etwas anderes gesehen haben" oder "über den Tellerrand schauen", sondern hat mir wahrscheinlich für die nächsten Jahre den nötigen Drall in Richtung der *eClassics* gegeben, der sicherlich auch diese Arbeit mitgeprägt hat.

Dr. **Ute Pietruschka** gilt mein Dank für zahlreiche Hilfestellungen zu semitischen Sprachen, wie Arabisch. 2011/12 konnte ich zwei Semester bei Ute an einem Arabisch-Kurs teilnehmen. Geholfen hat das insbesondere, um ein besseres Verständnis für Sprachfamilien und vor allem dem Studium antiker Sprachen zu bekommen. Vielen Dank für Deine sehr ausführlichen Erklärungen zu Übersetzungstechniken, die mir letztlich gezeigt haben, dass die maschinelle Übersetzung niemals richtig funktionieren wird. Im Nachhinein war genau das wahrscheinlich auch der Grund, warum ich mich in der Dissertation auf *Text Re-use* innerhalb einer Sprache konzentriert habe.

Auch Dr. **Monica Berti** hat im Hintergrund, wie sicherlich auch durch diverse Fußnoten explizit gekennzeichnet, nicht unerheblichen Anteil an der fachwissenschaftlichen Betreuung dieser Arbeit. Das umfasst das Finden von passenden Belegstellen für die Dissertation, fachwissenschaftlichen Erklärungen aber auch verschiedenen gemeinsamen Vorträgen sowie Publikationen. Monica hat sich im Rahmen dieser Dissertation zu "meinem kleinen Orakel für altertumswissenschaftliche Angelegenheiten" entwickelt. Zu nahezu jeder Tageszeit konnte ich auf sie zugehen und bekam immer ein promptes Feedback. Nicht grundlos kann ich bei Monica auch mit zwinkerndem Auge von meiner Arbeitsehefrau sprechen!

Thomas Eckart, Kollege und aus dem Berufsleben auch Freund geworden, danke ich für die sehr gute Zusammenarbeit insbesondere während des *eAQUA*-Projektes. Vieles wäre nicht ganz so erfolgreich gelaufen, wenn ich Thomas nicht mit im Projekt gehabt hätte.

Dem *eTRACES*-Team, namentlich in alphabetischer Reihenfolge **Markus Ackermann**, **Frederik Baumgardt**, **Petra Gamrath**, **Annette Geßner**, **Stefan Jänicke**, **Christian Kötteritzsch** und **Maria Moritz**, möchte ich insbesondere für die Unterstützung der letzten drei Monate danken, in welchen ich nahezu komplett nicht im Projekt stand, um diese Arbeit zu beenden.

Petra Gamrath und **Renate Schildt** gilt mein Dank dafür, mir regelmäßig den Papierkram, wie Dienstreiseanträge, abgenommen zu haben. Auch wenn es vermeintlich nur kleine Beträge sind, so macht es bei der Fülle meiner Dienstreisen ziemlich viel Zeit aus, in der ich mich mit anderen Dingen beschäftigen konnte. Auch der stabile Nachschub mit Kaffee soll gedankt und nicht als selbstverständlich hingenommen sein.

Für zahlreiche Korrekturvorschläge gilt mein Dank **Gerhard Heyer**, **Anett Büchler**, **Thomas Eckart**, **Annette Geßner**, **Thomas Efer**, **Maria Moritz**, **Volker Boehlke** sowie **Petra Gamrath**.

Letztlich möchte ich mich auch bei den 24 Testpersonen namentlich bedanken, die für ein Experiment im Rahmen dieser Arbeit zur Verfügung standen (in alphabetischer Reihenfolge): **Markus Ackermann** (*eTRACES*, Leipzig), **Frederik Baumgardt** (*eTRACES*, Halle/S.), **Volker Boehlke** (*CLARIN*, Leipzig). **Anett Büchler** (Leipzig), **Gabriele Büchler** (Delitzsch), **Thomas Eckart** (*CLARIN*, Leipzig), **Thomas Efer** (*eXCHANGE*, Leipzig), **Lars Gadegast** (Leipzig), **Annette Geßner** (*eTRACES*, Leipzig), **Katarina Jacob** (Rackwitz), **Ronny Jacob** (Rackwitz), **Stefan Jänicke** (*eTRACES*, Leipzig), **Christian Kötteritzsch** (*eTRACES*, Leipzig), **Matthias Leopold** (Deutsche Zentralbibliothek für Blinde, Leipzig), **Sebastian Lissmann** (*eAQUA*, Leipzig), **Maria Moritz** (*eTRACES*, Leipzig), **Clemens Neudecker** (Niederländische Nationalbibliothek, Den Haag, Niederlande), **Ute Pietruschka** (*CASG*, Halle/S.), **Elke Rosenkranz** (Markkleeberg), **Martin Schierle** (Ulm), **Thomas Stäcker** (Herzog-August-Bibliothek, Wolfenbüttel), **Lydia Steiner** (Institut für Medizinische Informatik, Statistik und Epidemiologie, Leipzig), **Sabine Thänert** (Deutsches Archäologisches Institut, Berlin) sowie **Diana Winger** (Leipzig).

Inhaltsverzeichnis

Zusammenfassung IV
Inhaltsverzeichnis XII
Abbildungsverzeichnis XVI
Tabellenverzeichnis XXI

1 Einführung **25**
1.1 Ein erster Überblick 26
1.2 Re-use in der Natur 28
1.3 Wissenstransfer im interdisziplinären Spannungsfeld 31
1.4 Information Overload vs. Information Poverty 32
1.5 Das "ACID for the eHumanities" Paradigma 38
1.6 Herausforderungen des textuellen Wissenstransfers auf geisteswissenschaftlichen Texten 41
1.7 Wissenschaftliche Einbettung des historischen Wissenstransfers in den Forschungsbereich der Informatik 46
1.8 Verwandte Themengebiete in der Informatik 50
1.9 Ausblick und Gliederung der Arbeit 52

2 Grundlagen **55**
2.1 Einführung 56
2.2 Humanities, Digital Humanities, eHumanities, Computer Science: Das 4-Sichten-Modell des *Historical Text Re-use* 57
2.3 Das 3-Generationen-Modell: Die Geschichte der Text Re-use Algorithmen . 62
2.4 Qualitätskriterien für Text Mining 62
2.5 Grundterminologien und Definitionen 64
2.6 Systematisierung des geisteswissenschaftlichen und informationstechnischen *Text Re-use* 68
2.7 Text Re-use Tasks 79
2.8 Noisy Channel Theorem und Conditional Kolmogorov Complexity 83

3 Historical Text Re-use Detection **87**
3.1 Einführung 88
3.2 Level 1: Segmentation 91
3.3 Level 2: Preprocessing 93
3.4 Level 3: Featuring 100
3.5 Level 4: Selection 104
3.6 Level 5: Linking 113
3.7 Level 6: Scoring 116
3.8 Level 7: Postprocessing 120
3.9 Wechselwirkungen zwischen den einzelnen Level 123
3.10 Text Re-use Compression 124

4 Zufall und Struktur **127**
4.1 Einführung 128
4.2 Probleme von Sprachmodellen 129
4.3 Evaluierung von Sprachmodellen 137
4.4 Noisy Channel Evaluation 141

4.5 Arten einer *Randomised Digital Library* 149
4.6 Eigenschaften der *Noisy Channel Evaluation* 151
4.6.1 *Mining Ability* in Abhängigkeit von der Größe einer *Digital Library* bei konstantem Parameterraum Θ 154
4.6.2 *Mining Ability* in Abhängigkeit vom Parameterraum Θ bei konstanter Größe einer *Digital Library* 158
4.6.3 Minimale und maximale *Mining Ability* $\mathcal{L}_{min}(\Theta)$ und $\mathcal{L}_{max}(\Theta)$ bei einem dynamischen Parameterraum Θ sowie unterschiedlich großen *Digital Libraries* 161
4.7 Einbettung dieses Kapitels in die gesamte Arbeit 163

5 Ergebnisse 167
5.1 Einführung 168
5.2 *Text Re-use* in der *Perseus Digital Library* 171
5.2.1 Level 3 - *Featuring*: Bigram Shingling vs. Unigram 172
5.2.2 Level 1 - *Segmentation*: *Sentence*-basierte und nicht überlappende Segmentierung vs. *Moving Window* mit fester Fenstergröße 172
5.2.3 Level 5 - *Scoring*: normalisierte Gewichtung vs. absolute Gewichtung des *Re-use Overlap* 173
5.2.4 Evaluierung auf der *Perseus Digital Library*: *Wieviel Homer steckt in Athenaeus*? 174
5.2.5 Zusammenfassung 175
5.3 System Evaluation 176
5.3.1 Evaluierung durch *Precision & Recall* gegen einen *Gold Standard* . . 178
5.3.2 Evaluierung durch *Noisy Channel Evaluation* 185
5.3.3 Evaluierung durch *Text Re-use Compression* 189
5.3.4 Zusammenfassung 192
5.4 Component & Aggregated Evaluation 201
5.4.1 Qualität der *Lemmatisation* 202
5.4.2 Qualität im Umgang mit paradigmatischen Relationen 203
5.4.3 Qualität im Umgang mit historischen Varianten 204
5.4.4 Qualität der *Digital Signature* 205
5.4.5 Qualität des *Linking* 209
5.5 *Noisy Channel Mining*: Extraktion paradigmatischer und historischer Schreibweisen 212
5.6 Zusammenfassung 215

6 Zusammenfassung 217
6.1 Ziele und Ergebnisse dieser Arbeit 218
6.2 Lessons Learnt 225
6.3 Weiterführende Aspekte und zukünftige Arbeiten 227

A χ^2-Tabelle 233

B Wissenschaftlicher Werdegang 245
B.1 wissenschaftlicher Lebenslauf 246
B.2 Praxiserfahrungen in der Informatik sowie in den *Humanities* 246
B.3 Auszeichnungen und Preise 247
B.4 Aktivitäten in *Advisory Boards*, als *Reviewer* und *Initiator*, in *Programmkomitees* sowie Unterstützung anderer Forscher und Projekte 247

B.5 Bearbeitete wissenschaftliche Projekte 248
B.6 Wissenschaftliche Akquise 249
B.7 Organisierte Workshops 249
B.8 Expert Talks, Invited Talks und Interviews 250
B.9 Besuchte Veranstaltungen 251
B.10 Bücher, Whitepaper und strategische Dokumente 252
B.11 Publikationen 252
B.12 Vorträge 254
B.13 Poster und Posterdemonstrationen 256
B.14 Lehrveranstaltungen 256
B.15 Betreute Abschlussarbeiten 257

Literaturverzeichnis **259**

Abbildungsverzeichnis

1.1 *Microview* einer Liste von Zitaten zur Unterstützung der geisteswissenschaftlichen Arbeit im Bereich der Textkritik - Platon Timaeus 91b7 ff. 35

1.2 *Dotplotview* zum paarweisen Vergleich zweier Werke, um systematisches Abschreiben zu identifizieren - Vergleich der Bücher Lukas und Markus aus der Bibel (Bild-Quelle: [Lee 2007]). 35

1.3 *Temperature Maps* eines Werkes bzgl. des *Text Re-use* in einer *Digital Library* zur Analyse, welche Teile eines Werkes besonders häufig referenziert bzw. wiederverwendet worden sind. 36

1.4 *Macroview* einer Digital Library, um einfach grobe Zusammenhänge zu explorieren wie bspw. der Korrelation zwischen dem Zitieren bestimmter Textpassage zu bestimmten Zeitpunkten, wie dem Neu- und dem Mittelplatonismus. Das linke obere Diagramm segmentiert einen *Re-use Graphen* nach der Zeit (x-Achse) und trägt die Häufigkeit für jede Zeitscheibe auf der y-Achse ab. Die rechte obere Darstellung zeigt die zitierenden Autoren auf. Die untere Darstellung repräsentiert die seitenweise Zerlegung des Werkes (Stephanus-Seiten) und trägt auf der y-Achse die Zitierhäufigkeit der jeweiligen Seite ab. 38

1.5 Die Abbildung zeigt Anmerkungen von *Axel Ahlberg's* Edition aus dem Jahre 1913 zu *Sallust's Catilinarian Conspiracy*. Mehr als 20 Verweise auf zitierende Textstellen sind im Apparat aufgelistet. 41

1.6 Shannon's *Noisy Channel*: Ein *Signal* $\mathcal{S}$ wird über einen *Noisy Channel* vom *Transmitter* zum *Receiver* übertragen. Das empfangene Signal $\mathcal{S}'$ weicht je nach *Rauschpegel* des *Noisy Channels* vom Original unterschiedlich stark ab. 47

1.7 Shannon's *Noisy Channel* mit einem künstlichen Störsignal (schwarze Box): Ein *Signal* $\mathcal{S}$ wird über einen *Noisy Channel* zum *Transmitter* übertragen. Das empfangene Signal $\mathcal{S}'$ weicht je nach dem *Rauschpegel* des Störsignals des *Noisy Channels* vom Original ab. 49

2.1 Vier-Säulen-Modell des *Historical Text Re-use*: Der Forschungsbereich ist durch die unterschiedlichen Fragestellungen der *Humanities*, *Digital Humanities*, *eHumanities* und der *Computer Science* definiert. Jeder dieser Bereiche hat unterschiedliche Fragestellungen an den *Text Re-use* (innerer Ring). Daraus ergeben sich vielfältige Wechselwirkungen zwischen den vier Bereichen (mittlerer Ring). Der äußere Ring ordnet existierende Projekte gemäß ihrer hauptsächlichen Fragestellungen an. 58

2.2 Es gibt sechs Parallelstellen zu Platon, Timaeus 91b7 ff, welche durch einen Algorithmus aufgedeckt werden konnten. Zu den insgesamt sieben Fundstellen im *Thesaurus Linguae Graecae* gibt es vier verschiedene Satzanfänge. Das erste Inhaltswort *metrai* (Gebärmutter) kommt in Platon erst an der siebten Stelle in der *Re-use Unit*. Die sechs Wörter davor wurden entweder angepasst oder in der Niederschrift entfernt. 60

2.3 Ein fachwissenschaftliches *Edge Type System* auf Basis eines dreistufigen Entscheidungsbaumes. Die Wurzel ist *Parallel Text*. Die erste Unterscheidung wird nach dem *Type* des *Text Re-use* gemacht. Die zweite Unterscheidung richtet sich nach der Größe (*Size*). Die dritte Entscheidung ist vom Grad der Veränderung zwischen beiden *Re-use Units* bestimmt. 77

2.4 Ein Auflistung der wichtigsten *Text Re-use Tasks*. *Text Re-use Tasks* können nach *Algorithmic Re-use Detection*, *Graph Based Tasks*, *Semantic Detection* und *Noisy Channel Mining* geclustert werden. 81

2.5 Systematisierung der *Re-use Variants*. In Anlehnung an die drei Textoperationen *Insertion*, *Substitution* sowie *Deletion* der *Levenshtein Distance* Metrik zeigt die Systematisierung diffizilere Textoperationen, die über ein *Philological Crowd Sourcing* gesammelt werden können. 85

3.1 Taxonomie des Level *Segmentation* für den *Historical Text Re-use*. Im Sinne eines Entscheidungsbaumes werden die beiden dominanten Entscheidungen a) überlappende oder disjunkte *Re-use Units* und b) Größe (inkl. statischer oder dynamischer Fenstergröße) hierarchisch dargestellt. 92

3.2 Taxonomie des Level *Preprocessing* für den *Historical Text Re-use*. Als Entscheidungsbaum werden verschiedene *Preprocessing*-Schritte dargestellt. . . 97

3.3 Gewichtung von *gerichteten Graphen* durch eine *PageRank*-ähnliche Technik (vgl. [Brin 1998]). Für x_3 kann sowohl y_1 als auch y_2 gelten. Auf Basis der Frequenzen $freq(y_1) = 10$ und $freq(y_2) = 20$ müsste sich für $y_{max} = y_2$ entschieden werden. Unter Berücksichtigung der zu y_i eingehenden Kanten werden die Frequenzen $freq(x_i)$ für alle eingehenden x_i zu y_j addiert. Unter dieser zusätzlichen Bedingung gilt nun $y_{max} = y_1$ als der vertrauensvollste Kandidat für x_3. 98

3.4 Konvertierung eines *ungerichteten Graphen* zu einem *gerichteten Graphen*. Während der sprachlichen Normalisierung des *Preprocessing* können *ungerichtete Graphen*, wie *WordNet* (vgl. [Miller 1995, Fellbaum 1998]), dazu eingesetzt werden, um entsprechende sprachliche Vielfalt zu harmonisieren. Hierzu wird auf Basis von Frequenzen $freq(x_i)$ eines Knotens x_i eine Gewichtung vorgenommen. 99

3.5 Taxonomie des Level *Featuring* für den *Historical Text Re-use*. Grundlegend kann zwischen den drei verschiedenen *Featuring*-Klassen *Syntactical Fingerprinting*, *Semantic Fingerprinting* sowie den *Non-statistic Approaches* unterschieden werden. 102

3.6 Taxonomie des Level *Selection* für den *Historical Text Re-use*. Die Abbildung zeigt die Klasse des *Global Selection Knowledge* sowohl im *Global* als auch *Local Selection Usage*. Grundlegende Techniken des *Global Selection Knowledge* können in den meisten Fällen in beiden *Selection Usage* eingesetzt werden. Alle hier vorgestellten Implementierungen können in der *TRACER*-Implementierung (vgl. [Büchler 2013a]) beliebig miteinander kombiniert werden. 109

3.7 Taxonomie des Level *Selection* für den *Historical Text Re-use*. Die Abb. zeigt die Klasse des *Local Selection Knowledge* sowohl im *Global* als auch *Local Selection Usage*. 111

3.8 Taxonomie des Level *Linking* für den *Historical Text Re-use*. Die Abbildung zeigt die Klassen des *Local Linking* und *Distributed Linking*. 116

3.9 Taxonomie des Level *Scoring* für den *Historical Text Re-use*. Diese Taxonomie schlüsselt verschiedene Techniken auf. Die Unterscheidung zwischen einem *Scoring* auf der *Word*- und *Feature*-Ebene hängt vom Einsatz ab. Während die *Feature*-Ebene aus technischer Sicht zu bevorzugen ist, sind für Anzeigen in Benutzerschnittstellen oftmals die *Word*-Ebene besser geeignet, da die *Scoring*-Werte einfacher nachzuvollziehen sind. 118

3.10 Taxonomie des Level *Postprocessing* für den *Historical Text Re-use*. Diese Taxonomie schlüsselt verschiedene Techniken aus der Bibliometrie, der phonetischen Analyse, den *Text Re-use Tasks* aus Abschnitt 2.7, den *Clusteranalysen*, sowie weiteren Techniken des *Graph Mining* auf. Gemeinsames Ziel aller *Postprocessing*-Techniken ist das Vereinfachen der Daten im Sinne des *Information Overload* vs. der *Information Poverty* aus Abschnitt 1.4. . 121

4.1 Über 90% aller in einer *Digital Library* enthaltenen Wörter werden zehnmal oder seltener beobachtet. Bedingt durch die zugrunde liegende *Power Law*-Verteilung werden bei einer Wortassoziationsanalysen statistische Probleme induziert. Den Abbildungen 4.1(a) bis 4.1(l) liegt eine *Bigram*-Analyse zugrunde, wobei die Wörter auf den beiden Achsen mit logarithmischer Skalierung abgetragen werden. Jedes beobachtete *Bigram* wird nach Formel 4.5 mit $f_{eps} \geq 1$ analysiert. Die schwarze Fläche entspricht den *Bigrams*, welche den Test bestanden haben. Die weiß markierten Flächen repräsentieren die *Bigrams*, welche den Test nicht bestanden haben oder nicht beobachtet worden sind. 132

4.2 Ausgehend von einer *Digital Library* mit *250M* Sätzen wird der Verlauf der Testbedingungen von $f_{epx} \geq 1$ bis $f_{epx} \geq 100$ abgebildet. 133

4.3 Versuchsaufbau der *Noisy Channel Evaluation*. Bei dieser Form der quantitativen Evaluierung werden die Ergebnisse sowohl eines natürlichen als auch eines zufälligen Signales verglichen und formen den *Score* der *Mining Ability* einer quantitativen Evaluierung. 143

4.4 *Mining Ability* $\mathcal{L}(\Theta)$ in dB für deutsche und englische *Bigrams* und *Co-occurrences* in Abhängigkeit von der Größe der *Digital Library* (Normkorpora der *Leipzig Linguistic Collection* vgl. [Biemann 2007a, Goldhahn 2012]). . . 155

4.5 *Mining Ability* $\mathcal{L}(\Theta)$ in dB für deutsche und englische *Bigrams* und *Co-occurrences* in Abhängigkeit vom Schwellwert des *Log-Likelihood-Ratio* -2λ. 159

4.6 *Minimum* und *Maximum Mining Ability* $\mathcal{L}_{min}(\Theta)$ und $\mathcal{L}_{max}(\Theta)$ für deutsche sowie englische *Bigrams* und *Co-occurrences*. 162

5.1 Taxonomie von Evaluierungstechniken für den *Historical Text Re-use*. Aus der *Biometrie* (vgl. [Maltoni 2009, BIMA 2012]) kann sowohl die *System* als auch die *Component Evaluation* adaptiert werden, um ein Verfahren bzw. ein Teil eines Verfahrens zu evaluieren. 169

5.2 Verteilung der Länge des *Re-use Overlaps* gemessen an 353 manuell gesammelten Datensätzen im Rahmen einer fachwissenschaftlichen Lehrveranstaltung. 173

5.3 Drei verschiedene Formen der Annotation für den *Historical Text Re-use*. Je nachdem wie gesichert ein *Text Re-use* angenommen werden kann, ist es möglich, durch das *quote*-Tag sowohl die Grenzen als auch die genaue Position des *Text Re-use* in *XML* festzuhalten. 174

5.4 Ergebnis der Evaluierung. Es wird sowohl nach Quelle, die *Odyssey* und die *Iliad* beide von Homer, als auch den drei Annotationsformen (vgl. Abb. 5.3) unterschieden. 175
5.5 *Precision-Recall-Plot* für *KJV*, *BBE*, *YLT* sowie *WBS* (Spalten) und *Word based Featuring*, *Bigram* sowie *Trigram Shingling* (Zeilen). Die Abbildungen reflektieren das unterschiedliche Verhalten von *Precision* und *Recall* beim paarweisen Vergleich von Bibelversionen. 184
5.6 Vergleich der Evaluierungsmetriken *Precision*, *Recall* und *F-Measure* in Abhängigkeit vom *Scoring*-Schwellwert t. Es wird der Unterschied in F_{max} bei der Analyse von *WBS* vs. *BBE* sowie *WBS* vs. *KJV* verglichen. 185
5.7 Verlauf des gefundenen *Text Re-use* in Abhängigkeit zum *Scoring*-Schwellwert t. Sowohl für das *Trigram* als auch *Bigram Shingling* sowie das *Word based Featuring* wird der insgesamt gefundene sowie der sich mit einem *Gold Standard* überlappende *Text Re-use* abgebildet. 194
5.8 *F-Measure* F in Abhängigkeit vom *Scoring*-Schwellwert t. In den Abbildungen wird das Verhalten des *F-Measure* gegen t für das *Trigram Shingling*, *Bigram Shingling* sowie für vier ausgewählte Bibelversionen dargestellt. In jedem Plot wird das Verhalten des *F-Measure* bzgl. der anderen sechs Bibelversionen aufgezeigt. 196
5.9 *Noisy Channel Evaluation* in Abhängigkeit vom *Scoring*-Schwellwert t. In den Abbildungen wird das Verhalten der *Mining Ability* $\mathcal{L}(\Theta))$ in *dB* gegen t für das *Trigram Shingling*, *Bigram Shingling* sowie für vier ausgewählte Bibelversionen dargestellt. In jedem Plot wird das Verhalten der *Mining Ability* bzgl. der anderen sechs Bibelversionen aufgezeigt. 197

Tabellenverzeichnis

1.1 Wechselwirkung zwischen Sprache und Kommunikation: Der *Historical Text Re-use* hängt in der Gegenwart sehr davon ab, ob bestimmte *Re-use Units* im Laufe der Zeit aufgeschrieben worden sind. Zielstellung einer ganzheitlichen Betrachtung des *Historical Text Re-use* würde sein, alle vier in der Tabelle genannten Quadranten in Betracht ziehen zu können. 42

2.1 *Node Types* für ausgewählte *Meme*: *Adage*, *Abstract*, *Anagram*, *Aphorism*, *Apophthegm*, *Battle Cry*, *Bonmot*, *Cliché*, *Definition* 71
2.2 *Node Types* für ausgewählte *Meme*: *Edition*, *Epigram*, *Epithet*, *Epitome*, *Fact*, *Flowery Phrase*, *Gnome*, *Idiom* 72
2.3 *Node Types* für ausgewählte *Meme*: *Joke*, *Koan*, *Law*, *Legend*, *Loanword*, *Mantra*, *Maxim*, *Meme*, *Metaphor*, *Motto* 73
2.4 *Node Types* für ausgewählte *Meme*: *Motto*, *Palindrom*, *Pangram*, *Parable*, *Paroimia*, *Phraseme*, *Platitude*, *Proverb*, *Punch Line* 74
2.5 *Node Types* für ausgewählte *Meme*: *Quip*, *Rant*, *Saw*, *Sententiae*, *Simile*, *Slogan*, *Template*, *Truism*, *Wit* 75

3.1 *7-Level-Architektur* des *Historical Text Re-use*. Die Tabelle bildet für die sieben Level *Segmentation*, *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring*, *Postprocessing* das benutzte Formelzeichen sowie den jeweiligen *Input* und *Output* der einzelnen Level ab. Bereits eingeführte Formelzeichen sind an den jeweiligen Stellen genannt. 90
3.2 *Selection Knowledge* vs. *Selection Usage*. Die Matrix vergleicht *Pros* und *Kontras* zwischen den jeweiligen Kategorien von *Selection*-Verfahren. Verfahren der vier Kategorien sind in den Abb. 3.6 und 3.7 abgebildet. *Global Selection Knowledge* bei *Local Selection Usage* bietet den besten Kompromiss aus genannten Vor- und Nachteilen. 107

4.1 Diese Tabelle stellt die Ergebnisse von Experimenten mit einer Normgröße von *1M*, *30M* und *100M* Sätzen dar. Untersucht werden die Anzahl derjenigen *Bigrams*, die die Testbedingung $f_{epx} \geq i$ mit $i \in \{1, 3, 5, 10, 20, 50, 75, 100\}$ erfüllen. 131
4.2 Diese Tabelle stellt die Ergebnisse aus Formel 4.7 dar. Hierzu wurden Experimente auf verschiedene Normgrößen, den Spalten, sowie zu vier verschiedenen Approximationen der Wahrscheinlichkeit $p(w_j)$, den Zeilen, durchgeführt. . 134
4.3 Vier mögliche Ausgänge einer Evaluierung gegen einen *Gold Standard*. . . . 138
4.4 Zusammenhang zwischen dem Fehler ε und der *Mining Ability* $\mathcal{L}_{Quant}$. . . . 145
4.5 Grobklassifikation von *Methoden* und deren *Mining Ability* $\mathcal{L}(\Theta)$ inklusive einer Wertung. Der angegebene Leistungsbereich und dessen Wertung basiert auf Erfahrungen im Umgang mit der *Mining Ability*. 146
4.6 Vergleich des qualitativen und quantitativen *Re-use*. 147
4.7 Grobklassifikation von *Randomisierungsmethoden*. Die erste Spalte kann als eine *Turing-Test*-Klassifikation verstanden werden, wobei $T1$ im Sinne eines *Turing-Tests* als einfach zu erkennen und $T5$ als ein schwieriger *Turing-Test* anzusehen ist. Je nach *Turing-Test*-Klassifikation sinkt oder steigt die eingeführte *Mining Ability* $\mathcal{L}(\Theta)$. 150

4.8 *Entropie-Test* ΔH^n für verschiedene Anzahl der Iterationen (Zeilen) sowie auf unterschiedlichen Größen der *Digital Library* (Spalten). 153

5.1 Testsystematik dieses Kapitels bezogen auf die *7-Level-Architektur* des *Historical Text Re-use*. Aufgelistet sind insgesamt sechs verschiedene Tests (1. Spalte), die in den folgenden Abschnitten (2. Spalte) durchgeführt werden. *L1* bis *L7* entsprechen den sieben Level *Segmentation*, *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring*, *Postprocessing*. 171
5.2 Sieben verschiedene Versversionen aus *Buch Genesis, Kapitel 1, Vers 1*. . . 177
5.3 Grundlegende sprachstatistische Kennzahlen für die 28632 gemeinsamen Verse der sieben eingesetzten Bibelversionen. 177
5.4 S_{ij}-Matrix für i *Preprocessing*- und j *Featuring*-Techniken, die im Rahmen der Evaluierung untersucht werden. 178
5.5 *Precision* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß der berechneten *Precision* zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt. 180
5.6 *Recall* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß des berechneten *Recall* zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt. 181
5.7 *F-Measure* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß des berechneten *F-Measure* zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt. 182
5.8 *Modified Noisy Channel Evaluation* in dB für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß der maximalen *Mining Ability* zwischen 0.0 (weiß) und 44.568 dB(schwarz) festgelegt. . 186
5.9 *Noisy Channel Evaluation* in dB für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. 187
5.10 *Modified Text Re-use Compression* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß der maximalen Kompression zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt. 190
5.11 *Text Re-use Compression* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. 191
5.12 Pearson's Korrelationskoeffizient $\rho(X;Y)$ zwischen verschiedenen Evaluierungsmetriken. 193
5.13 Pearson's Korrelationskoeffizient $\rho(X;Y)$ zwischen *Recall* R und *Text Re-use Compression* bzgl. unterschiedlicher Schwellwerte t sowie für drei *Featuring*-Techniken. 195
5.14 (t, F_{max})-Tupel für *Trigram Shingling* bei vier *Preprocessing*-Techniken im Vergleich von *KJV* zu den anderen Bibelversionen. 199
5.15 (t, F_{max})-Tupel für *Bigram Shingling* bei vier *Preprocessing*-Techniken im Vergleich von *KJV* zu den anderen Bibelversionen. 199
5.16 (t, F_{max})-Tupel für *Word based Featuring* bei vier *Preprocessing*-Techniken im Vergleich von *KJV* zu den anderen Bibelversionen. 199

5.17 *Perseus Tag System*. Die Tabelle repräsentiert die 14 verschiedenen *Part of Speech*-Tags des *Perseus Tag System*. 206
5.18 *Feature Density* pro *PoS*-Tag aus Tabelle 5.17. 207
5.19 *Linking*-Analyse mit unterschiedlichen *Featuring*-Techniken. Die Tabelle reflektiert die *Linking*-Analyse für das *Trigram Shingling*, *Bigram Shingling* sowie das *Word based Featuring*. Die Ergebnisse basieren auf dem *Preprocessing* ξ_4 der *System Evaluation* von Seite 176 ff. Sowohl *Unique Links* als auch *Linked Links* sind in Millionen (M) angegeben. 210
5.20 *Linking*-Analyse mit unterschiedlichen *Feature Density* $\mathcal{F}$. Die Tabelle reflektiert die *Linking*-Analyse für unterschiedliche *Feature Density* $\mathcal{F}$. Die Ergebnisse basieren auf dem *Preprocessing* ξ_4 der *System Evaluation* von Seite 176 ff. 211
5.21 *Linking*-Analyse mit einer wortartbasierten *Selection*. Die Wahl der zwei Mengen von berücksichtigten *PoS*-Tags (vgl. Tabelle 5.17 auf Seite 206) basiert auf Ergebnissen einer manuellen *Selection* von insgesamt 24 Probanden (vgl. Abschnitt 5.4.4). 211
5.22 Ergebnisse des *Noisy Channel Mining* auf sieben englischsprachigen Bibelversionen. Die Ergebnisse einer *Noisy Channel Mining*-Analyse sind insgesamt in neun verschiedene Relationstypen klassifiziert worden. Insgesamt sind 8193 der etwa 12000 extrahierten Assoziationen klassifiziert. 213

6.1 Vergleich durch Pro und Kontra für eine *Text Re-use Analysis* in der Korpuslinguistik sowie der Informatik. 226

A.1 χ^2_α-Signifikanzwerte für 1 bis 25 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 234
A.2 χ^2_α-Signifikanzwerte für 26 bis 50 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 235
A.3 χ^2_α-Signifikanzwerte für 51 bis 75 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 236
A.4 χ^2_α-Signifikanzwerte für 76 bis 100 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 237
A.5 χ^2_α-Signifikanzwerte für 101 bis 125 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 238
A.6 χ^2_α-Signifikanzwerte für 126 bis 150 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 239
A.7 χ^2_α-Signifikanzwerte für 151 bis 175 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 240
A.8 χ^2_α-Signifikanzwerte für 176 bis 200 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 241
A.9 χ^2_α-Signifikanzwerte für 201 bis 225 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 242
A.10 χ^2_α-Signifikanzwerte für 226 bis 250 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler). 243

Kapitel 1

Einführung

Contents

1.1 Ein erster Überblick 26
1.2 Re-use in der Natur 28
1.3 Wissenstransfer im interdisziplinären Spannungsfeld 31
1.4 Information Overload vs. Information Poverty 32
1.5 Das "ACID for the eHumanities" Paradigma 38
1.6 Herausforderungen des textuellen Wissenstransfers auf geisteswissenschaftlichen Texten 41
1.7 Wissenschaftliche Einbettung des historischen Wissenstransfers in den Forschungsbereich der Informatik 46
1.8 Verwandte Themengebiete in der Informatik 50
1.9 Ausblick und Gliederung der Arbeit 52

To know the breeding system is to know the genetic architecture of a species.
To know the evolution of a breeding system is to know how evolution works.

Lewis & Crowe, Evolution, (1955)

Text Re-use ist ein breiteres Forschungsfeld als oftmals im ersten Moment wahrnehmbar. Es umfasst einerseits *absichtlichen Re-use*, wie bspw. *Zitate*, *Paraphrasen* oder auch *Allusionen*. Andererseits zeigen entsprechende Techniken auch *unabsichtlichen Re-use*, wie *Idiome*, *Battle Cries*, *Multi Word Units* oder auch nur syntaktischen *Text Chunks*, an.

Dieses erste Kapitel gibt eine Einführung in den Bereich des historischen Text Re-use. Es wird mit verwandten Wissenschaften im Allgemeinen begonnen und zum Ende hin systematisch auf den *Historical Text Re-use* eingeschränkt. Abschließend folgt eine Gliederung der Arbeit sowie eine Abgrenzung verwandter Themen der Automatischen Sprachverarbeitung. Ziel dieses Kapitels ist im Besonderen, auf Spannungen hinzuweisen aber auch Möglichkeiten der *eHumanities* aufzuzeigen.

1.1 Ein erster Überblick

Text Re-use bedeutet Effektivität. Durch das Wiederverwenden von Texten, wie durch das *Single Sourcing* oder auch dem Plagiarismus, wird in erster Linie eine Kostenersparnis in Form von Zeit bewirkt. Auch wenn sowohl das *Single Sourcing* als auch der Plagiarismus zwei gegensätzliche Konzepte bzgl. der Rechtsauslegung sind, so haben sie die Methodik des Erkennens durch den *Text Re-use* gemeinsam.

Den Aspekt der juristischen Rechtsprechung, der durch *Plagiarismus* impliziert wird, sei im Rahmen dieser Arbeit vollständig unberücksichtigt. Um dies zu verstärken, wird der Fokus nicht auf moderne sondern historische Texte verlegt. Dies bringt einige Herausforderungen mit sich. Zwar ist es einfach, einen Spruch, wie *Geld stinkt nicht*, über die Jahrhunderte aufgrund seiner syntaktischen Festigkeit zu verfolgen, jedoch ist es viel häufiger der Fall, dass sich durch semantische Verschiebungen eines Wortes ein *Text Re-use* verändert. Das macht eine *Text Re-use Analysis* auf historischen Texten besonders herausfordernd. Dies wird durch eine größere Vielfalt an Schreibweisen sowohl zur gleichen Zeit im regionalen Kontext aber auch durch zeitabhängige Schreibvarianten, verschiedene Dialekte und Sprachevolution im Laufe der Jahrhunderte verstärkt. Alles in allem kann eine *Text Re-use Analysis* auf historischen Texten aufgrund ihrer großen Datenvolatilität als nicht statisch angesehen werden, weshalb im Rahmen dieser Arbeit auch *Historical Text Re-use* als Terminologie benutzt wird.

Historical Text Re-use kann vielschichtig betrachtet werden. Einerseits wurden in den letzten Jahrhunderten von Historikern und Philologen akribisch Belegstellen für *Text Re-use* in Form von entsprechenden Wörterbüchern gesammelt (vgl. Abb. 1.5 auf Seite 41), welche jedoch heutzutage oftmals nicht digital zur Verfügung stehen. Vielmehr kann das damit verbundene geisteswissenschaftliche Wissen als fachwissenschaftliche Erwartungen von Zitierabhängigkeiten genutzt werden. Andererseits stellt sich für die Informatik die Frage, wie robust entsprechende Techniken des *Text Re-use Mining* gegenüber der aufgezeigten Datenvolatilität sind. Insofern kann sowohl wegen der Datenvielfalt als auch einem über die Jahrhunderte vielfach untersuchten Textkanon der *Historical Text Re-use* auf historischen Dokumenten als der am höchsten anzusetzende *Gold Standard* in der gesamten Forschung zum *Text Re-use* aufgefasst werden.

Historical Text Re-use kann daher als ein Wissenschaftsbereich der in den letzten Jahren aufkommenden *eHumanities* angesehen werden. Die *eHumanities* resultieren aus drei verschiedenen Einflüssen. Einerseits ist ein Einfluss durch die Geisteswissenschaften mit der entsprechenden Wissenschaftstradition gegeben. Die *Digital Humanities* haben im Laufe der letzten zwei Jahrzehnte massiv historische Daten digitalisiert und diese entweder per *Open Access* aber zumindest über ein Portal bereitgestellt. Der letzte und jüngste Einfluss kommt aus der Informatik. Insbesondere der Bereich der *Automatischen Sprachverarbeitung* ist im Umgang mit historischen Texten am naheliegendsten. Für den Bereich des *Historical Text Re-use* bedeutet das im Speziellen, dass die Geisteswissenschaften mit dem bereits gesammelten Wissen als der *Gold Standard* bzgl. der Evaluierung angesehen werden kann, die *Digital Humanities* entsprechend die Texte bereitstellen und die Informatik im einfachsten Fall Techniken aus der Plagiarismusforschung adaptiert. Hierbei sind die entsprechenden Techniken nicht mehr Gegenstand, um Rechtsfragen über Urheberschaften zu klären, sondern ein nützliches Instrument, welches nicht nur Evidenzen von Transferwegen, sondern vielmehr auch einen fundamentalen Teil des sprachlich-kulturellen Erbes der Menschheit darstellt. Vielmehr wird durch die Rolle der *Digital Humanities* deutlich, dass in der Chronologie der *Historical Text Re-use* in Form von automatischen Methoden erst in der Gegenwart eine entsprechende Aufmerksamkeit entgegengebracht werden kann, da die Texte erst seit den letzten Jahren in einem hinreichend großen Umfang vorliegen. Ziel

und Gegenstand dieser Arbeit ist es, den Forschungsbereich des *Historical Text Re-use* zu definieren und zu strukturieren. Allem voran steht hierbei sowohl der Aspekt der genannten *Data Diversity* sowie der damit verbundenen und deutlich verkomplizierten Evaluierung von Ergebnissen einer *Text Re-use Analysis*.

Dieses einführende Kapitel soll dazu dienen, verschiedene Aspekte des *Re-use* zu definieren. Hierzu wird vom Allgemeinen hin zum Spezifischen sukzessive spezialisiert und der Fokus von *Re-use* im Allgemeinen auf den *Historical Text Re-use* im Speziellen gelegt.

Im Detail gibt dieses Kapitel einführend mit dem Abschnitt *Re-use in der Natur* (vgl. Abschnitt 1.2) Einblicke, dass unser gesamtes tägliches Leben von *Re-use* geprägt ist. Dies können *DNA*-Daten aber genauso auch architektonisches Wissen beim Hausbau sein. Ziel dieses Abschnittes ist es, dafür zu sensibilisieren, dass *Re-use* immer aus Gemeinsamkeit und Unterschied besteht. Einerseits können nur beispielhaft Bäume eines Typs in eine gemeinsame Klasse, der Baumart, klassifiziert werden. Andererseits gleicht kein Baum einer Art einem anderen. Hierdurch ergeben sich somit grundlegende Eigenschaften bzw. Anforderungen an die noch einzuführenden Konzepte der *Minutiae* und des *Re-use Nucleus* (beide vgl. Definitionen 16 und 17 auf Seite 112).

Der Abschnitt 1.3 benennt diverse Spannungsverhältnisse im Umfeld des *Text Re-use*. Diese umfassen nicht nur die Wechselwirkung zwischen der Informatik mit den Geisteswissenschaften, sondern insbesondere auch Wechselwirkungen innerhalb der Informatik, wie die mit der Plagiarismusforschung oder auch dem *Sequence Alignment* von *DNA*-Daten.

Abschnitt 1.4 stellt vier verschiedene Visualisierungen des *Historical Text Re-use* im Kontext von *Information Overload vs. Information Poverty* vor. *Information Overload* wird hierbei oftmals dadurch erzeugt, dass einem Nutzer zu viele Informationen angezeigt werden, was aus Sicht der Wahrnehmungspsychologie nachteilig ist. Andererseits muss die Frage gestellt werden, was eine Visualisierung im historischen Kontext, zu welchem der Verlust von zahlreichen Texten in der Vergangenheit gehört und somit lediglich eine Submenge aller Daten vorliegt, leisten kann. In Anbetracht der eben aufgezeigten *Information Poverty* liegt so die Gefahr einer Fehlinterpretation von Daten durch eine Visualisierung zugrunde. Dieser Abschnitt stellt vier Visualisierungen des *Historical Text Re-use* vor und diskutiert sie im Kontext des *Information Overload* und der *Information Poverty*.

In Abschnitt 1.5 wird das im Rahmen dieser Arbeit entstandene *ACID for the eHumanities* Paradigma vorgestellt. Das Paradigma enthält vier einfache Fragen nach *Acceptance*, *Complexity*, *Interoperability* und *Diversity*, welche bereits ausreichen, um ein Projekt oder eine Dissertation in den *eHumanities* erfolgreich zu gestalten. Sowohl die *Acceptance* durch die Fachwissenschaftler als auch die *Diversity* stellen besondere Herausforderungen dar. Vielmehr ist die *Diversity* von historischen Texten und dem unterschiedlichen und autorspezifischen Umgang mit *Text Re-use* eines der Kernprobleme dieser Arbeit. Dem schließt sich der Abschnitt 1.6 mit relevanten Herausforderungen aus dem Paradigma an.

Die wissenschaftliche Einbettung des *Historical Text Re-use* in das *Noisy Channel Theorem* von *Shannon* wird in Abschnitt 1.7 motiviert. Hierbei wird ein zitierter und ein zitierender Autor als *Source* und *Target* des *Noisy Channel Model* verstanden. Die bereits eingeführte *Data Diversity* wird durch den *Noisy Channel* repräsentiert. Einerseits kann so der *Noisy Channel* auf systematische Veränderungen, wie durch kopierende Mönche im Mittelalter oder eines Editors im Allgemeinen, aber auch die fachwissenschaftliche Textkritik modelliert werden. Andererseits dient das *Noisy Channel Model* auch dazu, ein künstliches Störsignal, das Rauschen, in den *Noisy Channel* einzuführen. Dabei wird die Fähigkeit einer *Text Re-use Analysis* sowohl auf einer *Digital Library* als auch durch ein Störsignal erzeugte *Randomised Digital Library* verglichen.

Abgeschlossen wird dieses Kapitel mit der Abgrenzung verwandter Themengebiete (vgl. Abschnitt 1.8) und dem Ausblick bzw. der Gliederung dieser Arbeit (vgl. Abschnitt 1.9).

1.2 Re-use in der Natur

Re-use ist in der Natur nicht nur vielfach beobachtbar, sondern er stellt auch eine elementare Komponente, gar eine Naturkonstante, dar. Obgleich *Re-use* in der Fauna, Flora, Natur oder auch von Menschenhand geschaffen beobachtet werden kann, geht es letztlich immer darum, erfolgreich Gelerntes und Erprobtes weiterzugeben. Im Grunde kann *Re-use* im Allgemeinen als ein *genetischer Algorithmus* verstanden werden, bei dem Erfolgreiches mit Erfolgreichem gekreuzt und dabei weniger Erfolgreiches oder gar Unbrauchbares entfernt wird. Was ist jedoch Erfolg, so dass eine Information wiederverwendet wird? Warum sind im Laufe der Evolution bestimmte Informationen gegen Einflüsse resistenter geworden als andere? Dies sind keinesfalls einfache und lineare Antworten bzw. Entscheidungen, da es sich beim *Re-use* implizit immer um ein *dynamisches* und *zeitabhängiges*, also ein *evolutionäres*, und *kein stationäres System* handelt. *Re-use* kann als ein fortbeständiger Prozess der Weitergabe von Wissen bzw. erfolgreich Gelerntem verstanden werden, der bedingt durch Anpassungen auf *äußere Einflüsse* im Sinne von *neue Anforderungen* das Wissen einem sukzessiven *evolutionären Wandel* aussetzt.

Ein kurzes Beispiel aus der Genetik soll diesen Prozess zu unterschiedlichen Zeitpunkten mit jeweils sehr konträren Anforderungen an die Leistungsfähigkeit eines Menschen verbildlichen. Ist ein gesundes Baby genau dann "erfolgreich", wenn es im Sinne eines *DNA-Re-use* keine Erbkrankheiten besitzt? In den meisten Fällen würde man diese Frage wahrscheinlich mit "ja" beantworten. Für eine gegenteilige Ansicht kann sich auf das Mittelalter berufen werden, als in Europa durch die Pest viele Menschen gestorben sind. Die gesunden und bis dahin genetisch sehr "erfolgreichen" Menschen sind sehr zahlreich gestorben, weil sie sich der Pest auf natürliche Weise nicht erwehren konnten. Eine signifikante Anzahl von Menschen hingegen haben die Pest überstanden. Kennzeichnend für die Überlebenden ist ein genetischer Defekt, der verursacht hat, dass diese Menschen dickflüssigere Schleimhäute besitzen, wodurch die Pest besser eingeschlossen und isoliert werden konnte[1]. Genau dieser kleine genetische Defekt hat sich zumindest zu dieser Zeit mit den speziellen Anforderungen als besonders erfolgreich herausgestellt. Wird heutzutage jedoch ein Kind von zwei Elternteilen gezeugt, welche beide diesen genetischen Defekt besitzen, so ist es sehr wahrscheinlich, dass das Kind eine sehr schwere Erbkrankheit haben wird, welche bedingt durch die eben genannte Vorgeschichte überdurchschnittlich oft im europäischen Raum auftritt.

Dieses Bewusstsein eines *Re-use* als *diachrones*, *dynamisches* und *evolutionäres* System zeigt bereits die Komplexität dieser Aufgabe für die Informatik auf. Bezugnehmend auf das eben genannte Pest-Beispiel ist es nicht allzu schwer, sich zu überlegen, welcher Aufwand dahinter steht, einen solchen Zusammenhang zwischen einem speziellen Gendefekt und einer der schlimmsten Epidemien in der Geschichte der Menschheit nachzuweisen. Vielmehr stellt sich neben solchen Einzelbeispielen die Frage, was zum Beispiel alles im Sinne einer *evolutionären Anthropologie* noch nicht bekannt ist.

In der *evolutionären Anthropologie* werden unter anderem Abstammungskarten erstellt. So wurde im Rahmen einer Ausgrabung 2008 in der Denisova-Höhle im Altai-Gebirge ein Fingerknochen gefunden[2]. Mit diesem Fingerknochen konnte zweifelsfrei nachgewiesen werden, dass der Homo Sapiens und der Neandertaler nicht nur zur gleichen Zeit gelebt haben, sondern auch in teilweise gleichen Gebieten angesiedelt waren und gemeinsame Nachkommen hatten. Durch einen einzigen Fingerknochen kann so das Wissen in Abstammungskarten sukzessive präzisiert werden. Der Punkt ist, dass die diachrone Eigenschaft des *Re-use* in den meisten Fachdisziplinen, speziell denen mit einem großen zeitlichen Fenster, oft-

[1] Diese Information geht auf ein persönliches Gespräch mit dem Humangenetiker Dr. med. Friedmar R. Kreuz aus Dresden zurück.

[2] vgl. `http://www.3sat.de/page/?source=/nano/natwiss/144355/index.html`

mals nicht auf alle Daten zugreifen kann, sondern nur auf einen relativ kleinen und überschaubaren Bestand an Informationen, die eher an den Strand geschwemmte Muscheln und Meeresreste erinnern, als an die bunte Farbvielfalt des Ozeans.

Eine möglichst starke Verbreitung von Wissen zur gleichen Zeit (vertikale Ausrichtung) sowie über einen möglichst großen Zeitraum (horizontale Ausrichtung) erlaubt eine gute Wahrnehmung sowohl zum aktuellen Zeitpunkt als auch in den meisten Fällen in der Zukunft. So gibt es beispielsweise heute keine Dinosaurier als *Primärquellen* mehr. Jedoch zeugen *fragmentarische Funde* von deren ehemaligen Existenz. Fragmente sind Überlieferungen aus vergangenen Zeiten, die entweder mutwillig oder auf natürliche Weise zerstört worden sind. Je größer der Verbreitungsgrad dieses genetischen oder anderen Wissens war, desto wahrscheinlich ist es, dass heutzutage solche Fragmente noch gefunden werden können. Ferner kann in der heutigen Zeit wahrscheinlich nur deswegen auf solche Fragmente zurückgegriffen werden, weil das entsprechende *Wissen zahlreich wiederverwendet* worden ist. Hierbei erhöht sich die *Vertrauenswürdigkeit*, um so mehr fragmentarisches Wissen vorliegt, welches Gleiches beschreibt. Entgegen modernen Diskussionen über unzulässigen Plagiarismus ist es im historischen bzw. anthropologischen Kontext wichtig, dass Texte bzw. jegliche andere Vervielfältigung möglichst oft kopiert worden sind.

Neben der eben erwähnten diachronen Eigenschaft des *Re-use* ist die *Dynamik* von Wissen die wohl schwierigste aller Eigenschaften. Auch diese Eigenschaft ist in der Natur zahlreich beobachtbar. Der Leser sei gebeten beim nächsten Waldspaziergang oder dem Fahren auf einer Landstraße, an welcher typischerweise Obstbäume gepflanzt sind, einmal zwei benachbarte Bäume der gleichen Art zu vergleichen. In den meisten Fällen wird feststellbar sein, dass trotz der gleichen *äußeren Einflüsse* wie Sonneneinwirkung, Regen und auch Beschneidungen durch den Menschen weitere artgleiche Bäume in den seltensten Fällen auch formgleich sind. Bäume, die am Rand eines Waldes stehen, sind beispielsweise kleiner und an der Wetterseite stärker "abgenutzt". Bäume in der Mitte eines Waldes sind aufgrund des Schutzes im Wald meist eher größer und die Äste beginnen näher zur Krone hin anzuwachsen. Sie werden beim näheren Betrachten immer unähnlicher. Der eine Baum hat eher eine spitze, der andere eine runde Silhouette. Der eine Baum wirkt "leer", der andere ist voll mit Blättern. Je länger der Beobachter hinschaut, umso mehr Unterschiede können festgestellt werden. Eine Analogie hierzu kann sich auch zur Tierwelt überlegt werden. Sind zwei Tiere der gleichen Art auch identisch in Form und Aussehen? Genau diese Eigenschaft der Dynamik, nicht nur in der Zeit, sondern auch die Vielfalt der parallelen Instanzen zur gleichen Zeit, erschwert das Erkennen des *Re-use* enorm. Die Frage, die sich unmittelbar aus der Dynamik des *Re-use* in Form und Aussehen ergibt, ist die Frage nach einer geeigneten *Signatur* (siehe Kapitel 2.5) zur Beschreibung dieses Wissens, die robust genug ist, um einen hinreichend großen *Adaptionsspielraum* für Wissen durch *äußere Einflüsse* behandeln zu können. Damit geht die Frage einher: Was sind die entsprechenden *Minutiae*[3] (vgl. Kapitel 2.5)? Es sei einmal angenommen, ein Biologe würde Bäume nur nach Aussehen, Form, Größe und ggf. vielleicht Früchten klassifizieren, was wäre dann bspw. ein junger Birnbaum, der keines der eben genannten Kriterien für ein ausgewachsenes Exemplare erfüllt? Dementsprechend existiert immer ein *Adaptionsspielraum* für Wissen, der durch *äußere Einflüsse* gesteuert werden kann und so die *natürliche Wechselwirkung* aus *Homogenität* und *Unterschied* abbildet.

Ein weiterer Aspekt dieses Gedankens ist die Sichtweise auf den *Re-use*. Die Dynamik-Eigenschaft des *Re-use* lässt Gleiches nicht immer gleich aussehen. Die Kernfrage ist jedoch,

[3] *Minutiae* [BIMA 2012, NSTC 2006] ist ein Begriff aus der *Biometrie*, welcher die Hauptkomponenten einer *biometrischen Signatur*, wie einem Fingerabdruck, definiert. Beim Fingerabdruck gibt es nach [Jain 2005] genau sieben Hauptkomponenten des menschlichen Fingerabdruckes (vgl. Abb. 3.4 auf Seite 99 in [Jain 2005]).

wie damit umgegangen werden muss? In einer eher typorientierten Makrosicht, in welcher die einzelnen Instanzen über die Sicherheit einer atomaren Eigenschaft einer *Minutiae* etwas aussagen, kann bspw. etwas über eine Eigenschaft dieses Typs (biologische Gattung bzw. Klassifikation) im Vergleich zu anderen Typen und deren Ähnlichkeit festgestellt werden. Hierbei ist die Dynamik eine sehr große Herausforderung, da eine Signatur eines Typs gefunden werden muss, der diese Vielfalt toleriert, ohne dabei ungenau zu werden. Dem steht eine *Instanz*-orientierte Mikrosicht entgegen, welche nicht mehr auf die Gemeinsamkeit aller Instanzen einer Art, sondern deren Abhängigkeiten, wie Vererbungen durch Vaterschaften, vergleicht. In der instanzbasierten Mikrosicht zählt somit nicht mehr die Gemeinsamkeit der Art, sondern etwaige Gemeinsamkeiten in der Dynamik wie zum Beispiel gemeinsam auftretende Auffälligkeiten bei genetischen Abweichungen in der Fauna oder geographische bzw. klimatische Unterschiede in der Flora.

Nicht nur die Natur, sondern auch der Mensch folgt einem Wiederverwendungsbestreben. So baut der Mensch Häuser auf der Basis von jahrhundertelangem Wissen über Architektur. Genau wie im vorherigen Beispiel mit den Bäumen, gleicht kein Haus dem anderen. Diese Individualisierung ist einerseits zur persönlichen Abgrenzung wichtig. Auf der anderen Seite folgen selbst solche individuellen Modifizierungen immer klaren architektonischen Regeln. Würden solche Regeln nicht beachtet werden, so könnte nicht garantiert werden, dass ein Haus einem Sturm standhalten könnte. Das gelernte architektonische Wissen leistet nicht nur Sicherheit Gewähr, sondern verfolgt in erster Linie auch *ökonomische Aspekte*, um die Kosten für einen solchen Bau zu reduzieren. Jede Form eines *Re-use*, sei es aus architektonischen, genetischen aber auch softwaretechnischen[4] bzw. textbasierten[5] Gründen, hat immer die Minimierung der Kostenfunktion zur Ursache.

Auch Menschen und Tiere lernen am besten durch das ständige Wiederverwenden von Wissen. So lernt ein Kleinkind Gegenstände und Personen zu benennen, indem einem Kind möglichst oft ein Gegenstand gezeigt wird und damit die Aussprache eines bestimmten Wortes korreliert. Ähnliches konnte der russische Wissenschaftler Pawlov in seinem Experiment mit Hunden ebenfalls zeigen[6]. Hierbei wurde ein Hund zur Fütterung immer einem akustischen Signal ausgesetzt. Nach geraumer Zeit hatte der Hund gelernt, dass es mit diesem Signal stets Futter gibt. Aufgrund dessen fing der Hund an, bereits beim Hören dieses Signals Speichel abzusondern, auch wenn es kein Futter gab. Im Gegensatz zum bereits erwähnten evolutionären Wandel von Wissen ist beim Erlernen dieses Wissens *Stabilität* wichtig und steht somit im Kontrast zur Evolution. In den beiden hier nur exemplarisch aufgeführten Beispielen würde weder das Kleinkind noch der Hund etwas lernen, wenn dem Kind entweder jedes Mal ein anderes Wort zu einem gezeigten Objekt oder dem Hund ein anderes Signal gegeben werden würde. Potentielles Wissen, das bereits beim Erlernen einer starken Volatilität ausgesetzt ist, wird schwer zu erlernen bzw. zu erfassen sein. Daher ist es im Kontrast zur historischen bzw. anthropologischen Evolution, die einen Wandel zulässt, wichtig, dass während der *Generierung des Wissens* eine *temporäre Stabilität* vorliegt.

Im Gegensatz zu strukturiertem Wissen, wie der Assoziation zwischen Futter und einer Klingel, liegt das meiste Wissen in unstrukturierter Form vor. Neben den bereits erwähnten Genomen sind auch Bilder, Musiknoten aber auch natürlichsprachliche Texte unstrukturiert. Dies heißt jedoch nicht, dass solche Daten zufällig auftreten bzw. generiert werden, sondern dass das strukturierte Wissen in einer linearen Sequenz abgespeichert wurde. Dies können bspw. Genomsequenzen aber auch Texte sein. Entgegen einer zufälligen Folge von Elementen eines Vokabulars W folgen solche linearen Sequenzen immer den Gesetzmäßigkeiten einer Sprache S.

[4] bspw. modulare Programmierung, um Code-Redundanzen zu vermeiden
[5] bspw. Plagiarismus oder Single-Sourcing
[6] vgl. `http://de.wikipedia.org/wiki/Pawlowscher_Hund`

Anhand dieser sehr wenigen Beispiele sollen dem Leser einführend die wichtigsten Eigenschaften des *Re-use* nahe gelegt werden. Weiterhin sind einige Analogien zu parallelen Wissenschaften, wie der Anthropologie, Biologie oder auch Soziologie, aufgezeigt worden. Beim Einstieg in das Thema lässt sich zu Beginn nur vermuten, dass es etwas wie einen *universellen Re-use* bzw. dass es eine Art *universellen Fingerprint*, quasi das *Re-use Pattern*, geben muss, der für den Erfolg einer Information sowohl im *Evolutionären Prozess* als auch der nötigen *Stabilität* steht.

1.3 Wissenstransfer im interdisziplinären Spannungsfeld

Auch wenn die einführenden Beispiele im ersten Moment dem Leser nicht relevant für diese Arbeit zu sein scheinen, so ist eigentlich genau das Gegenteil der Fall, da bereits einige der grundlegendsten Eigenschaften aufgezeigt werden. Neben den genannten Eigenschaften, wie das Arbeiten mit *Fragmenten*, dem *Wirkungsgrad*, dem *Unterschied von Gleichem*, der *angestrebten Effizienz* durch *Re-use*, dem Verhältnis aus *Funktion und Schönheit* oder der *Namensgebung*, beschreiben all diese Beispiele bei näherem Betrachten die wichtigste aller Eigenschaften: die *Interdisziplinarität*. Analysen zum Wissenstransfer in verschiedensten Formen sind meistens von interdisziplinärem Charakter geprägt.

Re-use zu messen, ist ein auf verschiedensten Medientypen, wie Bildern, Musik, DNA-Daten, Fingerabdrücken und auch Texten, relevantes Thema. Neben den jeweiligen Fachwissenschaften, wie der Anthropologie, der Medizin oder der Biologie, stehen diesen Forschungsbereichen qualifizierte Methoden der Massendatenanalyse aus der Informatik zur Verfügung. Der Leser sei gebeten, sich einmal kurz vorzustellen, dass in der Anthropologie eine Radio-Karbon-Analyse zur Bestimmung des Alters eines Knochens (vgl. Fingerknochen-Beispiel aus Kapitel 1.2) oder eine DNA-Analyse Ergebnisse mit einer Genauigkeit von nur 90% liefern würden. Auch wenn aus Zeitgründen das an dieser Stelle nicht näher ausgeführt werden kann, so sollte klar sein, dass solche Methoden gemeinhin nicht als akzeptiert angesehen werden, wenn sie eine derart schlechte Genauigkeit haben.

Aus diesem spezifischen Beispiel in der Interaktion zwischen verschiedenen Wissenschaftsbereichen können zwei Dimensionen des *interdisziplinären Spannungsfeldes* verallgemeinert werden.

- **Horizontales Spannungsverhältnis** (zwischen Disziplinen der Informatik): Entwickelte Methoden in unterschiedlichen Forschungsbereichen, wie der Biometrie, der DNA-Analyse oder der Plagiarismuserkennung, können oftmals nicht oder nur sehr schwer adaptiert werden, auch wenn die Methodik - hier der Wiederverwendung einer entsprechenden *Signatur* - sehr ähnlich scheinen. In den meisten Fällen scheitern solche Bemühungen schon bei der Beschreibbarkeit der entsprechenden Daten.

- **Vertikales Spannungsverhältnis** (zwischen der Informatik und den Fachdisziplinen): Sie entsteht dadurch, dass sich in den meisten Publikationen eine teilweise enorme Diskrepanz zwischen *fachwissenschaftliche Erwartung* bzw. *Notwendigkeit* und der *realen Leistungsfähigkeit* einer Methode auftut. Der Aufwand von einem technischen *State of the Art*, für welchen bspw. eine wissenschaftliche Publikation erstellt worden ist, hin zu einer anwendbaren Methodik zu kommen, wird insbesondere von der Informatik oftmals deutlichst unterschätzt.

Während dem *Horizontalen Spannungsverhältnis* meistens algorithmische *Harmonisierungsbemühungen* bzw. *Wiederverwendungen* zugrunde liegen, ist es beim *Vertikalen Span-*

nungsverhältnis oftmals genau die gegenteilige *fachspezifische Diversität*, auf die die Informatik nicht entsprechend reagieren kann.

So scheinen bspw. DNA-Re-use-Analysen und Plagiarismuserkennung im Sinne des *horizontalen Spannungsverhältnisses* sehr nah beieinander zu sein, jedoch ist auch offensichtlich, dass die zu analysierenden Sprachen kaum gegensätzlicher sein könnten. Während in der Plagiarismuserkennung auf Wortebene mit entsprechender Satzsyntax bzw. Morphologie zur Lemmatisierung gearbeitet werden kann, wird bei der DNA-Analyse mit einem kleinen Vokabular $V = (A, C, G, T)$ von nur vier Repräsentanten gearbeitet. Diese teilweise völlig unterschiedlichen Charakteristika und damit verbundenen Verteilungen bringen ganz andere Anforderungen zu Tage. Während bei Textdaten meist ein *Data Sparseness* sowie *viele seltene Ereignisse* (vgl. Kapitel 4) angenommen werden können, sind DNA-Daten wesentlich dichter und weniger *sparse*. Als unmittelbare Konsequenz dessen kann bspw. auf DNA-Daten die *Mutual Information* als statistisches Signifikanzmaß eingesetzt werden, welches aufgrund der vielen seltenen Ereignisse in Texten keine brauchbaren Ergebnisse liefert.

Das *vertikale Spannungsverhältnis* hingegen entsteht oftmals aus dem resultierenden Konflikt der digital zur Verfügung stehenden Daten, die teilweise exponentiell steigen, und der fehlenden Fähigkeit seitens der jeweiligen Fachwissenschaften, auf diese Menge von Daten mit traditionellen Methoden zu reagieren, genauso wie dass Modelle der Informatik oftmals nur die offensichtlichen Strukturen extrahieren (vgl. auch das Paradigma *ACID for the eHumanities* aus Abschnitt 1.5). So ist es ein Leichtes, Shakespeare's *to be, or not to be, that is the question* als *Text Re-use* mit Techniken des *Sequence Alignment* zu identifizieren. Jedoch ist dies für die Fachwissenschaften nicht von allzu großem Interesse, da dieses bereits ein bekanntes Zitat und auch die Frage nach dem *Archetyp* bereits geklärt ist.

Um befruchtende Forschung im *horizontalen* als auch *vertikalen Spannungsverhältnis* machen zu können, darf die Interdisziplinarität nicht als die Zusammenarbeit zweier Disziplinen verstanden werden, sondern als ein zusammenwachsender neuer Forschungsbereich. Als Grundlagen hierfür wurden spezielle Datenbanken und Verfahren für DNA- und Protein-Daten entwickelt. Die Bioinformatik ist hierbei sehr paradigmatisch für andere Disziplinen wie auch den *eHumanities*. Nach nunmehr etwa 15 Jahren aktiver Forschung im Bereich der Bioinformatik fanden Forscher erst unlängst diejenige 10 Zeichen lange Gensequenz, welche Epilepsie verursacht. Das Ziel für die Zukunft, und damit auch Auftrag für diese Arbeit, muss es sein, eben jene und insbesondere *vertikalen* Spannungsverhältnisse im Bereich der *eHumanities* aufzulösen. Wobei das Zusammenwachsen zweier Disziplinen nicht nur Auftrag, sondern auch Herausforderung ist. Letztlich arbeiten die Geisteswissenschaften und die Informatik methodisch diametral anders. Während die textorientierte Informatik mit Sprachmodellen arbeitet, welche meist die offensichtlichen Strukturen extrahieren, forschen die Geisteswissenschaften eher nahe am Text mit einzelnen Belegstellen. Letztlich haben beide Vorgehensweisen ihre Vor- und Nachteile. Wichtig hierbei ist jedoch, darauf hinzuweisen, dass durch die Digitalisierung die Menge an Text sukzessive steigt, so dass die traditionelle Arbeit mehr und mehr erschwert wird. Prof. Dr. Gregory Crane fragte in diesem Kontext 2006 danach *What can we do with a million books?* (vgl. [Crane 2006]).

1.4 Information Overload vs. Information Poverty

Mit dem Aufkommen des Computers und der damit verbundenen Massendatenhaltung sind wir heutzutage in einer *Digitalen Welt* zu Hause, in welcher wir tagtäglich vielen Informationen ausgesetzt sind. Auch in der Wissenschaft führen die Erfolge der Digitalisierung zunehmends zu signifikant größer werdenden Datenbeständen, was traditionelle fachwissenschaftliche Methode sukzessive weniger effektiv macht (vgl. Kapitel 1.3).

Dieser Abschnitt wird sich auf den Bereich des *Re-use* in den textorientierten *eHumanities* fokussieren. Die Massendigitalisierungen der letzten Jahre eröffnen neue Möglichkeiten, Algorithmen des *Historical Text Re-use Detection*[7] zu entwickeln, um so zumindest für den Bereich der Zitationsspuren, Paraphrasen und Allusionen, Texte mit Hypertext-Links zu versehen und damit ähnlich dem *World Wide Web* diese über Dokumente hinweg navigierbar dem Nutzer bereitzustellen.

Solche neuen Zugriffsformen sind nötig, um dem zunehmenden *Information Overload* entgegenzutreten. Der Leser stelle sich in einem Gedankenexperiment einmal folgendes vor: Ziel ist es, ein *Geflügeltes Wort* wie *gleich und gleich gesellt sich gern* auf seinen *Archetyp* bzw. sein frühestmögliches Auftreten zu untersuchen. Als Grundlage hierfür dienen alle digitalen Bücher in *Google Books*. Wie wahrscheinlich scheint es nun dem Leser, genau diese Information in einer großen digital zur Verfügung stehenden Bibliothek zu finden? Zweifelsohne ist es bereits sehr hilfreich, nicht nur nach einzelnen Wörtern des *Geflügelten Wortes* zu suchen, sondern nach dem gesamten *Geflügelten Wort* selbst. Dadurch wird die Treffermenge bereits deutlich reduziert. Jedoch bleibt sie dennoch aufgrund des enormen Datenbestandes unübersichtlich groß. Wenn nun ein Retrievalsystem keine chronologische Sortierung unterstützt, z. B. weniger aus technischen Gründen, sondern vielmehr, weil die Datierungsinformationen nicht vorhanden sind, dann wird trotz des Einsatzes des Computers die Suche nach dem *Archetyp* zu einer Suche nach der Nadel im Heuhaufen.

Dieses einführende Beispiel soll zeigen, dass für nahezu jedes *Mining*- und *Retrieval*-Verfahren, wie das Finden von *Text Re-use*, welches auf einen großen Datenbestand angewendet wird, ebenfalls eine große Menge von *Mining*-Ergebnissen berechnet werden kann. Allein das *Mining*-Verfahren hilft vielleicht, unbekannte Zitationen theoretisch aufzudecken, jedoch hilft es nicht dabei, dies systematisch zu ermöglichen. Im Grunde kommt der Nutzer, hier der Geisteswissenschaftler, "vom Regen in die Traufe", wodurch auch derzeit eine gewisse Abneigung solcher Methoden gegenüber zu spüren ist.

Neben dem *Information Overload*[8], dem sich ein Fachwissenschaftler zunehmend ausgesetzt sehen muss, herrscht zeitgleich auch eine *Information Poverty*. Historische Dokumente vor dem 10. Jahrhundert existieren heutzutage nahezu vollständig nur noch durch Abschriften. Je nachdem wie gut verbreitet die Abschriften waren, ist die Wahrscheinlichkeit gering, dass ein Text heutzutage noch erhalten ist. Was wir heute lediglich wissen, ist, dass sehr viele Texte in der Vergangenheit verloren gegangen sind (Details werden in Kapitel 1.6 erklärt). Wir schauen heute speziell auf die antiken Texte mit einer starken und nahezu willkürlichen Selektion speziell in den *Dark Ages*, in welchen viele Text zerstört wurden bzw. verloren gegangen sind. Im Grunde geben heutzutage die *Fragmentarischen Autoren* (vgl. [Berti 2009, Berti 2012]) eine ungefähre Mindestabschätzung dafür, auf welche Texte wir selbst mit Massendigitalisierungen nicht mehr zugreifen können.

Aus dem *Information Overload* und der *Information Poverty* lassen sich für die Informatik zwei fundamentale Anforderungen ableiten. Das *Information Overload* zeigt uns, dass Zugriffsformen benötigt werden, die die große Menge an Daten in geeigneter Form ganz im Sinne von Keim's Mantra *Overview first, zoom and filter, details-on-demand* aggregieren bzw. selektieren ([Keim 2002]).

Die *Information Poverty* lehrt uns zeitgleich, dass nicht jede Form des Minings bzw. der Visualisierung auch befruchtend für die Fachwissenschaften sein muss. Im eAQUA-

[7] *Historical Text Re-use* wird im Rahmen dieser Arbeit auch *Text Re-use* genannt, da sich die Arbeit ausschließlich mit historischen Texten beschäftigt.

[8] Die Grundidee dieses Kapitel geht auf die Präsentation *Information and books - growth and the circle of (de-)construction* von Reinhard Förtsch zurück, welche im Nomvember 2011 am Deutschen Archäologischen Institutes in Rom, Italien, zur Veranstaltung *Rezeption der Antike im Semantischen Netz* vorgetragen worden ist.

Projekt [Büchler 2008c][9] konnte so gelernt werden, dass sich insbesondere auf den antiken Texten diachrone Analysen bzw. eine Visualisierungen als wenig vorteilhaft für die *Humanities* herausgestellt haben. Dies liegt einfach im *Text Sample* begründet. Aufgrund der Überlieferungsgeschichte von Texten können wir keine seriöse Aussage über die Repräsentativität von heute noch erhaltenen antiken Daten machen. So führt bspw. eine diachrone Semantikanalyse, um Bedeutungsveränderungen zu bestimmen, dazu, dass nicht sicher gesagt werden kann, ob ein Unterschied wirklich aus einer solchen semantischen Verschiebung resultiert oder einfach nur bedingt durch ein nicht mehr nachvollziehbares *Sampling Bias* hervorgerufen wird. Es kann kurzum nichts zur Repräsentativität eines Korpus zum jeweiligen Zeitpunkt ausgesagt werden, so dass immer der Zweifel am Ergebnis bestehen bleibt und somit auch keine geisteswissenschaftliche bzw. hermeneutische Wissenschaft ermöglicht wird.

Im Kontext der *eHumanities* muss die Informatik verstehen, dass ein *Mining*-Ansatz bzw. eine Visualisierung einer genauen fachwissenschaftlichen Prüfung unterliegt. Hierbei steht die Frage im Vordergrund, ob damit wirklich Forschung in den Geisteswissenschaften ermöglicht werden kann oder ob im Sinne einer *Hammel-Nagel*-Methode seitens der IT ein Algorithmus bzw. eine Visualisierung nur auf Text angewendet werden soll.

Ein hervorragendes Beispiel im Rahmen des *Historical Text Re-use* stellt hierbei das Göttinger Teilprojekt des *eTRACES*-Projektes[10] dar. Ziel des Teilprojektes ist es, nah am Text das Erstellen von Kommentaren in Online-Editionen für *Zitate*, *Paraphrasen* oder *Allusionen* zu ermöglichen.

Aus dem Göttinger *eTRACES*-Teilprojekt können wir im Zusammenspiel mit Keim's Mantra lernen, dass es für jede geisteswissenschaftliche Problemstellung eine spezifische Nutzeroberfläche oder Visualisierung geben muss, die die Massendaten auf eine für den Menschen handhabbare Menge reduziert.

Für den *Historical Text Re-use* gibt es zahlreiche Fragen, wie die nach dem eingangs bereits erwähnten *Archetyp*, welche jeweils eine andere Nutzer-Oberfläche benötigen. Nachfolgend sollen einmal vier verschiedene geisteswissenschaftliche Fragestellungen betrachtet werden und wie eine entsprechende Visualisierung aussehen könnte. Hierbei werden die vier Visualisierungen von Moretti's *Close* zum *Distant Reading* sortiert (vgl. [Moretti 2005]).

Ein Hauptforschungsgebiet in den Geisteswissenschaften für den Einsatz von Daten des *Text Re-use* ist die *Textkritik* (vgl. [Maas 1960] und [Dover 1997] Seiten 45-85). Wesentlicher Bestandteil der Textkritik ist die Frage nach der Überlieferungsgeschichte eines Textes und wie sicher jede Textpassage im Original genauso geschrieben wurde. Zitationsspuren helfen hierbei nach Zeugen zu suchen, die den entsprechenden Text, ggf. auch noch das Original, lesen konnten. Je weniger Änderungen eine Textpassage trotz zahlreicher Zitate besitzt, desto wahrscheinlicher ist es, dass der Text sehr originalgetreu überliefert wurde. Es sei nun wieder auf das einführende Gedankenexperiment mit *Google Books* verwiesen. Je mehr gleiche bzw. sehr ähnliche Textpassage gefunden werden, desto schwieriger ist es, manuell genau diese Textstelle auf Unterschiede zu untersuchen. Abbildung 1.1 zeigt eine entsprechende Visualisierung, welche gleiche Textpassage übereinanderlegt und nur die Unterschiede in Form eines *Variantenzweiges* darstellt (vgl. Details in [Büchler 2010d]).

Der *Information Overload* wird nicht nur auf eine Liste von wenigen Zitaten reduziert, sondern insbesondere auch auf deren Abweichungen. Der Teilsatz aus Abb. 1.1 enthält neun Wörter. In den fünf gefundenen Zitaten werden insgesamt nur zwei unterschiedliche bzw. eingefügte Wörter identifiziert. Der vollständige Satz zu diesem Beispiel beinhaltet 94 Wörter, zu welchen in den insgesamt fünf Zitaten nur vier Abweichungen aufgedeckt werden

[9] Siehe hierzu auch die *eAQUA*-Webseite unter [Heyer 2008] und den eAQUA-Abschlussbericht [Heyer 2011c]

[10] vgl. `http://etraces.e-humanities.net/project-partners/gcdh-etraces.html`

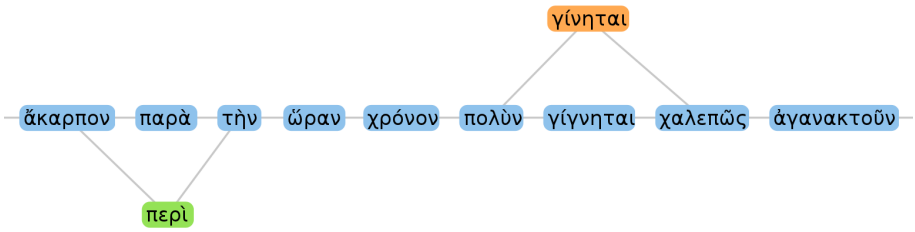


Abbildung 1.1: *Microview* einer Liste von Zitaten zur Unterstützung der geisteswissenschaftlichen Arbeit im Bereich der Textkritik - Platon Timaeus 91b7 ff.

können, was weniger als 1% Unterschied[11] ausmacht und somit ohne eine entsprechende Visualisierung deutlich schwerer zu identifizieren ist[12]. Die *Information Poverty* stellt hierbei die Grundmotivation der Textkritik dar.

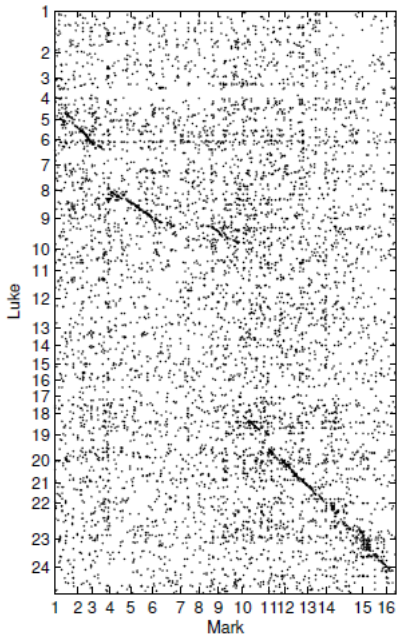


Abbildung 1.2: *Dotplotview* zum paarweisen Vergleich zweier Werke, um systematisches Abschreiben zu identifizieren - Vergleich der Bücher Lukas und Markus aus der Bibel (Bild-Quelle: [Lee 2007]).

Eine *Dotplotview* (vgl. Abb. 1.2) eignet sich für einen paarweisen Vergleich zweier Werke bzw. zweier Autoren. Hierbei ist von besonderem Interesse nicht nur zu wissen, dass es einen signifikanten *Re-use* gibt, sondern insbesondere, ob es eine systematische Abschrift gab. In den Humanities wird diese Methode bspw. in [Lee 2007] zum paarweisen Vergleich zweier

[11] 1% Unterschied bezieht sich hierbei auf die Anzahl der *Tokens*. Es werden vier Abweichungen in $5 \cdot 94$ *Tokens* beobachtet.

[12] Siehe hierzu auch `http://www.e-humanities.net/lectures/WS2010_1/SAMHT/2011-01-27-Halle-SGAHT_02-TextReuse-v1.pdf` Slides 39-42, die das Beispiel im Rahmen der Vorlesung *Softwaregestütztes Arbeiten mit historischen Texten* ausführlicher behandeln.

Bücher der Bibel, wie Markus und Lukas, herangezogen. Eine andere Anwendung findet sich in der Gnomologienforschung wieder (vgl. [Pietruschka 2012]). Gnomologien sind spezielle Spruchsammlungen. Werden hierbei mehrere Sprüche hintereinander kopiert, entstehen die in Abb. 1.2 dargestellten *diagonalen Muster*. Diese Muster unterstützen die Fachwissenschaften, zwei gnomologische Ordnungssysteme zu vergleichen und ggf. daraus Abstammungsbäume der Gnomologien abzuleiten.

Der *Information Overload* kann bei der *Dotplotview* auf zwei verschiedene Arten reduziert werden. Einerseits ist dies implizit gegeben, da der Fokus nur auf dem *Re-use* zwischen zwei Werken liegt. Andererseits kann der *Re-use* für diese Form des Vergleichs zweier Werke weiter reduziert werden, indem ein Rauschfilter all diejenigen Links beider Werke entfernt, die nicht Teil eines in Abb. 1.2 dargestellten *vertikalen Musters* sind. Die *Information Poverty* kann bei dieser Form der Darstellung vollständig ignoriert werden, da nur Werke miteinander verglichen werden können, die auch digital vorliegen. Diese wissenschaftliche Betrachtung ist unabhängig von der Frage, wie sich das *Sample* einer *Digital Library* zur *hypothetischen Grundgesamtheit* verhält.

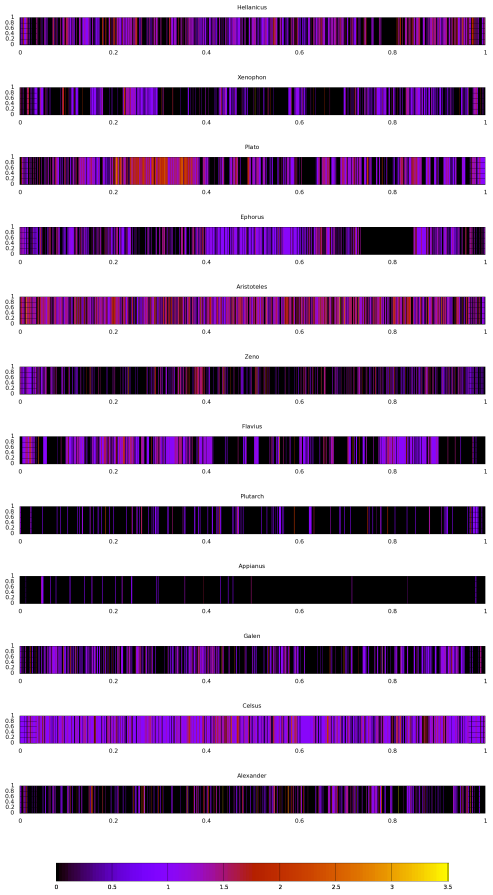


Abbildung 1.3: *Temperature Maps* eines Werkes bzgl. des *Text Re-use* in einer *Digital Library* zur Analyse, welche Teile eines Werkes besonders häufig referenziert bzw. wiederverwendet worden sind.

In den Geisteswissenschaften werden *Reading Lists* für die Studenten eines Griechisch-Kurses erstellt. Was sind gute Kriterien für die Auswahl der zu lesenden Textstellen? Oftmals ist es auch nur von Interesse, welche Textpassagen besonders häufig zitiert worden sind. Abb. 1.3 zeigt sogenannte *Text Re-use Temperature Maps* für 12 verschiedene Werke (vgl. [Büchler 2013c]). Hierbei wird auf der x-Achse das Werk nach Sätzen oder einer anderen *Segmentierungseinheit* zerlegt. Die Farbskala ist eine logarithmische Funktion[13] der *Re-use Frequency* dieser Textstelle (vgl. [Büchler 2013c]).

Der *Information Overload* wird bei einer *Temperature Map* Visualisierung auf Farben reduziert. Hierbei ist in einem ersten Schritt die einzelne Textstelle unbeachtet. Im Sinne von Moretti's *Distant Reading* liegt der Fokus darauf, welche Stellen eines Werkes besonders stark zitiert und welche Werke besonders breit wiederverwendet worden sind oder auch welche Werke keinen messbaren *Re-use* aufzeigen. Die *Information Poverty* ist hingegen stark von der Repräsentativität und Vollständigkeit einer *Digital Library* geprägt. Die Aussagekraft der Abb. 1.3 ist also sehr stark von den dahinterliegenden und insbesondere den zitierenden Texten beeinflusst. Der Leser stelle sich einmal in einem Gedankenexperiment vor, wie sehr das in Abb. 1.3 zugrunde liegende Korpus, nach verschiedenen Selektionsstrategien, wie zufällige Auswahl, Löschen von Werken aus bestimmten Epochen oder auch die Hinzunahme von weiteren Werken einer bestimmten literarischen Gattung, das Ergebnis verfälschen würde. Im Sinn der eingangs eingeführten *Information Poverty* muss eine solche Visualisierung immer mit der nötigen Randbetrachtung der benutzten Daten aus einer *Digital Library* einhergehen. Andernfalls besteht die potentielle Gefahr, dass hermeneutisch arbeitende Geisteswissenschaftler aus der Visualisierung falsche Schlüsse ziehen.

Oftmals ist es für ein erstes Explorieren eines Werkes bzw. einer *Digital Library* von Interesse, sich noch weiter vom Text zu entfernen als mit der zuvor genannten *Temperature Map*. Es sei die Frage nach dem Autor gestellt, welcher der größte *"Platon-Junkie"* ist? Dies impliziert nicht nur ein häufiges Zitieren Platons, sondern auch das Vermeiden der Wiedergabe von Texten anderer Autoren. Um diese Frage beantworten zu können, müsste ein traditionell arbeitender Geisteswissenschaftler die Werke von Platon auswendig lernen, um dann in allen zitierenden Autoren nach den entsprechenden Stellen zu suchen. Weiterhin müsste der Geisteswissenschaftler auch alle weiteren Texte auswendig können, um so auch Zitate in nicht-platonischen Werken zu identifizieren.

Ein weiteres Beispiel für einen komplexeren Zusammenhang ist in Abb. 1.4 dargestellt. Es gab in der Platon-Forschung zwei große Epochen: den *Mittelplatonismus* (grün) und den *Neuplatonismus* (gelb). Zu beiden Epochen werden verschiedene Stellen des Timaeus mit jeweils unterschiedlichen Themen zitiert. Während der Fokus des *Mittelplatonismus* durch Galen auf dem Ende des Timaeus liegt, in welchem weitgehend über die Anatomie des Menschen geschrieben worden ist, liegt der Fokus im *Neuplatonismus* auf den philosophischen Gedankengängen über das Sein.

Mit dem *Information Overload* und der *Information Poverty* verhält es sich sehr ähnlich wie bei den *Temperature Maps* aus Abb. 1.3. Während der *Information Overload* ebenfalls visuell durch Farben und interaktive Plots reduziert wird, benötigt diese Form der Explorierung einer *Digital Library* ebenfalls eine genaue Betrachtung der Textbasis und ihre Repräsentativität bzgl. der Ergebnisse.

[13] Die logarithmische Funktion dient lediglich dazu, der *Power-Law-Verteilung* der *Re-use Units* bzgl. ihrer *Re-use Frequency* entgegenzuwirken. Ohne die logarithmische Skalierung würde die *Temperature Map* nur schwarz bzw. dunkel sein. Dies ist durch die wenigen aber umso häufigeren *Frequency Peaks* im Vergleich zu den meist sehr niederfrequenten *Units* verursacht.

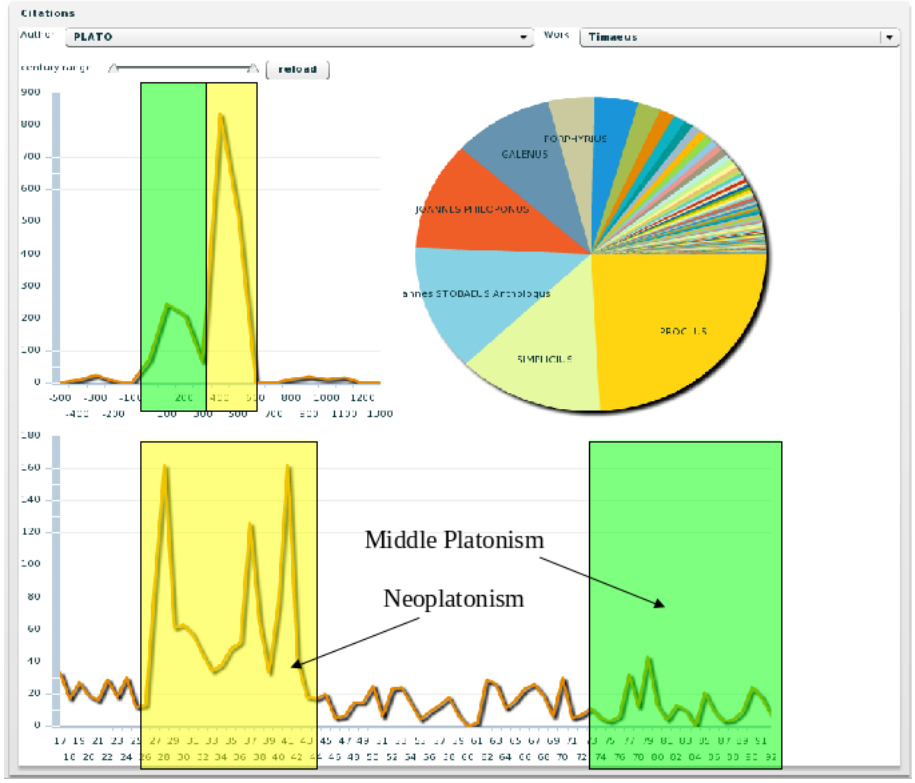


Abbildung 1.4: *Macroview* einer Digital Library, um einfach grobe Zusammenhänge zu explorieren wie bspw. der Korrelation zwischen dem Zitieren bestimmter Textpassage zu bestimmten Zeitpunkten, wie dem Neu- und dem Mittelplatonismus. Das linke obere Diagramm segmentiert einen *Re-use Graphen* nach der Zeit (x-Achse) und trägt die Häufigkeit für jede Zeitscheibe auf der y-Achse ab. Die rechte obere Darstellung zeigt die zitierenden Autoren auf. Die untere Darstellung repräsentiert die seitenweise Zerlegung des Werkes (Stephanus-Seiten) und trägt auf der y-Achse die Zitierhäufigkeit der jeweiligen Seite ab.

Die vier zuvor genannten Beispiele zeigen einführend auf, dass es in den *eHumanities* nicht nur um das Kombinieren von Daten und Verfahren hin zu Anwendungen gehen kann. Vielmehr werden gute Visualisierungen benötigt, die zukünftig die Arbeit der Geisteswissenschaften unter der Prämisse ständig steigender, digital vorliegender Daten und damit dem *Information Overload* erleichtern bzw. die Möglichkeiten geben, neue Forschungsfragen zu stellen. Schaffen es die *eHumanities* nicht, für Methoden die nötige Akzeptanz und Nutzung in den Geisteswissenschaften zu bekommen, so wird dieser noch junge Forschungsbereich eine reine "Elfenbeinturmforschung" werden.

1.5 Das "ACID for the eHumanities" Paradigma

Was macht eine erfolgreiche Forschung in den *eHumanities* aus? In den Abschnitten 1.3 und 1.4 ist bereits auf zwei Hauptprobleme der *eHumanities* eingegangen worden. Einerseits muss es das intrinsische Interesse sein, diese junge Disziplin im Sinne des *vertikalen Spannungsverhältnisses* nicht mehr als interdisziplinär zu betrachten, sondern vielmehr als einen neuen Forschungsbereich, der aus der Überlappung der Geisteswissenschaften und der

Informatik hervorgegangen ist. Andererseits müssen automatisierte Methoden immer auch unter der Berücksichtigung von *Information Overload* und *Information Poverty* betrachtet werden, da nicht alles, was technisch machbar ist, auch zeitgleich einen fachwissenschaftlichen Mehrwert bedeutet. Dies heißt im Detail, dass Methoden geschaffen werden müssen, die Fragen beantworten und nicht Zweifel am Ergebnis schaffen.

Seit 2008 wurde in Leipzig aus den *eHumanities*-Projekten *eAQUA*[14] und *eTRACES*[15] gelernt, wie mit dem *Spannungsverhältnis*, dem *Information Overload* und der *Information Poverty* umgegangen werden kann. Im Projektmanagement sind Techniken, wie die *SWOT*-Analyse oder das *Waterfall*-Modell, bereits hinreichend etabliert. Jedoch reichen beide Methoden bei *eHumanities*-Projekten nicht aus. Die Erfahrungen wurden zum *ACID for the eHumanities* Paradigma (vgl. [Büchler 2012a][16]) zusammengefasst, welches keine Garantien für eine erfolgreiche eHumanities-Forschung geben kann. Jedoch soll es helfen, bereits genannte Probleme während der Projekt- und Forschungsphase zu identifizieren. Diese gesamte Arbeit unterliegt dem *ACID for the eHumanities* Paradigma. *ACID* ist hierbei eine Abkürzung für *Acceptance*, *Complexity*, *Interoperability* und *Diversity*. Hinter diesen vier Konzepten steht jeweils folgende Frage:

- **Acceptance:** Jede IT-nahe Forschung ist ein Irrweg, wenn die Methode von den Fachwissenschaften nicht akzeptiert wird. Daher stellt sich insbesondere im Kontext des *vertikalen Spannungsfeldes* die Frage, wie im Kontext der *eHumanities* die Akzeptanz für eine Methode gesteigert werden kann.

- **Complexity:** Die fachwissenschaftliche Forschung ist meist sehr komplex und wenig auf einzelne Algorithmen reduzierbar. So ist das Erkennen von *Text Re-use Traces*[17] lediglich eine Aufgabe des *Historical Text Re-use Detection*. Andere Aufgaben, die meist von der IT unberücksichtigt bleiben, sind bspw. das Typisieren von *Text Re-use Units* bzw. den *Links* zwischen zwei *Text Re-use Units*. Weiterhin zählen zu den *Text Re-use Task*[18], die bereits genannte *Archetype Detection*, die *Intention Detection*, die *Meaning Detection* oder auch die *Direction Detection*. Damit geht die Frage einher, welche Aufgaben für die Forschung nötig sind.

- **Interoperability:** Insbesondere die Infrastrukturinitiativen *CLARIN*[19], *DARIAH*[20] oder *Bamboo*[21] müssen sich der Aufgabe stellen, die in den letzten 20 Jahren digitalisierten und oftmals weit verstreuten Daten so zu homogenisieren, dass jedwede Verfahren über ein einheitliches Format auf die Daten zugreifen können[22] (vgl. [Pansch 2010, Eckart 2011]. Daher muss sich die Frage nach der *Interoperability* der Daten und Verfahren gestellt werden.

- **Diversity:** Die herausforderndste Säule des *ACID*-Paradigmas behandelt den Umgang der Diversität in der Benutzung menschlicher Sprache. Menschen benutzen Sprache nicht gleich, so dass oftmals das aus der Informatik propagierte *Sprachmodell* bestenfalls die dominanten jedoch nicht alle möglichen Daten extrahiert.

[14] vgl. `http://eaqua.e-humanities.net`
[15] vgl. `http://etraces.e-humanities.net`
[16] Das *ACID for the eHumanities* Paradigma wurde erstmals im Rahmen des CLARIN-D M12 Workshops in Leipzig vorgestellt und ist zum Zeitpunkt des Druckes dieser Arbeit noch nicht publiziert.
[17] *Text Re-use Traces* wird nachfolgend nur noch *Traces* genannt, da aus dem Kontext dieser Arbeit klar ist, dass es sich um Spuren des *Text Re-use* handelt.
[18] Die Text Re-use Tasks werden in Kapitel 2.7 genauer erklärt.
[19] vgl. `http://www.clarin.eu/`
[20] vgl. `http://www.dariah.eu/`
[21] vgl. `http://www.projectbamboo.org/`
[22] siehe auch vgl. `http://laquat.cerch.kcl.ac.uk/`

Im Kontext des *Historical Text Re-use Detection* kann das *ACID for the eHumanities* Paradigma sehr vielschichtig umgesetzt werden. Im Rahmen des *eTRACES*-Projektes wird so die *Acceptance* durch vier Aspekte verbessert:

1. **nutzernahe Entwicklung:** Wie bereits erwähnt wird *Text Re-use* im Rahmen des Göttinger Teilprojektes in *eTRACES* nah am Interesse der Fachwissenschaften entwickelt. Hierbei wird das Erstellen von Kommentaren bzgl. Spuren von Zitaten, Paraphrasen oder Allusionen so unterstützt, dass der Nutzer beim Erstellen einer *Online Edition* sich bereits entsprechende Textstellen anzeigen lassen kann und nicht erst Satz für Satz nach *Parallelstellen*[23] suchen muss. Hierdurch kann die *Acceptance* sowohl durch ein konkretes Problem, wie das der Erstellung einer *Online Edition*, aber auch durch den Geschwindigkeitsvorteil der automatisierten Methode erreicht werden.

2. **Annotation von Daten:** Im Rahmen der eingangs genannten Projekte hat es sich als vorteilhaft herausgestellt, die Fachwissenschaftler zu Beginn erst einmal einen *Gold Standard* für das *Benchmarking* erstellen zu lassen. Dadurch wird nicht nur ein Verständnis dafür generiert, was untersucht werden soll, sondern es hilft insbesondere auch, *Anforderungscluster* zu bilden. In der Regel bedeutet dieser Schritt, dass mehr als eine Technik notwendig ist, so dass anschließend die Ergebnisse mehrerer Algorithmen wieder zusammengesetzt werden können.

3. **Realistische Erwartungen:** Nahezu einhergehend mit dem *Annotieren von Daten* werden auch realistische Erwartungen geweckt. So ist ein *Syntactic Text Re-use*, wie Zitate, meist sehr einfach zu erkennen. Ein *Semantic Text Re-use*, wie *Allusionen* und einfachere *Paraphrasen*, ist oftmals noch technisch realisierbar. Ein *Cognitive Text Re-use* in Form von semantisch entfernteren *Paraphrasen* oder *Analogien* hingegen ist nahezu nicht mehr automatisch umsetzbar. So wäre es zweifelsohne wünschenswert, *Re-use Units*, wie *like will to like* und *Birds of same feather flock together*, aufeinander verlinken zu können, jedoch ist dies zumindest technisch derzeit nicht realistisch.

4. **webbasierte Text Re-use Debugger:** Oftmals scheitert die *Acceptance* von Methoden des *Text Mining* daran, dass die Fachwissenschaften nicht präzise verstehen, was in der *Black Box* des *Text Mining* genau passiert. Im Rahmen des *eTRACES*-Projektes wird hierzu gerade paradigmatisch an einem webbasierten *Text Re-use Debugger*, ähnlich einem Debugger für Programmiersprachen in einer Entwicklungsumgebung, gearbeitet, welcher die einzelnen Zwischenschritte transparent und nachvollziehbar im Browser darstellt.

Die *Complexity* und die *Diversity* sind essenzielle Bestandteile dieser Arbeit und werden detailliert in den Kapitel 2.6 und 2.7 behandelt. Beide setzen sowohl fundiertes Wissen über das Korpus sowie solide Erfahrung mit den Methoden voraus.

Ziel des *ACID for the eHumanities* Paradigmas ist es, sich durch die vier Säulen *Acceptance*, *Complexity*, *Interoperability* und *Diversity* eine gesunde Forschungsumgebung zu schaffen. Dies inkludiert neben dem Verständnis für die fachwissenschaftlichen Problemstellungen auch das Verständnis und die Erfahrung dafür, wie *Text Re-use* verwendet wird.

[23] Eine *Parallelstelle* im geisteswissenschaftlichen Sinn ist eine Textpassage, die den gleichen Inhalt wiedergibt. Hierbei kann es sich nach fachwissenschaftlicher Prüfung sowohl um ein Zitat als auch nur um zwei Zeugen bspw. ein und desselben Events handeln.

1.6 Herausforderungen des textuellen Wissenstransfers auf geisteswissenschaftlichen Texten

Text Re-use im Speziellen und *eHumanities* im Allgemeinen wurden bisher unter dem Aspekt des *interdisziplinären Spannungsfeldes* (vgl. Abschnitt 1.3), dem *Information Overload* und der *Information Poverty* (vgl. Abschnitt 1.4) sowie den Erfolgschancen (vgl. Abschnitt 1.5) betrachtet. *Historical Text Re-use* zeichnet sich jedoch im Vergleich zu benachbarten Disziplinen, wie dem *Plagiarismus*, dem *Software Re-use*, der menschlichen *DNA-Analyse* aber auch der Analyse des menschlichen *Fingerabdruckes* dadurch aus, dass aufgrund der enormen Zeitspannen *Traces* sowohl viel evolutionärer Dynamik ausgesetzt waren als auch, dass über die Jahrhunderte ein enormes und oftmals nicht digitales Wissen über *Parallelstellen* eines *Text Re-use* gesammelt wurde. So werden in [Büchmann 2007] akribisch speziell *Geflügelte Wörter*, in [Schemann 2000] Deutsche *Redensarten*, in [Apfel 2010] speziell *Zitate* und *Weisheiten* sowie in [Alsleben 2007] ebenfalls Zitate und Redewendungen zusammengestellt. Einerseits liegt bei manchen Sammlungen der Fokus auf der Erklärung eines *Text Re-use*. Andererseits stellen andere Werke reine Sammlungen von Parallelstellen dar (vgl. Abb. 1.5[24]). Allein die hier genannten Werke haben ein Gesamtvolumen von über 100.000 Datensätze. Ergänzt wird dieses Wissen durch kleinere Spezialsammlungen wie [Wagner 2010] und [Wagner 2011b], die spezielle *Redewendungen* aus dem Mittelalter bzw. der Bibel beinhalten sowie Sammlungen, wie in [Spears 1998], von amerikanische *Phrasen* und *Idiomen*.

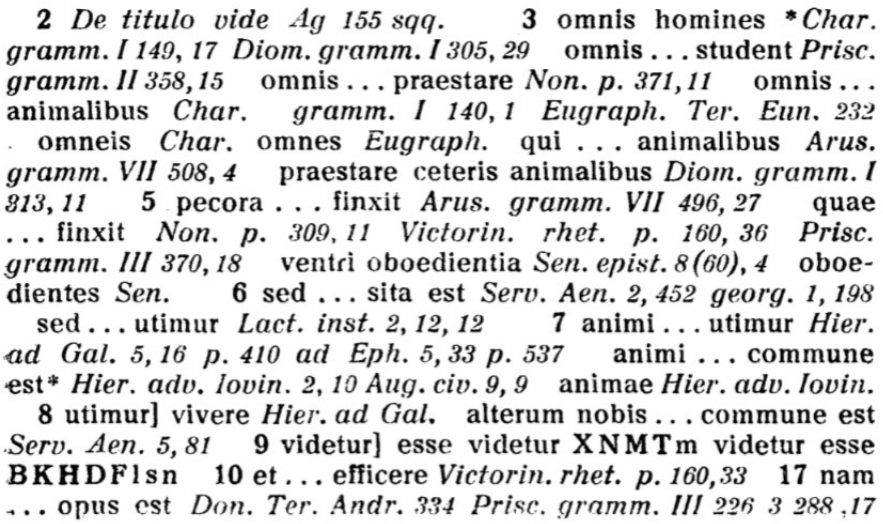

2 *De titulo vide Ag 155 sqq.* 3 omnis homines **Char. gramm. I 149, 17 Diom. gramm. I 305, 29* omnis . . . student *Prisc. gramm. II 358, 15* omnis . . . praestare *Non. p. 371, 11* omnis . . . animalibus *Char. gramm. I 140, 1 Eugraph. Ter. Eun. 232* omneis *Char.* omnes *Eugraph.* qui . . . animalibus *Arus. gramm. VII 508, 4* praestare ceteris animalibus *Diom. gramm. I 313, 11* 5 pecora . . . finxit *Arus. gramm. VII 496, 27* quae . . . finxit *Non. p. 309, 11 Victorin. rhet. p. 160, 36 Prisc. gramm. III 370, 18* ventri oboedientia *Sen. epist. 8 (60), 4* oboedientes *Sen.* 6 sed . . . sita est *Serv. Aen. 2, 452 georg. 1, 198* sed . . . utimur *Lact. inst. 2, 12, 12* 7 animi . . . utimur *Hier. ad Gal. 5, 16 p. 410 ad Eph. 5, 33 p. 537* animi . . . commune est* *Hier. adv. Iovin. 2, 10 Aug. civ. 9, 9* animae *Hier. adv. Iovin.* 8 utimur] vivere *Hier. ad Gal.* alterum nobis . . . commune est *Serv. Aen. 5, 81* 9 videtur] esse videtur XNMTm videtur esse BKHDFlsn 10 et . . . efficere *Victorin. rhet. p. 160, 33* 17 nam . . . opus est *Don. Ter. Andr. 334 Prisc. gramm. III 226 3 288 ,17*

Abbildung 1.5: Die Abbildung zeigt Anmerkungen von *Axel Ahlberg's* Edition aus dem Jahre 1913 zu *Sallust's Catilinarian Conspiracy*. Mehr als 20 Verweise auf zitierende Textstellen sind im Apparat aufgelistet.

All die hier genannten und ungenannten Bücher zeigen valide auf, dass es schon immer ein intrinsisches Interesse am Sammeln von *Text Re-use* gab. So ist es auch nicht verwunderlich, dass heutzutage geisteswissenschaftlich orientierte Projekte, wie das *Corpus der*

[24] Dieses Beispiel ist von Prof. Dr. Gregory Crane im Rahmen der gemeinsamen Publikation [Büchler 2012c] recherchiert und beigesteuert worden. In dieser Arbeit wurde die valide Argumentationslinie dieses Beispiels übernommen.

Arabischen und Syrischen Gnomologien [Pietruschka 2012] und *Sharing Ancient Wisdoms* [Roueché 2010], welche beide im griechischen, arabischen und syrischen Sprachraum ethische und moralische Weisheitssprüche sammeln, essenziell zur internationalen Community gehören, an welcher sich die *eHumanities* zwangsläufig messen lassen müssen.

Was kann beim Recherchieren aus all diesen Nachschlagewerken gelernt werden? In erster Linie untermauern diese Nachschlagewerke die *Complexity* und *Diversity* (vgl. Abschnitt 1.5). Sie machen aber auch deutlich, dass der *Historical Text Re-use* eine Wechselwirkung aus gesprochener und geschriebener Sprache ausgesetzt ist bzw. war, eine Vielzahl von Varianten besitzt, absichtlich oder nicht absichtlich genutzt wird bzw. wurde, aber auch der allgemeine Umgang mit historischen Sprachen sich als schwieriger herausstellt, als bei Sprachen mit aktiven Muttersprachlern, da nichts als gesichert angenommen werden kann.

Sprache wird nicht nur geschrieben, sondern natürlich auch gesprochen. Wir alle produzieren im täglichen Leben wesentlich mehr gesprochene als geschriebene Sprache. Im Kontext des *Text Re-use* ergibt sich somit die grundlegende Frage nach dem *Transfermedium*. Der Leser sei einmal auf ein Gedankenexperiment mitgenommen. Wahrscheinlich jeder Erwachsene kennt das *Geflügelte Wort Gleich und Gleich gesellt sich gern.* Wie oft hat der Leser dieses *Geflügelte Wort* bisher schreibend verwendet und wie oft sprechend? Wie würde es sich mit den Phraseologismen *Geld stinkt nicht!* oder *jemandem auf's Dach steigen* aussehen? Es sollte offensichtlich sein, dass wir wesentlich öfter bestimmte *Re-use Units* mündlich anstatt geschrieben wiedergeben. Da mündliche Kommunikation meist sehr kurzlebig ist, stellen die heute noch erhaltenen historischen Dokumente nicht nur einen fragmentarischen Ausschnitt des *Historical Text Re-use* dar, sondern sind insbesondere auch kein repräsentativer Textausschnitt (siehe *Information Poverty* in Abschnitt 1.4). Letztlich ergibt sich aus der *gesprochenen* und *geschriebenen Sprache* sowie der *synchronen* und *asynchronen Kommunikation* eine Matrix, die in der Tabelle 1.1 dargestellt ist.

		Sprache	
		gesprochen	geschrieben
Kommunikation	synchron	**Beispiele**: Gespräch, mündlicher Dialog **Eigenschaft**: temporär, kurzlebig	**Beispiele**: schriftlicher Dialog, Chat **Eigenschaft**: langlebig
	asynchron	**Beispiele**: Monolog, Filme, Theateraufführungen, Weitergabe von Ritualen und Traditionen, Sozialisierungen **Eigenschaft**: temporär, kurzlebig	**Beispiele**: Text Re-use und Knowledge Transfer **Eigenschaft**: langlebig, kann aber auch nach dem Aussterben wieder reaktiviert werden.

Tabelle 1.1: Wechselwirkung zwischen Sprache und Kommunikation: Der *Historical Text Re-use* hängt in der Gegenwart sehr davon ab, ob bestimmte *Re-use Units* im Laufe der Zeit aufgeschrieben worden sind. Zielstellung einer ganzheitlichen Betrachtung des *Historical Text Re-use* würde sein, alle vier in der Tabelle genannten Quadranten in Betracht ziehen zu können.

Die in Tabelle 1.1 dargestellten Quadranten entsprechen hierbei unterschiedlichen Forschungsbereichen bzw. Disziplinen. Das wünschenswerte Ziel einer ganzheitlichen Betrachtung des *Historical Text Re-use* wäre, alle vier Quadranten in Betracht ziehen zu können, auch wenn dies sicher speziell für den oberen linken Quadranten (gesprochene Sprache mit synchroner Kommunikation) aus Tabelle 1.1 heutzutage äußerst schwierig ist. Jedoch kann

eine Analyse der gesprochenen asynchronen Kommunikation in der Gegenwart Aufschluss darüber geben, ob bestimmte *Text Chunks* wie *Phraseologismen* oder *Redewendungen* über die Jahrhunderte innerhalb der Sprache mündlich weitergegeben wurden. In diesem Sinne hat der Autor dieser Arbeit im Juli 2012 ein kleines Experiment mit einigen Kollegen[25] der Philologien[26] durchgeführt. Ausgangspunkt war die Frage von Anett Büchler[27], ob der Autor wüsste, wo der Ursprung der Phrase *Äpfel mit Birnen vergleichen* herkommt. All die eingangs genannten Nachschlagewerke konnten keinen Aufschluss darüber geben. Eine Recherche bei Google Books ergab, dass das früheste Auftreten im Deutschen auf ein Märchen der Gebrüder Grimm (frühes 19. Jh.) zurückgeht, was aber als *Archetyp* zu unwahrscheinlich erschien. Daraufhin wurden die genannten Philologen befragt. Sowohl im Lateinischen, Altgriechischen als auch im Arabischen wurden keine schriftlichen Belege gefunden. Jedoch konnte identifiziert werden, dass es für diese Redewendung in sehr vielen modernen europäischen und nordamerikanischen Sprachen vergleichbare Phrasen gibt. So scheint es im Russischen und Arabischen zwar kein entsprechendes Pendant zu geben. Im Italienischen *confrontare le mele con le pere*[28] und Englischen *comparing apples with oranges* gibt es jedoch Varianten, die auffällig gleich sind. Zu irgendeinem Zeitpunkt in der Vergangenheit wurden Birnen und Orangen konzeptuell ausgetauscht, ohne jedoch den Sinn zu verändern. Auch wenn der *Archetyp* nicht gefunden werden konnte, so lassen die gesammelten Informationen und das Wissen über Abhängigkeiten zwischen Sprachen durch ihre Sprachfamilie die Vermutung aufkommen, dass der *Archetyp* wahrscheinlich im Mittelalter in Europa in der lateinischen Philosophie der Gelehrten zu findet sein wird.

Dieses Beispiel soll zwei Dinge aufzeigen. Erstens, ist es sehr schwierig, die *Trace* eines *Text Re-use* nachzuvollziehen. Dies gilt insbesondere dann, wenn Sprachgrenzen überbrückt werden müssen. Zweitens, auch wenn die asynchrone gesprochene Kommunikation der Gegenwart weitestgehend außerhalb der Forschung zum *Historical Text Re-use* bleibt, so sind die Ähnlichkeiten von *Phrasen* und *Redewendungen* zwischen Sprachen und deren sprachfamiliären Abhängigkeiten gute Indikatoren für das *Backwards Tracing* ebendieser. Zukünftig wäre es dementsprechend wünschenswert, im Rahmen eines *Philological Crowd Sourcing*, *Text Re-use* in verschiedenen Sprachen aufeinander zu verlinken, um die Möglichkeit des *Backwards Tracing* erstmals zu ermöglichen. Diese Methode sei insbesondere im historischen Kontext der Produktionskosten von geschriebenen Texten betrachtet. Heutzutage sind die Produktionskosten von Geschriebenem durch den Buchdruck und noch viel mehr durch das Internet sehr gering. Speziell in der Antike und dem Mittelalter jedoch sind die Kosten, geschriebenen Text zu produzieren, ungleich größer gewesen. So bleibt nur die Vermutung, dass speziell bis zur Einführung des Buchdruckes der *Text Re-use* stärker mündlich als schriftlich verbreitet worden ist, was die Forschung im Rahmen dieser Arbeit sichtlich erschwert und Ergebnisse immer unter dem Aspekt dieser Herausforderung betrachtet werden müssen.

Eine weitere Aufgabe besteht darin, wie *Text Re-use* gemessen wird bzw. was *Text Re-use* überhaupt ausmacht? Einerseits kann er wie in wissenschaftlichen Publikationen[29] oder dem *Philosophical Text Re-use* (vgl. [Büchler 2010e, Büchler 2013c]) auf Satzebene sowie auf einem kleineren *Moving Windows* (vgl. [Büchler 2012c]) gemessen werden. Andererseits ist es eine Herausforderung, automatisch zwischen *Language Re-use* und *Text Re-use* zu unterscheiden. Insbesondere der sehr kurze *Re-use*, wie der innerhalb der *Perseus Digital*

[25] in alphabetischer Ordnung: Monica Berti (Italien), Federico Boschetti (Italien), Neil Coffee (USA), Gregory Crane (USA), Ute Pietruschka (Deutschland), Bruce Robertson (Kanada)

[26] Historische Sprachen: altgriechisch, latein, arabisch, syrisch. Moderne Sprachen: englisch, deutsch, italienisch, französisch

[27] Anett Büchler ist die Ehefrau des Autors und hat keinen wissenschaftlichen Hintergrund.

[28] Diese italienische Variante ist ein wortwörtliche Übersetzung der deutschen Phrase.

[29] vgl. `http://etraces.e-humanities.net/project-partners/gesis-etraces.html`

Library (vgl. [Crane 1985, Büchler 2012c]), erschwert es deutlich, eine Unterscheidung zwischen einer *Co-occurrence*, einem *Bigram* oder einem *Trigram* und kurzem *Text Re-use*, wie *(Pecunia) non olet.*[30] oder *veni, vidi, vici*, machen zu können. Das Problem dieser Separierung wird durch syntaktische Bausteine, wie *im Namen unseres Herren Jesus Christus*, oder auch *Multi Word Units*, wie *König Alexander der Große*, verstärkt. Die ganze Tragweite dieser Aufgabe wird deutlich, wenn sich der Leser einmal den Spracherwerb von Kleinkindern betrachtet. In der Kommunikation zwischen Kindern und Eltern werden zu Beginn Phrasen wie *Spielen gehen* oder *Oma fahren* ausgetauscht, die dann später zu ganzen Sätzen formuliert werden. Angesichts dessen, dass wir über solche Phrasen lernen, uns bereits im Kindesalter zu artikulieren, sollte es dem Leser ein Leichtes sein, sich zu überlegen, wie sehr unsere Sprache von solchen Kurzphrasen und syntaktischen Satzbausteinen durchzogen ist. Dementsprechend schwierig ist es, den *Text Re-use* von solchem *Language Re-use* zu separieren.

Mit der *Diversity* (vgl. Abschnitt 1.5) der fachwissenschaftlichen Disziplin geht einher, dass der *Text Re-use* unterschiedlich gemessen werden muss. Einerseits besteht eine *Re-use Unit* aus der Philosophie, wie Shakespeare's *To be, or not to be, that is the question*, oft aus sehr allgemeinsprachlichen Wörtern. Ein Entfernen aller Stoppwörter würde diesen *Spruch* unauffindbar machen. In der Historiographie hingegen wird *Text Re-use* sehr viel freier eingesetzt. Wird so bspw. der Text über ein Ereignis, wie einen Krieg oder einen Konflikt, wiederverwendet, so sind oftmals lediglich die *Named Entities*, wie Personen, Datierung, Namen für den Konflikt oder Krieg, stabil. Während in der Philosophie der *Re-use* meist sehr nah am Original bleibt, was eine wortwörtliche oder nahezu wortwörtliche Analyse auf Basis von *Ngrams* nahelegt, so empfiehlt es sich, in der Historiographie *Named Entities* als *Re- use Features* zu benutzen. Auch wenn diese verschiedenen Vorgehensmuster am Beispiel dem Leser vielleicht plausibel erscheinen, so sei dem Leser empfohlen, sich zu überlegen, wie dieses Vorgehen auf einer *Digital Library* berechnet werden kann, wenn mehrere dieser unterschiedlichen *Re-use Styles* in einem meist heterogenen Textbestand enthalten sind. Mehrere Durchläufe, die die unterschiedlichen *Re-use Styles* messen und am Ende die Ergebnisse im Sinne eines *Hybrid Text Re-use* zusammenfügen, sind hierbei nur der zweite Schritt. Die Herausforderung liegt in einem ersten Schritt darin, die enthaltenen *Re-use Styles* oder zumindest die, die für eine Forschungsfrage von Interesse sind, zu erkennen (vgl. *Diversity* des *ACID for the eHumanities* Paradigmas aus Abschnitt 1.5).

Eine letzte Herausforderung des *Historical Text Re-use* ist die Diskrepanz zwischen *statistischer Aussagefähigkeit* eines *Bi-* oder *Trigrams* und der *sprachspezifischen Relevanz*. Insbesondere *Altgriechisch* und *Latein* haben eine wesentlich geringere syntaktische Festigkeit als modernere Sprachen. Im Gegenzug besitzen diese Sprachen eine meist komplexere Morphologie, die wiederum über die morphologischen Dependenzen die Abhängigkeiten zwischen den Wörtern abbildet. Es sei nun einmal folgendes Gedankenexperiment gemacht. Es sei ein Bigram wie *non olet!* gegeben. Es wird ein Signifikanzmaß, wie das *Log Likelihood Maß* [Dunning 1993] oder die *Mutual Information* [Church 1989], eingesetzt, um eine Aussage über die statistische Aussagefähigkeit zu machen. Alle nötigen Parameter, wie die Häufigkeit beider Wörter und die Häufigkeit des gemeinsamen Auftretens, seien als identisch angenommen zu *not smell*, der Englisch wortwörtlichen Übersetzung. Numerisch würden beide Beispiele die gleiche statistische Signifikanz haben. Jedoch ist selbst bei gleichem Signifikanzwert *non olet* aufgrund der wesentlich freieren Wortstellung aus der sprachspezifischen Gegebenheit im Lateinischen deutlich relevanter als im Englischen, welches eine stärkere syntaktische Festigkeit besitzt. Dieses Beispiel soll illustrieren, dass statistische Maße, wie sie zahlreich in der Automatischen Sprachverarbeitung eingesetzt

[30] dt.: Geld stinkt nicht!

werden, nur sehr bedingt vergleichbar sind. Erst recht können Schwellwerte selbst bei gleichem Maß nicht einfach adaptiert werden. Dies birgt letztendlich die Gefahr, dass während der Erstellung der *Digital Signature* einer *Re-use Unit* auf keinerlei Erfahrungen der Vergangenheit zurückgegriffen werden kann. Ferner soll an dieser Stelle offen bleiben, wie in einem solchen Szenario evaluiert werden kann. Es sei jedoch bereits auf die Kapitel 4 und 5 verwiesen.

In diesem Abschnitt wurden nur sehr fragmentarisch einige Herausforderungen, wie der Umgang mit *gesprochener* vs. *geschriebener* Sprache, die *Diversity* des *Re-use Styles* oder auch der Umgang mit *statistischen Maßen*, genannt und ausgeführt. Ergänzt durch den komplizierteren Umgang im Vergleich zu modernen Sprachen mit verschiedenen Schreibweisen, bedingt durch Dialekte oder Sprachevolution bzw. semantischen Varianten, die teilweise deutlich von der heutigen Bedeutung abweichen oder auch der Umgang mit der *Information Poverty* (vgl. auch [Cayless 2010]), muss an dieser Stelle die Frage nach dem Sinn und der Zielstellung dieser Arbeit im Bereich der *eHumanities* gestellt werden. Letztlich zeigen immerhin Berichte, wie [Babeu 2011], auf, dass es noch weit mehr Aufgaben gibt, als hier ausgeführt oder auch nur genannt wurden.

Auch wenn es angesichts der gegebenen Herausforderungen erdrückend erscheint, im Bereich des *Historical Text Re-use* eine Promotion zu schreiben, so sind es genau diese Herausforderungen, welche das Thema für die Informatik erst richtig interessant machen. Der Leser sei einmal angehalten, die hier genannten Aspekte im Kontexkt der *Plagiarismusforschung* zu betrachten. Es sollte deutlich werden, dass die Schwierigkeiten im *Historical Text Re-use* ungleich schwerer sind. Auch wenn es faktisch unrealistisch ist, alle Aufgaben im Rahmen dieser Arbeit zu lösen, so ist es das unmittelbare und intrinsische Interesse dieser Arbeit, den Forschungsbereich des *Historical Text Re-use* nicht nur für die Informatik aufzuzeigen, sondern insbesondere auch erstmals zu strukturieren. So fällt bspw. grundlegend bei der Literaturrecherche auf, dass es zwar einige Arbeiten in diesem Wissenschaftsbereich gibt, jedoch fehlt es an jedweder grundlegender Systematisierung. Für die Systematisierung des *Historical Text Re-use* dient neben Einflüssen aus der Informatik und den Geisteswissenschaften insbesondere die *Biometrie*. In Anlehnung an diese verwandten Wissenschaften ist neben der Evaluierung (siehe Abschnitt 5.1) auch das erste Glossar zum *Historical Text Re-use* entstanden.

Die Bedeutung dieser fachwissenschaftlichen Herausforderungen sei an dieser Stelle einmal der Methodik der Informatik gegenübergestellt. Einerseits sei eine in der Informatik übliche Evaluierung mit *Precision*, *Recall* oder *F-Measure* gegeben, die bspw. *Text Re-use* in Zahlen ausdrückt und dabei oftmals systematische Probleme eines Ansatzes wie die *Diversity* der verschiedenen *Text Re-use Styles* hinter den Zahlen versteckt. Andererseits sei auf den in diesem Abschnitt bereits genannten *Archetyp* bzw. die *Bedeutung* eines *Text Re-use* verwiesen. Immer wieder konnte der Autor im Rahmen seiner Arbeiten feststellen, dass es ein großes gesellschaftliches Interesse am *Cultural Heritage*, unserem kulturellen Erbe, speziell in Europa gibt. Das im Rahmen dieses Abschnittes genannte Beispiel *Äpfel mit Birnen vergleichen* und dem Interesse von Anett Büchler nach dem Ursprung und dem damit verbundenen Verständnis über die Semantik bzw. dessen semantische Benutzung zeigen nur exemplarisch das gesellschaftliche Interesse am *Cultural Heritage* auf. So kann das meist negative Sentiment[31] des *Geflügelten Wortes Gleich und Gleich gesellt sich gern* auf Homer im 8. Jh. v. Chr. zurückgeführt werden. Das Sprichwort *Geld stinkt nicht!* kann auf einen Dialog von Kaiser Vespasian mit seinem Sohn im 1. Jh. zurück verfolgt werden, in welchem Vespasian's Sohn auf die Einführung einer öffentlichen Latrinensteuer zu ihm sagte, dass man für die Benutzung der öffentlichen Toiletten, um Urin für die Lederherstellung zu sam-

[31] Stimmung bzw. Bedeutung

meln, doch keine Steuer erheben kann. Die kurze Antwort von Kaiser Vespasian war: *Geld stinkt nicht!* Der Phraseologismus *jemandem auf's Dach steigen* kann auf eine germanische *Tradition* zwischen dem 7. und 9. Jh. zurückgeführt werden. In jener Zeit sind die jungen Männer eines Dorfes jemandem auf das Dach gestiegen, der nicht der gesellschaftlichen Norm des jeweiligen Dorfes entsprach, um dieses abzudecken, so dass Kälte und Regen ins Haus gelangten. Bereits im Mittelalter wurde diese Form der gesellschaftlichen Willkür in frühen konstitutionellen Texten verboten. Das kulturelle Erbe dieses Spruches reicht bis hin zur aktuellen Verfassung der Bundesrepublik Deutschland, in welcher es im Artikel 13 GG. heißt: *Die Wohnung ist unantastbar.*

Diese wenigen Beispiele sind nicht etwa ausgewählte *Sprüche*, *Phrasen* und *Redewendungen*, die eine besondere Geschichte hervorbringen. Vielmehr steht hinter nahezu jedem *Historical Text Re-use* eine solche Geschichte, die entdeckt und aufgedeckt werden will. Gemessen am gesellschaftlichen Interesse und den eben aufgeführten Potentialen, sowohl für die Benutzung von Sprache, dem kulturellen Erbe, als auch dem gesellschaftlichen Interesse, sind die in diesem Abschnitt genannten Herausforderungen vielmehr Auftrag als Aufgabe für zukünftige Forschungen. Letztlich impliziert diese Erkenntnis, dass es im Kontrast zur Informatik, in welcher oftmals nur Sprachmodelle verglichen werden, wie *Mein Sprachmodell ist im F-Score um 2% besser als dein Sprachmodell.*, in den *eHumanities* nicht um den Vergleich von Modellen gehen kann, sondern diese als "funktionierend" angenommen werden müssen, um entsprechende Fragestellungen beantworten zu können.

Auch wenn die Möglichkeiten und Perspektiven des *Cultural Heritage* im Sinne des *Historical Text Re-use* sehr vielversprechend scheinen, so muss jedem Leser klar sein, dass das Forschungsperspektiven der nächsten 10 Jahre und mehr sind. Beim großen Bruder der *eHumanities*, der Bioinformatik, hat es ebenfalls 10 - 15 Jahre gedauert, bis auf der Basis der digital vorliegenden Daten und entsprechenden Methoden ernsthafte Ergebnisse, wie das Auffinden der Gensequenz, die Epilepsie verursacht (vgl. Abschnitt 1.3), möglich waren. Daher muss diese Arbeit als eine Grundlagenarbeit verstanden werden, die Folgearbeiten ermöglichen wird, um sich dem Aufdecken des *Cultural Heritage* zu stellen.

1.7 Wissenschaftliche Einbettung des historischen Wissenstransfers in den Forschungsbereich der Informatik

Für den *Historical Text Re-use* sind im Kontrast zur *Plagiarismusforschung* große Zeiträume von teilweise mehreren Jahrtausenden eine grundlegende Konstante. In diesen großen Zeitspannen sind die Texte entweder verloren gegangen (vgl. *Information Poverty* in Abschnitt 1.4 und [Cayless 2010]) oder wurden zumindest Veränderungen ausgesetzt. Um die Auswirkungen dieser Zeitspanne richtig verstehen zu können, muss der Prozess der Textüberlieferung von der Antike in die Moderne verstanden werden. Heutzutage gibt es nahezu keine Originale mehr. Jeder Text, den wir heute noch überliefert wissen, wurde im Laufe der Zeit zahlreich abgeschrieben. Daraus entsteht ein sogenanntes *Stemmata*[32] eines Textes, zu welchem ähnlich der genetischen Abstammungskarte nicht alle, insbesondere die frühen, Informationen vorhanden sind. Bei jedem Abschreiben werden absichtliche und unabsichtliche Veränderungen der Vorlage vorgenommen. Diese Veränderungen sind sehr

[32] Ein *Stemmata* kann als ein Familienbaum für einen Text verstanden werden. Hierbei entsprechen die Knoten dieses gerichteten Graphen den Abschreibern. Die Kanten zwischen den Knoten repräsentieren die Abschreibungsverhältnisse zwischen den Werken. Ein *Stemmata* ist grundlegender Bestandteil der *Textkritik*.

vielschichtig und bilden in den Geisteswissenschaften den Forschungsbereich der Textkritik (vgl. [Dover 1997]). Veränderungen können einerseits durch im Mittelalter abschreibende Mönche verursacht worden sein, die Texte falsch kopiert oder Rechtschreibfehler bzw. ungewöhnliche Schreibweisen in bester Absicht korrigiert haben[33]. Andererseits kann der Herausgeber einer modernen und digitalen Edition eines Textes, unleserliche Textstellen eigenmächtig vervollständigt haben, ohne dabei auf dessen unsichere Überlieferung in Form der *Leiden Convention* hinzuweisen.

Um derartig komplexe Prozesse und Abhängigkeiten abbilden zu können, wird der Forschungsbereich des *Historical Text Re-use* in Shannon's *Noisy Channel* (siehe Abb. 1.6) eingebettet (vgl. [Shannon 1948]). Shannon arbeitete in den 40iger und 50iger Jahren des 20. Jahrhunderts bei einem US-amerikanischen Telefonunternehmen. In jener Zeit beschäftigte er sich mit der Übertragung von Sprachsignalen. Hierbei wird ein Signal $\mathcal{S}$ von einem *Transmitter* über einen verrauschten Kanal gesendet. Ein *Receiver* empfängt schließlich das Signal $\mathcal{S}'$. Daraus resultierte die Frage danach, wie groß die Abweichung Δ mit

$$\Delta \leq |\mathcal{S}' - \mathcal{S}| \tag{1.1}$$

maximal sein kann, damit das übertragene Signal verstanden wird. In Shannon's Fall besteht der *Noisy Channel* beispielsweise aus sphärischen Störungen oder dem Überlagern von verschiedenen Signalen.

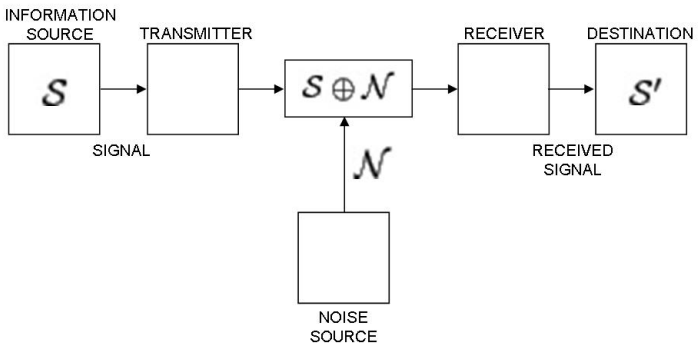


Abbildung 1.6: Shannon's *Noisy Channel*: Ein *Signal* $\mathcal{S}$ wird über einen *Noisy Channel* vom *Transmitter* zum *Receiver* übertragen. Das empfangene Signal $\mathcal{S}'$ weicht je nach *Rauschpegel* des *Noisy Channels* vom Original unterschiedlich stark ab.

Das Modell des *Noisy Channel* kann ebenfalls auf die Übertragung von *Text Re-use* adaptiert werden. Der *Transmitter* ist hierbei der Autor aus der Antike. Der *Receiver* stellt den in der Gegenwart noch erhaltenen Text dar. Der *Noisy Channel* repräsentiert jede Form der eingangs genannten Änderungen (vgl. auch die Kapitel 2.6 und 2.8). Zusätzlich zu diesen eher editorischen Aspekten soll ein *Transmitter* nicht nur als ein ganzes Werk betrachtet werden, sondern auch deutlich kleinere Textpassagen, wie Sprüche, Phraseologismen und Redewendungen. Die möglichen Veränderungsoperationen, wie *Omissions* oder *Deletions* (vgl. Abschnitt 2.6), bleiben die gleichen. Aufgrund der starken zeitlichen Ausdehnung kann der *Noisy Channel* heute selbst noch ein Signal an den *Receiver* übertragen, welches bereits

[33] Ungewöhnliche Schreibweisen sind oftmals gute Indikatoren für bestimmte Autoren, bestimmte geographische Orte oder auch Epochen. Daher wird in den Geisteswissenschaften immer eine möglichst originalgetreue Version des Textes bevorzugt.

nicht mehr existiert, wie die *fragmentarischen Autoren* (vgl. [Berti 2009, Berti 2012])[34] oder auch Philosophen wie *Sokrates*, die nie ein Wort aufgeschrieben haben und nur noch aus dem geschriebenen *Text Re-use* von Autoren wie *Platon* überliefert worden sind.

Mit der wissenschaftlichen Einbettung des *Historical Text Re-use* gehen zahlreiche neue Fragen einher:

1. Wie ähnlich müssen sich zwei *Re-use Units* sein, damit das gesendete und empfangene Signal auch als hinreichend ähnlich erkannt werden kann?

2. Wenn ein gesendetes und empfangenes Signal als hinreichend ähnlich erkannt werden kann, welche systematischen Veränderungen im *Noisy Channel* können extrahiert werden?

3. Wie "gut" werden im Sinne eines *Turingtests* (vgl. Abschnitt 4.5) zwei Textstellen aufeinander gelinkt, wenn sich ein übertragendes Signal mit einem künstlichen *Noisy Signal* interferiert?

Der ersten Frage nach Mindestähnlichkeit zweier *Re-use Units* kann sich zweigeteilt genähert werden. Seitens der Geisteswissenschaften ist die Antwort von den eingangs erklärten Veränderungen durch Abschreiber und Editoren bzw. dem Grad der Verfremdung durch den *Syntactic Re-use*, *Semantic Re-use* oder *Cognitive Re-use* bestimmt (vgl. Abschnitt 2.6). Aus der Sicht der Informatik können solche Veränderungen als eine Art *Levenshtein-Distance* (vgl. [Levenshtein 1966]) verstanden werden. Vielmehr ist die *Levenshtein Distance* eine Spezialform der *Conditional Kolmogorov Complexity* (vgl. [Kolmogorov 1963, Kolmogorov 1998, Li 2008, Fortnow 2001]). Die *Conditional Kolmogorov Complexity* stellt die Frage nach dem *minimalsten Programm*, welches einen Text $\mathcal{T}$ nach $\mathcal{T}'$ transformiert, was einer Systematisierung der Veränderungsoperationen des *Noisy Channels* entspricht (vgl. Kapitel 2.8).

Mit der Beschreibung des *minimalsten Programmes* im Sinne der *Conditional Kolmogorov Complexity* kann an dieser Stelle nicht nur ein Ausblick auf die Kapitel 2.8 und 5.5 gegeben werden, sondern zeigt auch die Rolle der *Conditional Kolmogorov Complexity* im Kontext der oben genannten zweiten Frage. Heutzutage wissen wir oft nicht mehr, wer welche Veränderungen vorgenommen hat. Jedoch ermöglicht das *minimalste Programm* zu analysieren, welche der dort aufgelisteten Operationen systematisch einem Autor, einem Abschreiber oder auch einem Editor zugeordnet werden können. So bewirkt die ganzheitliche Analyse des *Noisy Channel* mit der *Conditional Kolmogorov Complexity* völlig neue Fragestellungen. Neben den bereits einschlägig diskutierten Schreib- oder semantischen Varianten können auch Transmissionsfehler (vgl. *Error Mining*) oder auch speziell *Watermarks* von Autoren, Abschreibern und Editoren erkannt und aufgezeigt werden. Ein personenspezifischer *Watermark* wurde beispielsweise von Johann Sebastian Bach in all seine Werken eingebaut. In jedem Werk von Bach taucht die Notensequenz "*b a c h*" auf. *Digital Watermarks* werden immer öfter auch von Editoren eingesetzt, um ihre Rechte an bestimmten Texten zu sichern und ein unerwünschtes Kopieren anhand der im Text verborgenen *Watermarks* zu identifizieren. Bei Navigationssystemen sind diese *Watermarks* künstlich eingesetzte Orte, die es in der realen Welt nicht gibt. In Textdaten können *Watermarks* sowohl im XML (vgl. [Ng 2005]) als auch über eine *Zero-Watermarking* (vgl. [Jalil 2010]) im Text selber versteckt werden. Mit dem hier skizzierten Ansatz wäre es nebenbei auch möglich, ebensolche *digital* und *non-digital Watermarks* als Teil einer Veränderung im *Noisy Channel*

[34] *Fragmentarische Autoren* sind Autoren bzw. Werke, von denen in der Gegenwart keine Texte mehr erhalten sind. Die Existenz dieser Autoren ist nur deswegen bekannt, weil sie in noch überlieferten Texten zitiert worden sind (vgl. *Information Poverty* aus Kapitel 1.4).

und somit als Teil des *minimalsten Programmes* der *Conditional Kolmogorov Complexity* zu identifizieren. Dieser Prozess wird im Rahmen dieser Dissertation *Noisy Channel Mining* genannt.

Die dritte aufgezeigte Frage nach der *Mining Competence* eines Algorithmus auf künstlich verrauschten Daten impliziert das Interesse nach der Qualität im Sinne von *Negative Results* (vgl. Abb. 1.7). Oftmals wird im Kontext des *Text Mining* nur anhand eines *Gold Standards* ein *Benchmark* für ein Algorithmus durchgeführt. Erstens bleibt immer die Frage offen, wie gut der *Gold Standard* auf die *Digital Library* passt. Zweitens und wesentlich kritischer ist zu sehen, dass es im Bewusstsein der bereits ausführlich diskutierten *Diversity* (vgl. das Paradigma *ACID for the eHumanities* aus Abschnitt 1.5) faktisch keinen *Gold Standard* für das *Historical Text Re-use* geben kann, der einer gesamten *Digital Library* genügt.

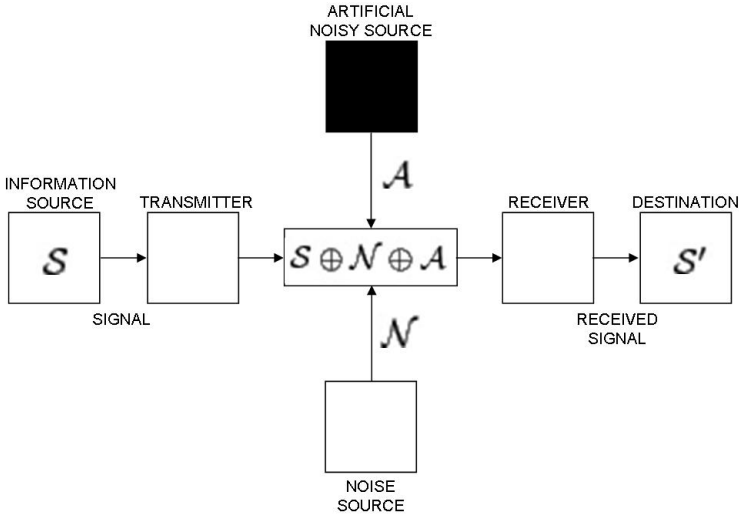


Abbildung 1.7: Shannon's *Noisy Channel* mit einem künstlichen Störsignal (schwarze Box): Ein *Signal* $\mathcal{S}$ wird über einen *Noisy Channel* zum *Transmitter* übertragen. Das empfangene Signal $\mathcal{S}'$ weicht je nach dem *Rauschpegel* des Störsignals des *Noisy Channels* vom Original ab.

Mit dem *Artificial Noisy Channel* soll das Paradigma der *Quantitative Evaluation* im Bereich des *Text Mining* eingeführt werden. Hierbei ist es nicht nur wichtig, ein gutes Ergebnis in einer *Qualitative Evaluation* gegen einen *Gold Standard* zu erzielen, sondern auch eine möglichst geringe Menge an *Mining*-Daten, im Sinne des *Historical Text Re-use* bei einer derartigen *Quantitative Evaluation* zu berechnen. So werden auf diese Weise zwei gegensätzliche Kräfte bzw. Interessen abgebildet: das *Text Re-use Detection* vs. *Counter Text Re-use Detection*. Während ersteres das eigentliche Ziel ist, hilft die fehlende Angreifbarkeit im Sinne des *Counter Text Re-use Detection* dabei, die Algorithmen sukzessive stabil zu machen. Diese Methode wurde im Rahmen dieser Dissertation aus der Wechselwirkung zwischen Internetsicherheitsfirmen und Hackern angepasst. Letztlich ist es genau dieser wechselseitige Wettbewerb und nicht die IT- Firmen selbst, welcher den hohen Standard an IT-Sicherheit im Internet heutzutage gewährleistet. Auch muss im Kontext des *Counter Text Re-use Detection* ein Algorithmus noch sinnvolle Ergebnisse liefern, wenn er mit

einer textsynthetisch generierten *Digital Library*[35] im Sinne eines *Turing*-Tests absichtlich getäuscht werden soll[36] (vgl. Kapitel 5.1).

Zusammenfassend kann festgehalten werden, dass das *Noisy Channel Model* von Shannon sehr vielseitig einsetzbar ist. Einerseits zum *Noisy Channel Mining*, welches "systematisches Rauschen" in Form von menschlichen Veränderungen aus dem Kanal aufzeichnet. Andererseits um künstlich Rauschen dem Kanal hinzuzufügen, wodurch getestet werden kann, wie gut eine *Text Re-use Analysis* funktioniert, wenn das Signal kein *Text Re-use* enthält bzw. bis zu welchem Rauschlevel noch etwas gefunden werden kann. Dies ist bspw. in der Wechselwirkung zwischen OCR-Digitalisierung und jedweder Form von *Text Mining* von besonderem Interesse. Auf der einen Seite entfernt jeder OCR-Postprocessing-Schritt nicht nur OCR-Fehler, sondern auch Rechtschreibfehler sowie dialektische oder sprachevolutionäre Varianten, so dass eine eher passivere OCR-Postkorrektur wünschenswert ist. Auf der anderen Seite sind die meisten *Mining*-Techniken im Ansatz nicht stabil genug, um mit solch verrauschten Daten umgehen zu können.

Vielmehr ergeben sich gerade im Kontext der *Textkritik* durch das *Noisy Channel Mining* weitere attraktive Verwendungen dieses Modells. So kann der *Noisy Channel* Ansatz dazu eingesetzt werden, um festzustellen, welcher Autor in seinen Texten oftmals sehr nahe am Original zitiert bzw. in welcher Epoche dies zum Zitierstil gehört hat. Gerade durch die starke *Information Poverty* ergeben sich neue Synergien für die *fragmentarischen Autoren*, um die Vertrauenswürdigkeit eines Textzeugens entsprechend mathematisch zu gewichten.

Auch wenn der Fokus bzgl. des *Noisy Channel Mining* im Rahmen dieser Dissertation auf sprachliche und semantische Varianten aus Zeitgründen beschränkt bleiben muss, so sollte dieser Abschnitt nicht nur der Einbettung dieser Arbeit dienen, sondern er soll vielmehr die Grundlage für den Forschungsbereich des *Historical Text Re-use* mit all seinen Facetten darstellen.

1.8 Verwandte Themengebiete in der Informatik

Forschung ist nicht unabhängig von verschiedenen Einflüssen. Insbesondere angesichts des fächerübergreifenden Forschungsbereiches der *eHumanities* ist dies in besonderer Weise gegeben. Dieser Abschnitt soll dem Leser dazu dienen, noch einmal aus der Sicht der Informatik die *Diversity* durch verwandte Forschungsbereiche gänzlich erfassen zu können.

Die technisch stärkste Überlappung des *Historical Text Re-use* besteht mit der *Plagiarismusforschung*. Auf algorithmischer Ebene sind Plagiarismus-Techniken oftmals sehr auf *Duplicates* und *Near-duplicates* ausgelegt. Dies geht einfach mit der Zielstellung einher, einen Autor beweisführend des geistigen Diebstahls zu überführen. Bei einer Paraphrase oder Allusion ist dies bereits deutlich schwieriger. Der *Historical Text Re-use* hingegen ist oftmals wesentlich diffiziler durch unterschiedlich große *Re-use Units* wie ganze Sätzen oder auch nur wenige Wörter. Neben diesen eher technischen Überlappungen sind jedoch die Zielstellungen wesentlich anders gesetzt. Beim Plagiarismus kommt es nicht nur auf den eigentlichen *Text Re-use* an, sondern insbesondere auf die Frage, ob dieser *Re-use* als eben jener kenntlich gemacht worden ist. Beim *Historical Text Re-use* hingegen steht das *Cultural Heritage* im Vordergrund. So kann es bspw. sein, dass es in bestimmten Epochen üblich war, den Ursprung im Sinne der Abgrenzung einer intellektuellen Elite nicht mit anzugeben und davon auszugehen, dass ein Gelehrter wissen muss, von wem der *Text Re-use* übernommen worden ist. Speziell in diesem Punkt unterscheiden sich der *Historical Text Re-use* und der *Plagiarismus* deutlich.

[35] bspw. aus der Kombination einer Markov-Kette mit einem $argmax$-Ansatz

[36] Siehe hierzu auch die aus der Biometrie adaptieren Qualitätskriterien aus Kapitel 2.4.

String-Suchalgorithmen (vgl. [Boyer 1976]) sowie Methoden und Ansätze der String-Ähnlichkeit (vgl. [Jin 2002, Bocek 2007]) werden benutzt, um bestimmte Textpassagen zu finden bzw. ähnliche Patterns zu matchen. Diese meist auf Buchstabenebene funktionierenden Algorithmen sind der Ursprung des *Text Re-use*. Diese Techniken wurden dann federführend von der Bioinformatik zum *Sequence Alignment* bspw. von DNA-Daten weiterentwickelt (vgl. [Higgins 2000]), welche noch heute oftmals auch im *Text Re-use* eingesetzt werden (vgl. [Olsen 2011]). In den 70er Jahren des 20. Jahrhunderts ging aus ursprünglichen Suchalgorithmen (vgl. [Boyer 1976]) auch der Forschungsbereich der *Text Compression* hervor (vgl. [Ziv 1977]). Letztlich bedeutet jeder in einer *Digital Library* gemachte *Re-use*, dass dadurch der Text komprimiert werden kann. Je größer der *Text Re-use* ist, desto größer ist auch das *Compression Ratio*, welches noch heute eingesetzt wird, um zu einer *Digital Library* den enthaltenen *Re-use* zu quantifizieren (siehe Abschnitt 3.10).

Im Kontext eines *Syntactical Text Re-use* gibt es Überlappungen zu den Forschungsbereichen der *Phraseologismen* (siehe u. a. [Cowie 1998]) bzw. der *Multi Word Expressions* (siehe u. a. [Dias 2005]). Während Phraseologismen expliziter Bestandteil des *Historical Text Re-use* sind, so sind *Multi Word Expressions*, wie *König Alexander der Große*, oftmals Teil des Ergebnisses einer *Text Re-use Analysis*, die jedoch eher als störend empfunden werden. Deren Separierung stellt eine offene Forschungsfrage dar.

Im Sinne eines *Semantic Text Re-use* gibt es inhaltliche Überlappung mit dem Forschungsbereich des *Topic Detection and Tracking* (vgl. [Allan 2002]), dem *Information Flow* (vgl. [Metzler 2005, Radford 2009]), dem *Passage Retrieval* [Tellex 2003] und der *Document Similarity* (vgl. [Lee 2005]). Im Wesentlichen kann der *Historical Text Re-use* von eben genannte Disziplinen durch die Größe des *Re-use Overlaps* bzw. der Größe der *Re-use Units* abgegrenzt werden. Er ist dadurch ausgezeichnet, dass der *Re-use Overlap* mindestens zwei, tendenziell drei bis vier, Wörter im Minimum umfasst. Bei der Größe der *Re-use Unit* reicht oftmals bereits eine satzsegmentierte Betrachtung aus. In manchen Fällen, wie dem *Text Re-use* in der Historiographie, ist auch ein größeres Fenster für die *Re-use Unit* angemessen.

Topic Detection and Tracking kann vom *Text Re-use* insofern abgegrenzt werden, als dass der *Overlap* zwischen zwei Textstellen nur ein Wort ist bzw. durch *Multi Word Expressions* auf einige wenige Wörter beschränkt bleibt. Beim *Text Re-use* hingegen ist der *Re-use Overlap* deutlich größer.

Information Flow, *Passage Retrieval* und *Document Similarity* unterscheiden sich vom *Text Re-use* durch die Fenstergröße der *Re-use Unit*. Während, wie bereits erwähnt, *Text Re-use* meist auf Satzebene bzw. in bestimmten Satzteilen beobachtet werden kann, sind die Beobachtungsfenster beim *Passage Retrieval* mit einer Betrachtung auf Absatzebene sowie dem *Information Flow* und der *Document Similarity* meist auf Dokumentebene anzusetzen. Einerseits kann bei dieser Betrachtung so der Eindruck entstehen, dass diese Themengebiete alle bis auf die Segmentierung stark voneinander abhängig sind. Andererseits muss sich auch vor Augen gehalten werden, was die jeweilige Zielstellung ist. Beim *Text Re-use* geht es um absichtliche und unabsichtliche Zitationsspuren[37]. Das Hauptziel des *Passage Retrieval* ist es, dass Ergebnis einer Suchmaschine zu verbessern. Oftmals werden dem Nutzer von Suchmaschinen Ergebnisse angezeigt, die zwar alle relevanten Keywörter beinhalten, welche sich jedoch über das ganze Dokument verteilen und somit nicht in einem Kontext stehen. Das *Passage Retrieval* kann als eine Spezialform des *Information Retrieval* verstanden werden, welches die Treffermenge auf Dokumente so reduziert, dass die Keywörter alle im gleichen Textabschnitt, der Passage, enthalten sind. *Document Similarity* hat im Gegensatz zum *Text Re-use* nicht das Ziel, festzustellen, ob bspw. ein Zeitungstext eine Dublette eines anderen Artikels ist, sondern inhaltlich gleiche Dokumente zu finden.

[37] Absichtliche Zitationsspuren sind beispielsweise *Zitate*, *Paraphrasen* und *Allusionen*. Unabsichtliche Zitationsspuren sind unter anderem *Geflügelte Wörter* und *Redewendungen*.

Auf der einen Seite ergeben sich durch diese Verwandtschaften interessante Möglichkeiten, zwischen den unterschiedlichen Disziplinen Erfahrungen auszutauschen. Auf der anderen Seite muss jedoch immer auch die Zielstellung eben diskutierter Forschungsbereiche im Blick behalten werden, was oftmals bedeutet, völlig unterschiedliche Interessen verfolgen zu müssen.

1.9 Ausblick und Gliederung der Arbeit

Die grundlegende und an das *ACID for the eHumanities* Paradigma angelehnte Fragestellung dieser Arbeit ist der Umgang mit der *Diversity* einer *Text Re-use Analysis*. Wie kann beim *Historical Text Re-use* mit den zahlreichen Schreibvarianten umgegangen werden? Weiterhin stellt das automatische Evaluieren von größeren Ergebnismengen einer *Text Re-use Analysis* unter der *Data Diversity* sowohl Gegenstand als auch eine Herausforderung dieser Arbeit dar. In diesem Sinne ist die Dissertation in fünf weitere Kapitel unterteilt.

Ziel des Kapitels 2 (*Grundlagen*) ist es, neben dem Aufstellen der grundlegenden Definitionen, die Umsetzung des *ACID for the eHumanities* Paradigmas im Rahmen dieser Arbeit aufzuzeigen. Abschnitt 2.6 reflektiert hierzu die *Diversity* des *Historical Text Re-use* anhand verschiedener Knoten- und Kantentypen eines *Re-use Graph*. Abschnitt 2.7 zeigt hingegen die *Complexity* auf. Das Kapitel 2 wird durch das *Vier-Sichten-Modell* der *Humanities*, *Digital Humanities*, *eHumanities* sowie der *Computer Science* auf den *Historical Text Re-use* ergänzt. Weiterhin werden aus der Biometrie in Abschnitt 2.4 Qualitätskriterien für die *Text Re-use Analysis* adaptiert.

Kapitel 3 (*Historical Text Re-use*) führt die *7-Level-Architektur* des *Historical Text Re-use* ein, die aus den Level *Segmentation* (vgl. Abschnitt 3.2), *Preprocessing* (vgl. Abschnitt 3.3), *Featuring* (vgl. Abschnitt 3.4), *Selection* (vgl. Abschnitt 3.5), *Linking* (vgl. Abschnitt 3.6), *Scoring* (vgl. Abschnitt 3.7) sowie *Postprocessing* (vgl. Abschnitt 3.8) besteht. Abgeschlossen wird das Kapitel von Abschnitt 3.10, in welchem aufgezeigt wird, wie die durch *Text Re-use* erzeugte Redundanz einer *Digital Library* ausgenutzt werden kann, um ein Maß für den enthaltenen *Text Re-use* durch die *Text Re-use Compression* zu bestimmen.

Kapitel 4 (*Zufall und Struktur*) hat mehrere Ziele im Rahmen dieser Arbeit. Einerseits wird in Abschnitt 4.2 aufgezeigt, dass es signifikante Nachteile von probabilistischen Sprachmodellen für den *Text Re-use* gibt, woraus in der Konsequenz die Motivation hinreichend gegeben ist, auf entsprechende statistische Modelle zu verzichten. Andererseits wird die auf die wissenschaftliche Einbettung des *Historical Text Re-use* in das *Noisy Channel Model* aufsetzende *Noisy Channel Evaluation* eingeführt (vgl. Abschnitt 4.4). Ziel dieser Technik ist es, die Fähigkeit einer *Text Re-use Analysis* und perspektivisch auch jedem anderen *Mining*-Verfahren zu messen, wie gut zwischen einer natürlichsprachlichen und einer zufälligen Struktur unterschieden werden kann.

In Kapitel 5 (*Ergebnisse*) werden die Resultate von unterschiedlichen *Text Re-use Analysis* reflektiert. Dies umfasst Ergebnisse sowohl auf Basis der *Perseus Digital Library* (vgl. Abschnitt 5.2) aber auch manuelle Selektionen von Wörtern an deutschen Redewendungen (vgl. Abschnitt 5.4.4). Das weitere Kapitel ist stark von einer *Text Re-use Analysis* auf sieben verschiedenen englischsprachigen Bibelversionen bestimmt. Die sprachliche Varianz reicht vom archaischen Englisch des 16. Jahrhunderts bis hin zu modernen Editionen der Gegenwart. Die *Text Re-use Analysis* wird in Kapitel 5 sowohl aus einer ganzheitlichen *System Evaluation* aber auch aus einer einzelnen *Component Evaluation* betrachtet, welche an die Vorgehensweise in der Biometrie angelehnt ist.

Kapitel 6 (*Zusammenfassung*) fasst nicht nur die Ergebnisse dieser Arbeit zusammen, sondern gibt insbesondere auch einen realistischen Ausblick darüber, was zukünftige For-

schungsschwerpunkte aber auch weiterführende Anwendungen sein können, wie die Adaption der *PageRanking*-Technik für einen *Text Re-use Graph* basierend auf sonst unstrukturierten Daten, um aus dem Text heraus ein *Cultural Heritage aware Ranking* abzuleiten.

Da mit dieser Arbeit das Forschungsfeld des *Historical Text Re-use* begründet werden soll, ist sie auf Grundlagen ausgerichtet. Hierdurch wird impliziert, dass sie niemals das Thema vollständig bearbeiten sondern vielmehr nur den ersten Schritt darstellen kann.

Kapitel 2

Grundlagen

Contents

2.1 Einführung **56**

2.2 Humanities, Digital Humanities, eHumanities, Computer Science: Das 4-Sichten-Modell des *Historical Text Re-use* **57**

2.3 Das 3-Generationen-Modell: Die Geschichte der Text Re-use Algorithmen **62**

2.4 Qualitätskriterien für Text Mining **62**

2.5 Grundterminologien und Definitionen **64**

2.6 Systematisierung des geisteswissenschaftlichen und informationstechnischen *Text Re-use* **68**

2.7 Text Re-use Tasks **79**

2.8 Noisy Channel Theorem und Conditional Kolmogorov Complexity **83**

Von einem Autor abzuschreiben ist ein Plagiat, von mehreren abzuschreiben ist Forschung.

Wilson Mitzner, (1876-1933)

Disziplinübergreifende Forschung bedarf einer Systematisierung der beteiligten Wissenschaften. Dieses Kapitel legt den Grundstein für den Forschungsbereich des *Historical Text Re-use* im Bereich der *eHumanities*. Das umfasst sowohl die Abgrenzung der Interessen der *Humanities*, *Digital Humanities*, *eHumanities* sowie der *Computer Science* als auch deren Wechselwirkungen. Damit geht insbesondere die *Complexity* und die *Diversity* (vgl. *ACID for the eHumanities aus Abschnitt 1.5*) einher.

2.1 Einführung

Text Re-use hat viele Gesichter. Wir alle generieren sowohl im privaten als auch beruflichen Leben Unmengen von Daten. Die meisten geschriebenen oder geäußerten Daten werden nicht wiederverwendet und haben einen reinen *Single Use*. Einerseits werden Daten, wie u. a. Berichterstattung über Events, sowohl heutzutage als auch im historischen Kontext, doppelt und dreifach produziert. Andererseits geht vieles von dem *Single Use* wieder verloren. Daraus ergibt sich die unmittelbare Frage, warum bestimmte *Text Chunks* wiederverwendet werden und andere nicht. In jedem Fall entstehen durch den *Text Re-use* in einer *Digital Library* teilweise enorme informationstechnische Redundanzen, was eine weitere Motivation darstellt, die Arbeit in Shannon's *A Mathematical Theory of Communication* (vgl. [Shannon 1948]) einzubetten.

Eine weitere Facette des vielseitigen *Historical Text Re-use*, und der damit verbundenen informationstechnischen Redundanz, ist die Größe des *Re-use*. Bereits in der Antike war man darum bemüht, gewisse textuelle Muster, die als erfolgreich und gut gegolten haben, wieder zu verwenden. Neben dem naheliegenden *Local Text Re-use* (vgl. [Seo 2008]), der durch ein *Quote*, eine *Paraphrase* oder eine *Allusion* gegeben ist, wurde insbesondere der größere *Template Text Re-use* bereits in der Antike bspw. in den *Delphischen Freilassungsurkunden* von Sklaven (siehe auch [Reinhold 2011]) oder in den dokumentarischen Papyri der *Duke Databank of Documentary Papyri* (siehe auch [NYUDL 2012]) eingesetzt. Hierbei wurden bereits sehr früh, ähnlich einem Serienbrief oder einem Urteil eines Gerichtes, vorgefertigte Vorlagen, die Schablonen, benutzt.

Sowohl im *Local Text Re-use* als auch im globaleren *Template Text Re-use* oder *Edition Text Re-use* sind die beiden Eigenschaften der *Locality* und der *Order* grundlegend. Ersteres beschreibt, dass bestimmte gemeinsame Wörter insbesondere beim *Template Text Re-use* und *Edition Text Re-use* lokal innerhalb eines kleinen Fensters zusammen auftreten. Ein gutes Gegenbeispiel ist ein Online-Artikel der *finanzen.net*[1]. Jener Artikel wurde von *Sven Reuse*[2] geschrieben, der zu Beginn des Artikels als Autor genannt wird. Am Ende des Artikels, also reichlich 700 Wörter später, taucht dann das Wort *Text* auf. Google Alert meldete dies auf *Text Re-use*, auch wenn der Artikel inhaltlich nicht einmal ansatzweise mit dem Thema verwandt ist. Dieses Beispiel soll zeigen, was die Konsequenzen für das Ergebnis einer *Text Re-use Analysis* sein kann, wenn die *Locality*-Eigenschaft unberücksichtigt bleibt (siehe auch Abschnitt 3.2). Die *Order* des *Global Text Re-use*, also u. a. dem *Template Text Re-use* und *Edition Text Re-use*, beschreibt, dass es in diesen Fällen immer parallele Sequenzen von *Text Re-use* in zwei digitalen Werken gibt.

Ziel dieses Kapitels soll es sein, die Grundlagen für die Forschung im *Historical Text Re-use* zu legen. Dies inkludiert neben einem Vergleich der Forschungen des *Text Re-use* in den *Humanities*, *Digital Humanities*, *eHumanities* und *Computer Science* (vgl. Abschnitt 2.2), den Aufbau einer *Re-use Taxonomy* (vgl. Abschnitt 2.6) als auch der Grundlagenarbeit im Bereich der *Re-use Terminology* (vgl. Abschnitt 2.5).

[1] vgl. `http://www.finanzen.net/nachricht/zinsen/DIPS-Kolumne-Zinsniveau-Quo-Vadis-1911519`

[2] Der Autor benutzt bisher immer die Schreibweise *Re-use* anstelle von *Reuse*. Dies liegt einerseits im deutschen Nachnamen *Reuse* begründet. Andererseits ist der Autor auf Konferenzen mehrfach gefragt worden, was *Reu-se* bedeutet. Um sowohl zwischen der Silbentrennung als auch dem deutschen Nachnamen und dem *Text Re-use* eindeutig zu unterscheiden, hat sich der Autor entschieden, *Re-use* zu benutzten.

2.2 Humanities, Digital Humanities, eHumanities, Computer Science: Das 4-Sichten-Modell des *Historical Text Re-use*

In Abschnitt 1.3 wurden die *eHumanities* als eine neue und aufkommende Disziplin mit dem Hinweis eines aus den *Humanities* und der *Computer Science* zusammenwachsenden Forschungsbereiches eingeführt. Ziel dieses Kapitels ist es, aktuelle Arbeiten bzw. den Forschungsstand des *Text Re-use* in den *Humanities*, *Digital Humanities*, *eHumanities* und der *Computer Science* sowohl zu reflektieren, als auch Dependenzen zwischen jenen Bereichen aufzuzeigen. Die Abgrenzungen zwischen *Digital Humanities* und *eHumanities* sind nicht immer klar. In den meisten Fällen resultiert dies daraus, dass persönliche und politische Interessen die Definition beeinflussen. Arbeitet bspw. ein Geisteswissenschaftler bereits in den *Digital Humanities*, wenn er nur digitale Korpora bzw. computerbasierte Methoden benutzt? Wohl kaum, da sonst die unmittelbare Folge wäre, dass es keine *Humanities* mehr gibt, da jeder Fachwissenschaftler im 21. Jahrhundert mehr oder minder mit digitalen Daten und Methoden arbeitet bzw. arbeiten muss. Aufgrund der meist politischen Interessen hinter den Terminologien werden die vier Bereiche *Humanities*, *Digital Humanities*, *eHumanities* und *Computer Science* nicht nach Interessen, sondern nach Aufgaben bzw. den damit verbundenen Fragestellungen aufgeteilt, um den Forschungsstand sowie existierende Projekte in diesem Sinne ansiedeln zu können. Das damit verbundene Interesse ist, sich von politischen und persönlichen Zuordnungen zu lösen. Eine unmittelbare Konsequenz ist, dass der Autor das Interesse verfolgt, weg von "ich bin *Digital Humanist*, weil" und hin zu "gemäß meiner Aufgaben arbeite ich in den *Humanities*, *Digital Humanities*, *eHumanities* und *Computer Science* mit folgender Verteilung" zu kommen.

Abb. 2.1 zeigt den Forschungsbereich des *Historical Text Re-use* im jeweiligen Kontext der vier Forschungsbereiche *Humanities*, *Digital Humanities*, *eHumanities* und *Computer Science* auf. Der innere Kern definiert hierbei die Hauptaufgaben, die mittlere Scheibe daraus resultierende Aufgaben und die äußere Scheibe ordnet gemäß ihrer Ausrichtung einige relevante Projekte an.

Die *Humanities* arbeiten bereits seit Jahrhunderten im Bereich des *Historical Text Re-use*[3] (vgl. Ausführungen zur Abb. 1.5 auf Seite 41). Im Vordergrund steht in den *Humanities* eine fachwissenschaftliche Fragestellung, welche mit Hilfe historischer Texte versucht wird zu beantworten. Relevante fachwissenschaftliche Fragestellungen sind unter anderem

- **Textkritik**: Aufgrund dessen, dass speziell antike Texte nur in seltenen Fällen noch im Original vorliegen, existieren die meisten historischen Texte nur noch durch Abschriften. Durch jede Abschrift kommen verschiedenste Veränderungen in Form und Inhalt in die Kopie. Die *Textkritik* (vgl. [Maas 1960]) stellt die Frage danach, wie original ein Text ist und welche Textstellen in besonderer Weise verändert worden sind. Hierbei hilft *Text Re-use*, um entsprechende Parallelstellen zu finden (vgl. [Dover 1997, Dué 2009]). Je ähnlicher sich die Parallelstellen sind, desto mehr steigt die Wahrscheinlichkeit, dass ein Werk nahe am Original überliefert worden ist[4].

- Finden von **Parallelstellen**: Neben der *Textkritik* werden Parallelstellen auch benötigt, um die Authentizität bestimmter Informationen zu prüfen. So sind in der Historiographie verschiedene Parallelstellen notwendig, um von einer einzelnen Meinung über bspw. einen Krieg oder Konflikt hin zu einer intersubjektiven Ansicht über das Ereignis zu kommen. In [Bigwood 1983] wird über verschiedene Zeugen der Schlacht

[3] In den *Humanities* ist der Begriff des *Text Re-use* der *Intertextuality* ähnlich (vgl. [Allen 2011]).
[4] Die Textkritik ist Gegenstand des *Homer Multitext* Projektes (vgl. Abb. 2.1).

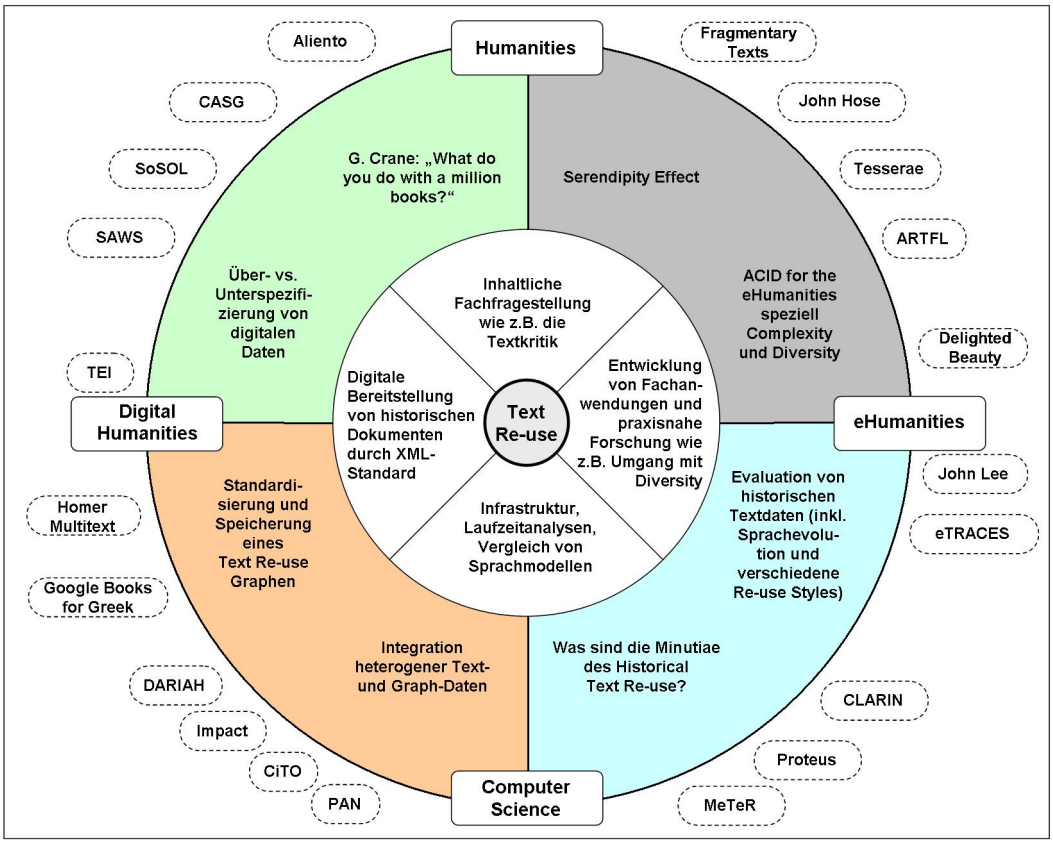


Abbildung 2.1: Vier-Säulen-Modell des *Historical Text Re-use*: Der Forschungsbereich ist durch die unterschiedlichen Fragestellungen der *Humanities*, *Digital Humanities*, *eHumanities* und der *Computer Science* definiert. Jeder dieser Bereiche hat unterschiedliche Fragestellungen an den *Text Re-use* (innerer Ring). Daraus ergeben sich vielfältige Wechselwirkungen zwischen den vier Bereichen (mittlerer Ring). Der äußere Ring ordnet existierende Projekte gemäß ihrer hauptsächlichen Fragestellungen an.

von *Kunaxa* zwischen *Cyrus* und seinem jüngeren Bruder *Artaxerxes II* in der Nähe von *Babylon* verglichen. Antike Zeugen sind neben dem griechischen Historiker *Xenophon*, der *Cyrus* auf seiner Reise nach Asien begleitet hat, auch *Sophaenetus*, *Ctesias* und *Diodorus*. In deren Texten werden Gemeinsamkeiten dieser Schlacht gefunden, jedoch weichen die Rahmenparameter, wie bspw. die Größe der Armeen, ab[5].

- **Lines of Transmission**: Bestimmte Informationen fließen über verschiedene kulturelle und sprachliche Grenzen bis hin zur Moderne. Diese *Lines of Transmission* aufzudecken, ist bspw. das Interesse vom *Corpus der Arabischen und Syrischen Gnomologien* (vgl. *CASG* in Abb. 2.1,[Pietruschka 2012]), *Sharing Ancient Wisdoms* (vgl. *SAWS* in Abb. 2.1, [Roueché 2010]) oder *Aliento* (vgl. *Aliento* in Abb. 2.1 [Varol 2010]).

[5]Dieses Beispiel wurde dankenswerterweise von meiner Kollegin Dr. Monica Berti recherchiert und bereitgestellt.

- **Nachwirkungen** von Autoren: Die Wichtigkeit bestimmter Autoren kann durch deren Nachwirkungen ausgedrückt werden. In diesem Kontext werden Zitationsverhältnisse analysiert und nach bestimmten Epochen klassifiziert[6] (vgl. [Büchler 2010d]).

- **Abhängigkeitsverhältnisse**: Texte werden nicht nur willkürlich, sondern oft auch systematisch, von einander abgeschrieben. So wird in [Lee 2007] das Verhältnis des Buches Markus und des Buches Lukas aus der Bibel miteinander verglichen. In [Hose 2004] werden die *Dead Sea Scrolls* mit einer hebräischen Version der Bibel verglichen. In [Pietruschka 2012] werden verschiedene Gnomologien, also Sammlungen von Sinn- und Weisheitssprüchen, auf Abschreibabhängigkeiten analysiert.

- **Fragmentary Texts**: Da im Laufe der Zeit viele Text verloren gegangen sind, existiert in der Gegenwart eine große Menge an *Incomplete Text Re-use*, welcher zwar ein *Target* jedoch kein *Source* mehr besitzt (vgl. [Berti 2009]). Diese Fragmente zu sammeln, kann der nachträglichen Rekonstruktion verloren gegangener Werke dienen.

- **Template Detection**: Bereits in der Antike wurden bestimmte erfolgreiche Textschablonen, die *Re-use Templates*, benutzt, um bspw. Freilassungsurkunden von Sklaven ganz im Sinne heutiger Serienbriefe zu erstellen (vgl. [Reinhold 2011]).

Das sind nur einige geisteswissenschaftliche Themen im Bereich des *Historical Text Re-use*. Komplementär dazu beschäftigen sich die *Digital Humanities* mit der Frage der digitalen Bereitstellung und Speicherung von *Text Re-use Traces* (vgl. Abb. 2.1). Die *Digital Humanities* sind mit der Verbreitung des Computers aus den *Humanities* hervorgegangen[7]. Dies beinhaltet unter anderem

- die **Digitalisierung** von historischen Daten durch OCR (vgl. *Google Books for Greek* und *IMPACT* in Abb. 2.1, [Boschetti 2009, Neudecker 2011]) bzw. die digitale Aufbereitung durch ein *Philological Crowd Sourcing* wie im *SoSOL*-Projekt,

- die **Digitale Konservierung** der Daten in ein maschinenlesbares Textformat wie TEI (vgl. [TEIC 2012]),

- die **Speicherung** von *Text Re-use Graphen* durch Technologien des *Semantic Web* (vgl. [Jordanous 2012]), Canonical Text Services (vgl. [Smith 2010]) oder innerhalb des TEI-Dokumentes selbst durch das *quote*-Tag (vgl. [TEIC 2012]).

Hierbei ist für die *Digital Humanities* nicht die Herkunft der *Text Re-use Data*, egal ob manuell oder automatisch gesammelt, von Interesse. Das können von Hand erstellte Annotationen (vgl. [TEIC 2012]), Daten aus einer *Crowd Sourcing* Umgebung (vgl. [Crane 2012, Beaulieu 2012]) aber auch Ergebnisse automatischer *Mining*-Techniken (vgl. [Olsen 2011, Büchler 2012c, Coffee 2012a]) sein. Bezüglich des *Historical Text Re-use* liegt der Fokus der *Digital Humanities* auf der Konservierung und Bereitstellung der *Text Re-use Data*, die dann Gegenstand von wissenschaftlichen Arbeiten der *Humanities* sein können. Die Schwierigkeit dieser vermeintlich einfachen Aufgabe kommt mit der Entscheidung, wo und wie öffnende und schließende *quote*-Tags für den *Text Re-use* gesetzt werden. Abb. 2.2 zeigt exemplarisch den Anfang eines Platon-Zitates (dritte Zeile). Diese Textstelle ist insgesamt sechsmal belegbar zitiert worden. Von diesen insgesamt sieben Textstellen gibt es, wie in Abb. 2.2 dargestellt, vier verschiedene Anfänge des Zitates. Das Entscheiden über das Setzen

[6] Siehe hierzu auch Abb. 1.4. Hier korrelieren die Zitationsverhältnisse mit dem Mittel- und Neuplatonismus.

[7] Noch heute ist oft beobachtbar, dass die meisten *Digital Humanists* einen geisteswissenschaftlichen Hintergrund haben.

von öffnenden und schließenden Tags ist somit oftmals ein Urteil darüber, ob ein bestimmtes Stoppwort hinzugenommen wird oder nicht. Die Aufgabe wird umso herausfordernder, wenn der *Text Re-use* paraphrasiert wurde.

αἱ μῆτραί τε καὶ ὑστέραι λεγόμεναι ...
αἱ δὲ ἐν ταῖς γυναιξὶν μῆτραί τε καὶ ὑστέραι λεγόμεναι ...
αἱ δ' ἐν ταῖς γυναιξὶν αὖ μῆτραί τε καὶ ὑστέραι λεγόμεναι ...
αἱ δ' ἐν ταῖς γυναιξὶ μῆτραί τε καὶ ὑστέραι λεγόμεναι ...

Abbildung 2.2: Es gibt sechs Parallelstellen zu Platon, Timaeus 91b7 ff, welche durch einen Algorithmus aufgedeckt werden konnten. Zu den insgesamt sieben Fundstellen im *Thesaurus Linguae Graecae* gibt es vier verschiedene Satzanfänge. Das erste Inhaltswort *metrai* (Gebärmutter) kommt in Platon erst an der siebten Stelle in der *Re-use Unit*. Die sechs Wörter davor wurden entweder angepasst oder in der Niederschrift entfernt.

Der Fokus der *Computer Science* bzgl. einer *Text Re-use Analysis* liegt auf dem Vergleich von *Text Re-use Language Models* (vgl. [Seo 2008]), effizientem Indexing von *Features* (vgl. [Huston 2011]), bestimmten *Selection*-Strategien (vgl. [Schleimer 2003]) oder auch flexibleren *Features* (vgl. [Siefkes 2004]). Speziell im Fokus von Infrastrukturinitiativen, wie *CLARIN-D*[8], sind auch *Laufzeitverhalten* sowie das Verhalten auf großen Datenmengen von Interesse. Eine *Brute-Force-Methode* des *Historical Text Re-use* wäre, jede *Re-use Unit*, bspw. jeden Satz, mit jeder anderen *Re-use Unit* zu vergleichen, welche von der Komplexitätsklasse $O(n^2)$ ist. Dieses quadratische Verhalten zeigt bereits das grundsätzliche Performanzproblem des *Text Re-use* auf. Selbst in optimierten Umgebungen, die nur diejenigen *Re-use Units* auf Ähnlichkeit überprüft, die mindestens ein gemeinsames *Feature* im *Re-use Overlap* beinhalten, bleibt die *quadratische Komplexität* erhalten. Eine unmittelbare Konsequenz daraus ist, dass das *Laufzeitverhalten* einer *Text Re-use Analysis* bei steigender Größe einer *Digital Library* nicht durch Parallelisierung signifikant verbessert werden kann[9]. *Text Re-use* auf *Big Scale Data*, wie dem Web, ist insofern eine eigenständige Herausforderung (vgl. [Bendersky 2009]). Die Frage nach einer performanten *Signatur* einer *Re-use Unit* und das damit verbundene Sprachmodell sind wesentlicher Bestandteil der Interessen der *Computer Science*. Für letzteres sei das *MeTeR*-Projekt[10] genannt, welches den *Information Flow* zwischen Nachrichten misst. Neben zahlreichen Publikationen und Programmen für die Plagiarismuserkennung wird von der Bauhaus-Universität Weimar jährlich die *PAN Challenge*[11] für Plagiarismus und *Text Mis-use* veranstaltet.

Das *Single Sourcing* wird oftmals bei technischen Dokumentationen eingesetzt. Das Ziel ist hierbei *Text Re-use* sukzessive so zu minimieren, dass es keine Redundanzen verschiedener Absätze in mehreren Texten gibt, sondern nur modulare Textbausteine, die für eine technische Dokumentation zusammengesetzt werden. Hierbei steht insbesondere der zusätzliche Wartungsaufwand, bedingt durch die Redundanzen, im Vordergrund.

Die *eHumanities* sind die jüngste der vier Disziplinen. Im Zuge der Massendigitalisierung kam letztlich in den Geisteswissenschaften mit der Frage "*What do you do with a million books?*" (vgl. [Crane 2006]) die Diskussion bzw. vielmehr das Bestreben auf, was nach der Massendigitalisierung und der XML-basierten Bereitstellung von historischen Dokumenten durch die *Digital Humanities* folgt und welche Konsequenzen sich daraus für die *Humanities* ergeben.

[8] vgl. `http://www.clarin.eu`
[9] Sofern die Menge der Rechner nicht auch quadratisch anwächst.
[10] MeTeR: Measuring Text Re-use
[11] vgl. `http://pan.webis.de/`

Das Bundesministerium für Bildung und Forschung (BMBF) fasst den Begriff der *eHumanities* in der 2011 erschienenen Bekanntmachung[12] wie folgt zusammen:

> *"... Die eHumanities verstehen sich als Summe aller Ansätze, die durch die Erforschung, Entwicklung und Anwendung moderner Informationstechnologien die Arbeit in den Geisteswissenschaften erleichtern oder verbessern wollen. ..."*

Um die vom BMBF angesprochene Erleichterung und Verbesserung in den Geisteswissenschaften zu erreichen, muss der Fokus der *eHumanities* auf einer praxisnahen Forschung liegen, die sowohl geisteswissenschaftliche Fragen beantworten (*Humanities*) und diese auf digitalisierten Daten der *Digital Humanities* anwenden kann als auch die Skalierbarkeit von Algorithmen auf *Big Scale Data* ermöglicht. Der Fokus der *eHumanities* im Bereich *Text Re-use* liegt auf dem Umgang mit der menschlichen *Diversity* (vgl. [Heyer 2011a, Olsen 2011, Coffee 2012a], vgl. auch die Projekte *eTRACES*, *ARTFL* und *Tesserae* in Abb. 2.1), die in einem starken Kontrast zu den *Language Models* der Informatik stehen, der Entwicklung fachwissenschaftlicher Anwendungen (vgl. [Heyer 2011a, Cheesman 2012]) sowie der Aggregation von verteilten Daten aus den *Digital Humanities*. Gerade letzteres wurde in Projekten wie LaQuat[13] oder Forschungsarbeiten (vgl. [Pansch 2010]) thematisiert. Unter diesem Aspekt sei noch einmal auf das *ACID for the eHumanities* Paradigma in Kapitel 1.5 auf Seite 38 verwiesen. Die *Diversity* des Paradigmas ist beim *Text Re-use* durch die unterschiedlichen *Text Re-use Styles* definiert. Die *Interoperability* entspricht den Arbeiten von *LaQuat* oder auch der Infrastrukturinitiativen *CLARIN*, *DARIAH* oder *Bamboo*. Die *Complexity* ergibt sich aus den Rahmenbedingungen der *Humanities* nach speziellen Anforderungen (siehe hierzu das noch folgende Kapitel 2.7). Während die drei genannten Säulen des *ACID for the eHumanities* Paradigmas aus konkreten Aufgaben bzw. Anforderungen resultieren, stellt die *Acceptance* der *Humanities* die "Königsdisziplin" dar. Um die *Acceptance* für quantitative Methoden, die *Big Scale Data* verarbeiten können, zu bekommen, muss sich sowohl das extrahierte Wissen mit dem existierenden Wissen decken als auch muss es einfach sein, neues Wissen zu erlangen. Im Kontext des *Text Re-use* heißt letzteres im Speziellen, dass es besonders einfach sein muss, neue und unerwartete Zitationsverhältnisse aufzudecken (vgl. *Uncovering Serendipity* in [Büchler 2013b]). Dies stellt eine Wechselwirkung zwischen den *Humanities* und den *eHumanities* dar.

Es ist offensichtlich, dass diese vier Säulen nicht autonom agieren, sondern einen vierdimensionalen Arbeitsbereich aufspannen. Für die *eHumanities* ergeben sich somit Wechselwirkungen zu den *Humanities*, *Digital Humanities* und der *Computer Science*.

Im Kontext des *Historical Text Re-use* ergeben sich im Wesentlichen zwei Abhängigkeiten mit den *Digital Humanities*. Die Massendigitalisierung durch OCR ist sehr fehlerbehaftet. Daraus ergibt sich die Frage, wie stark die digitalisierten Texte mit OCR-Fehlern behaftet sein können, damit eine *Text Re-use Analysis* durchgeführt werden kann. Weiterhin ist das automatische Taggen von *Text Re-use* im Dokument aufgrund der genannten Schwierigkeit des Bestimmens von Anfang und Ende nahezu unmöglich. Daher muss es das Ziel sein, die Datenmenge eines *Text Re-use Graphs* unter Berücksichtigung von *Identifiern*, wie den CTS-ID's, außerhalb der Texte, in einem geeigneten Graph-Format abzuspeichern (vgl. [Smith 2010]). Mit der *Computer Science* stehen die *eHumanities* beim Entwickeln von *Language Models*, die gegen die *Diversity* resistenter werden, in einer Wechselwirkung.

Gemessen an den in diesem Kapitel genannten Aufgaben (vgl. Abb. 2.1) der *Humanities*, *Digital Humanities*, *eHumanities* und *Computer Science* für den *Text Re-use* kann diese Arbeit wie folgt verstanden werden: 50% der Themen decken sich mit den Aufgaben der

[12] vgl. `http://www.bmbf.de/foerderungen/16466.php`
[13] vgl. siehe auch `http://laquat.cerch.kcl.ac.uk/`

eHumanities (siehe Kapitel 3 und 5), 30% mit Aufgaben der *Computer Science* (siehe Kapitel 4 und die Abschnitte 2.8 sowie 5.1), 15% mit Aufgaben der *Digital Humanities* (siehe Abschnitt 2.6) und zu etwa 5% deckt sich diese Arbeit mit den *Humanities*.

2.3 Das 3-Generationen-Modell: Die Geschichte der Text Re-use Algorithmen

Der Forschungsbereich des *Text Re-use* erstreckt sich nunmehr über drei Generationen von wissenschaftlichen Methoden, die systematisch bis in die 70er Jahre des 20. Jahrhunderts zurückverfolgt werden können. Methoden der *1. Generation* (Stringology), wie der *Boyer-Moore-Suchalgorithmus* (vgl. [Boyer 1977]), das Speichern in *Directed Acyclic Word Graph* (vgl. [Crochemore 1997]) bzw. Suffix-Trees (vgl. [McCreight 1976]) sowie die *Text Compression* (vgl. [Ziv 1977]) und den damit verbundenen theoretischen Arbeiten, wie die zur *Kolmogorov Complexity* (vgl. [Kolmogorov 1998]), basieren auf der Buchstabenebene (vgl. [Crochemore 2003]). Diese Verfahren ermöglichen einerseits ein *Fuzzy-Matching*. Auf der anderen Seite hat sich jedoch auch schnell herausgestellt, dass die Wahl eines Buchstaben als kleinste bedeutungstragende Einheit für den *Text Re-use* sehr ungünstig ist, da hierbei immer eine komplexe Information und deren Wiederverwendung gemessen werden soll. Durch das Ersetzen eines Wortes durch ein Synonym sind diese Methoden in ihrer Funktionalität beschränkt und motivieren vielmehr die *2. Generation* der linguistischen Verfahren, die seit etwa Mitte der 90er unter dem Einfluss von Arbeiten aus dem *Information Retrieval* entwickelt worden sind. Hierbei ist das kleinste Element nicht mehr ein Buchstabe sondern das Wort, welches in verschiedenen Flexionen sowie Synonymen im Text auftreten kann. Im Wesentlichen lassen sich diese Verfahren je nach Anwendungsszenario in syntaktische und semantische Verfahren einteilen (vgl. [Gaizauskas 2001, Clough 2002]). Auch wenn entsprechende Vorteile der linguistischen Algorithmen offensichtlich sind, so hängt deren Qualität deutlich stärker von der Qualität des *Preprocessings* (z. B. Satzsegmentierung und Tokenisierung) ab. Ferner ist die nützliche Eigenschaft des *Fuzzy-Matching* verloren gegangen. 2008 führten Seo und Croft (vgl. [Seo 2008]) die Methode der *Discrete Cosine Transformation* (DCT) ein, welche sowohl syntaktische als auch semantische Eigenschaften der *2. Generation* in einem Verfahren abbilden kann, aber auch die nützliche Eigenschaft des *Fuzzy-Matching* der Verfahren der *1. Generation* bewahrt. Für eine Weiterführung der DCT können *Diskrete Fourier Transformationen* sowie Verfahren der *Wavelet*-Technologie verwendet werden (vgl. [Büchler 2011a]).

2.4 Qualitätskriterien für Text Mining

In der Wissenschaft werden oft Objekte miteinander verglichen. Grundlagen hierfür sind immer wiederkehrende Strukturen, die sowohl durch die Natur als auch durch den Menschen gegeben sind. Diese Strukturen speziell in Texten zu identifizieren und zu extrahieren, ist Gegenstand des *Text Mining*.

In der Astronomie wird bspw. nach erdähnlichen Planeten gesucht. In der Bild- und Signalverarbeitung werden Photos oder Satellitenaufnahmen ausgewertet, um archäologischen *Re-use* (vgl. [Förtsch 2010]) oder die Bewegung von Eisplatten zu beobachten (vgl. [Clausi 2002]). Weiterhin werden biometrische Daten eines Menschen gemessen und gespeichert, um diese Person jederzeit eindeutig zu identifizieren (vgl. [Tuyls 2007, Maltoni 2009]). So können beliebig viele Messpunkte eines menschlichen Fingerabdruckes genommen werden. Jedoch stellt sich immer die Frage, ob jeder Messpunkt genauso gut und wichtig ist wie

der andere. So sind bei einem menschlichen Fingerabdruck Messpunkte, die zwei parallele Rillen darstellen, eher unwichtig. Messpunkte, die davon abweichen, wie Windungen, Falten oder Narben, sind deutlich charakteristischer.

Zurückkommend zum Thema dieser Arbeit soll an diesen Beispielen verdeutlicht werden, dass es je nach Fragestellung (Verfahren) und Beobachtungsmaterial (Daten) nie die beste oder eine gute *Re-use Analysis* geben kann, sondern dass es lediglich eine Menge von Methoden gibt, die je nach Forschungsfrage und Datenlage unterschiedlich gut funktionieren. Letztlich werden durch *Text Re-use* entweder syntaktisch bzw. semantisch auffällige Strukturen identifiziert, die wiederum durch die ökonomische Wiederverwendung in der Natur, aber auch durch den Menschen verursacht, auftreten. Im Rahmen dieser Arbeit stehen sowohl neben unterschiedlichen Fragestellungen (vgl. Abschnitt 1.9 und Kapitel 5) auch die Anwendbarkeit eines Verfahrens auf unterschiedlichen Textsorten im Vordergrund.

So kann einführend an dieser Stelle auf frühe Ergebnisse aus dem eAQUA-Projekt verwiesen werden. Mit dem eher für Gräzisten ausgelegten Verfahren des *Longest-Common-Ngram* (vgl. Kapitel 3.4), welches die längste gemeinsame Wortfolge bestimmt, konnten auf den philosophischen Texten in Platon's Timaeus selbst bei einer geringen Satzähnlichkeit von etwa 30% gute bis sehr gute Ergebnisse erzielt werden. Auf den historischen Texten der Atthidographen, die wiederum weltliche Ereignisse und Beobachtungen niedergeschrieben haben, hatte die gleiche Methode mit gleichen Parametern deutlich schlechtere Ergebnisse von etwa 20%[14].

Diese einführenden Beispiele zeigen die Komplexität des *Re-use* im Allgemeinen und des *Text Re-use* im Speziellen auf. Die grundlegende Fragestellung insbesondere im Kontext der *Diversity* menschlicher Sprache und der mit auf historischen Texten verbundenen *Sprachevolution* ist die, wie in einem solchen Umfeld etwas gemessen und evaluiert werden kann.

Daher sollen aus der *Biometrie* einige Qualitätsmerkmale für eine *Re-use Analysis* übernommen werden (vgl. [Jain 2005, Maltoni 2009]). Ziel dieser *Qualitätskriterien* ist es, das entsprechende Vorgehen, die Fragestellung aber auch einzelne Analysemethoden auf bestimmten Daten genauer zu beleuchten.

Insgesamt können acht[15] Qualitätskriterien[16] aufgestellt werden:

- *Acceptability* - Eine *Featuring*- und *Selection*-Methode muss von Dritten (Geisteswissenschaftlern) akzeptierbar sein. Sobald es Zweifel an den Techniken gibt, muss geklärt werden, ob diese Zweifel berechtigt sind bzw. mit welcher Wahrscheinlichkeit der Zweifel begründet ist.

- *Circumvention* - Eine *Text Re-use Analysis* muss möglichst robust gegenüber Attacken sein (vgl. *Text Re-use* vs. *Counter Text Re-use* aus Abschnitt 1.7). Bei einer *Text Re-use Analysis* gib es im Detail zwei wesentliche Aspekte. Zum einen bedeutet dies im Kontext der *Conditional Kolmogorov Complexity*, dass auch bei einem "größeren" *minimalen Programm*, welches die Veränderungen beschreibt, inhaltlich gleiche Textpassagen gefunden werden. Zum anderen soll eine *Text Re-use Analysis* auch robust gegenüber *Rauschen* sein (vgl. *Artificial Noisy Channel* in Abb. 1.7 aus Abschnitt 1.7). Das bedeutet, dass die *Re-use Analysis* nicht einen *Re-use Match* anzeigen soll, wenn es keinen *Re-use gibt* (Reduzierung der *False Positive*).

[14] Die Zahl ist nicht empirisch belegt, sondern Teil eines mündlichen Feedbacks. Vielmehr kommt es an dieser Stelle der Arbeit nicht auf die exakte Evaluierung an, da die Ergebnisse derart diametral auseinanderliegen, dass zu erwarten ist, dass eine qualifizierte Analyse dies nur bestätigen würde.

[15] Sieben der Kriterien sind aus der *Biometrie* (vgl. [Jain 2005, Maltoni 2009]) übernommen. Die *Selection* wurde vom Autor aufgrund ihrer zentralen Rolle mit aufgenommen.

[16] Qualitätskriterien sind in alphabetischer Ordnung aufgelistet.

- *Collectability* - Die relevanten *Features* eines *Fingerprints* müssen einfach erfassbar bzw. messbar sein. Besonders bei nur noch fragmentarisch erhaltenen Texten, wie Papyri oder diversen Inschriften, die nicht oder nur selten in der Vergangenheit dupliziert worden sind, fehlen heutzutage ganze Wort- oder Satzteile, so dass dementsprechend nicht jedes mögliche *Feature* gemessen werden kann.

- *Performance* - Die Performance umfasst sowohl die *Leistungsfähigkeit* als auch die *Geschwindigkeit* einer *Text Re-use Analysis*.

- *Permanence* - Die *Features* sollten resistent gegenüber Zeit sein. Im Kontext dieser Arbeit können sich Texte nicht mehr ändern, jedoch wurden sie über mehrere Jahrhunderte durch die *Transmission* mehr oder weniger verändert. Grundlegend soll jedoch bleiben, dass die Information im Kern gleich bleibt. Mit diesem *Qualitätskriterium* geht die Frage nach dem "Warum wird etwas *re-used*?" einher.

- *Selection* - Die *Signatur*, also eine Selektion bzw. Reduzierung aller messbaren *Features* eines *Fingerprints* einer *Re-use Unit*, muss diese bestmöglich repräsentieren und möglichst nah an den *Re-use Nucleus*, also der theoretisch bestmöglichen *Signatur* eines *Fingerprints*, herankommen.

- *Distinctiveness*[17] - Verschiedene unabhängige *Re-use Units* dürfen nicht die gleichen *Features* haben, sondern sollten sich bzgl. ihrer *Features* möglichst deutlich unterscheiden.

- *Universality* - Jeder Text bzw. jedes Werk sollte mit den gleichen *Feature Extraction*-Techniken, wie einer *Bigram*-Analyse, messbar sein. Das damit verbundene Ziel ist die Sicherstellung von globalen Eigenschaften der *Features*. Die *Universality* ist jedoch auf *OCR*-Massendaten bedingt durch Scanfehler eingeschränkt.

Im Rahmen dieser Arbeit wird immer wieder auf diese *Qualitätskriterien* verwiesen. Vielmehr werden verschiedene Analyseschritte auf der Basis dieser acht Eigenschaften untersucht.

2.5 Grundterminologien und Definitionen

Text Re-use ist der Forschungsbereich des Wiederverwendens von Textpassagen. Entgegen anderer Disziplinen, wie die *DNA-Analyse* oder auch Forschungen im Bereich des menschlichen Fingerabdruckes, ist die Forschung des *Text Re-use* im Vergleich weit zurück. Natürlich gibt es dazu zahlreiche Arbeiten, jedoch zeichnet sich diese Forschung dadurch aus, dass ein Algorithmus auf Daten angewendet und evaluiert wird. Im Vergleich zu anderen Disziplinen wird jedoch schnell deutlich, dass diese Forschungsbereiche viel besser strukturiert sind. So gibt es in der Forschung zum *menschlichen Fingerdruck* ein wohldefiniertes Glossar (vgl. [BIMA 2012, NSTC 2006]). Bei der Lektüre solcher Glossare wird schnell deutlich, wie unstrukturiert der *Forschungsbereich* des *Text Re-use* ist. Was sind bspw. *D-Points* oder *Minutiae* beim *Historical Text Re-use*? Oder: Wie können die zahlreichen Evaluierungen für den *Text Re-use* adaptiert werden?

Dieser Abschnitt führt die wesentlichsten Terminologien ein. Um diesen Abschnitt nicht 20 oder 30 Seiten stark werden zu lassen, liegt der Fokus lediglich auf den grundlegendsten Terminologien. In Kapitel 3 wird die Menge der Definitionen sukzessive erweitert.

[17] Dieses Qualitätskriterium wird in der Fachliteratur auch *Uniqueness* genannt. Da jedoch *Uniqueness* sehr restriktiv ist, wird im Rahmen dieser Arbeit die Terminologie *Distinctiveness* benutzt.

Die hierbei aufgebaute Menge an Terminologien hat im Wesentlichen drei Haupteinflüsse. Einerseits werden Terminologien der *Computer Science* benutzt, um Mining-Techniken bzw. mathematische Ausdrücke einzubringen. Terminologien aus den *Humanities* werden eingebracht, um Daten zu beschreiben. Terminologien aus der forensischen Biometrie werden schließlich adaptiert (vgl. speziell Kapitel 3), um vielen bisher nicht genau benannte Komponenten und Evaluierungen einen Namen zu geben.

Um eine *Digital Library* D_S einer Sprache S auf *Text Re-use* untersuchen zu können, muss ebendiese zu einzelnen *Re-use Unit*, wie in Definition 1, segmentiert werden (vgl. auch Abschnitt 3.2).

Definition 1 (Re-use Unit)**.** *Sei eine Digital Library D_S einer Sprache S gegeben. Eine Re-use Unit s_i ist die geordnete und vollständige Zerlegung einer Digital Library D_S mit $D_S = s_1 s_2 s_3 ... s_{n-1} s_n$.*

Hierbei wird *Text Re-use* zwischen den gewählten *Re-use Units* bestimmt. Eine Segmentierung einer *Digital Library* auf *Re-use Units* hat weiterhin die tiefere Intention, dass dadurch die Eigenschaft der *Locality* gewahrt wird. Dies bedeutet im Detail, dass auf möglichst geringem Raum eine möglichst große Dichte an gemeinsamen Merkmalen beobachtet werden kann (vgl. Abschnitt 1.8).

In diesem Sinne kann *Text Re-use* als eine Abbildungsfunktion ϕ_Θ zwischen zwei paarweisen *Re-use Units* s_i und s_j mit dem Parameterraum Θ wie folgt verstanden werden:

Definition 2 (Text Re-use)**.** *Seien zwei Re-use Units s_i und s_j der Digital Library D_S gegeben. Es liegt Text Re-use vor, wenn eine Methode ϕ_Θ existiert, die mit einem Parameterraum Θ die beiden Re-use Units s_i und s_j verlinkt und die Kante in die Menge E eines Re-use Graphs $G = (V, E)$ einfügt.*

Aus der Definition 2 können zwei Dinge abgeleitet werden. Im Sinne eines *Syntactic*, *Semantic* und *Cognitive Text Re-use* muss es nicht immer eine Abbildungsfunktion ϕ_Θ geben, die jede Form eines *Text Re-use* aufeinander projizieren kann. Insbesondere in Anbetracht der geisteswissenschaftlichen *Completeness* (vgl. Definition 18 auf Seite 124) ist dies zu berücksichtigen. Weiterhin impliziert die Definition auch, dass es sowohl automatische, semi-automatische aber auch manuelle Methoden bspw. durch ein *Philological Crowd Sourcing* geben kann [Dué 2009, Pietruschka 2012, Roueché 2010]. Letztlich muss es das Ziel sein, die größtmögliche *Completeness* erreichen zu können. So wird es für automatische Methoden der *Computer Science* auch in den nächsten 10 oder 20 Jahren noch eine Herausforderung bleiben, eine Funktion ϕ_Θ zu bestimmen, die einen *Cognitive Text Re-use*, wie bspw. "*Like will to like*" und "*Birds of same feather flock together*", systematisch aufeinander abbilden kann. Alle automatischen Methoden, einen *Text Re-use* ϕ_Θ festzustellen, werden in Definition 3 unter *Text Re-use Mining* zusammengefasst.

Definition 3 (Text Re-use Mining)**.** *Seien zwei verschiedene Re-use Units s_i und s_j gegeben. Text Re-use Mining umfasst alle automatischen Methoden, um zwei Re-use Units s_i und s_j durch syntaktische, semantische und stilistische Gemeinsamkeiten bzgl. des Parameterraumes Θ als ähnlich zu bestimmen.*

Durch die Segmentierung einer *Digital Library* D_S in verschiedene *Text Re-use Units* $s_1 s_2 s_3 ... s_{n-1} s_n$ kann ein *Text Re-use*, wie in Formel 2.1, als *Adjacency Matrix* A dargestellt werden. Hierbei steht ein Eintrag $A_{i,j}$ mit den Indexen i und j für die *Re-use Units* s_i und s_j. Während ein Wert von 1 in der *Adjacency Matrix* für einen *Text Re-use* ϕ_Θ steht, signalisiert eine 0, dass kein *Text Re-use* existiert bzw. identifiziert werden konnte.

Aus der *Adjacency Matrix* in Formel 2.1 können unmittelbar folgende Eigenschaften des *Text Re-use* abgeleitet werden:

- *Self Similarity*: Jeder *Re-use Unit* s_i ist in der *Adjacency Matrix* A mit A_{ii} zu sich selbst ähnlich bzw. identisch, wodurch in einer *Digital Library* D_S die Hauptdiagonale immer mit 1 gesetzt ist (vgl. Definition 2).

- *Symmetrie*: Die *Adjacency Matrix* A ist bzgl. der Hauptdiagonalen symmetrisch.

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \tag{2.1}$$

Durch die *Self Similarity* einer *Re-use Unit* mit sich selbst sind alle Elemente der Hauptdiagonale $A_{i,i}$ immer 1. In einer (n,n)-*Adjacency Matrix* sind somit immer, wie in Formel 2.2 dargestellt, anteilig $\frac{1}{n}$ aller Werte auf 1 gesetzt.

$$\frac{n}{n^2} = \frac{1}{n} \tag{2.2}$$

Auch wenn gilt

$$0 = \lim_{n \to \infty} \frac{1}{n}, \tag{2.3}$$

so stellt die *Self Similarity* ein *systematisches Grundrauschen* dar. Insbesondere im Kontext der *Randomised Digital Libraries* (vgl. Abschnitt 4.5), in welchen systematisch Eigenschaften einer Sprache S durch Zufall ersetzt werden, verfälscht die *Self Similarity* das Ergebnis dennoch. Auch in weiterführenden Analysen, wie der *Dotplotview* aus Abb. 1.2 auf Seite 35 oder der *Text Re-use Compression* aus Abschnitt 3.10, werden die Ergebnisse durch die *Self Similarity* verfälscht.

Aus diesem Grund wird die *Adjacency Matrix* A in eine *Re-use Adjacency Matrix* R ohne *Self Similarity* nach Formel 2.4 transformiert.

$$R = A - E \tag{2.4}$$

Die in Formel 2.1 abgebildete *Adjacency Matrix* kann durch Formel 2.4 in die *Re-use Adjacency Matrix* (Abk. *RAM*) in Formel 2.5 umgerechnet werden. Hierbei sind die Einträge $R_{i,i}$ nun immer 0. Besonders bei großen *Digital Libraries*, wie *Google Books*, mit mehreren Millionen *Re-use Units* wird auf diese Weise die Menge der *Re-use Data* deutlich reduziert.

$$R = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix} \tag{2.5}$$

Der Vorteil einer *Re-use Adjacency Matrix* aufgrund ihres *binären Inhaltes* ist, dass sie relativ kompakt gespeichert werden kann. Jedoch hat sie auch den Nachteil, dass nur binär zwischen 0 und 1 also *Non Text Re-use* und *Text Re-use* unterschieden werden kann. Weiterhin ist nicht jeder *Text Re-use* auch gleich gewichtig bzw. gleich gesichert.

Die θ-Funktion transformiert für jedes (s_i, s_j)-Paar eine *Re-use Adjacency Matrix* durch $L_{i,j} = \theta(s_i, s_j)$ in einer *Re-use Similarity Matrix* L wie in Formel 2.6 um.

$$L = \begin{pmatrix} 0 & 0 & \theta(s_1, s_3) & 0 & 0 & \cdots & \theta(s_1, s_{n-2}) & 0 & 0 \\ 0 & 0 & \theta(s_3, s_3) & 0 & 0 & \cdots & 0 & 0 & 0 \\ \theta(s_3, s_1) & \theta(s_3, s_2) & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & \theta(s_5, s_{n-1}) & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \theta(s_{n-3}, s_1) & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta(s_{n-1}, s_5) & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix} \tag{2.6}$$

Während die *Adjacency Matrix* A und die *Re-use Adjacency Matrix* R zur Hauptdiagonalen symmetrisch sind, ist dies bei der *Re-use Similarity Matrix* L von der Wahl des Ähnlichkeitsmaßes θ abhängig. Liegt θ ein symmetrisches Maß zugrunde, so ist auch L bzgl. der Hauptdiagonalen symmetrisch. Falls andernfalls jedoch θ nicht symmetrisch ist und $\theta(s_i, s_j) \neq \theta(s_j, s_i)$ gilt, dann ist auch L nicht symmetrisch (vgl. Abschnitt 3.7).

Angesichts großer *Digital Libraries* von teilweise mehreren Millionen *Re-use Units* s_i ist die Schreibweise und Speicherung des *Text Re-use* als *Adjacency Matrix* A, *Re-use Adjacency Matrix* R oder *Re-use Similarity Matrix* L aufgrund des quadratischen Speicherverbrauches oftmals nicht sehr effizient, weshalb die *Re-use Data* vereinfacht in einem *Re-use Graph* G dargestellt werden können.

Definition 4 (Re-use Graph)**.** *Sei eine Digital Library D_S einer Sprache S gegeben. Ein Re-use Graph $G_{D_S,\phi_\Theta} = (V_{D_S}, E_{D_S,\phi_\Theta})$ besteht aus der Menge $V_{D_S} = \bigcup s_i$ aller Re-use Units s_i einer Digital Library D_S sowie der durch ein Text Re-use ϕ_Θ mit dem Parameterraum Θ generierten Menge der assoziierten Kanten $E_{D_S,\phi_\Theta} \subset V_{D_S} \times V_{D_S}$ zwischen zwei Elementen $s_i \in V_{D_S}$ und $s_j \in V_{D_S}$ mit der Eigenschaft $i \neq j$.*

Während die Menge V der geordneten Menge der *Re-use Units* s_i einer linearen *Digital Library* D_S entspricht, wird die Menge E als die Menge der *Edges* oder *Relations* (s_i, s_j)

zwischen paarweise verlinkten *Re-use Units* s_i und s_j bezeichnet. Ein *Re-use Graph* G_{D_S,ϕ_Θ} stellt damit eine Hypertextstruktur auf linearen Texten dar (vgl. Anwendung eines *Text Re-use Graph* für die Adaption des *Google PageRanking* Algorithmus in Abschnitt 6.3).

Ein *Re-use Graph* ist immer ein *ungerichteter Graph*. Dies bedeutet, dass zwar die Ähnlichkeit zweier *Re-use Units* berechnet wird, jedoch ein *Re-use Graph* keine Auskunft per se darüber gibt, welche *Re-use Unit* einer *Edge* die andere wiederverwendet. Aus einem ungerichteten *Re-use Graph* einen gerichteten *Re-use Graph* zu machen, ist Gegenstand des *Direction Detection* (vgl. Abschnitt 2.7).

Der Vorteil des *Re-use Graph* G_{D_S,ϕ_Θ} im Vergleich zu einer Matrix-Schreibweise ist, dass sämtliche 0-Werte der *Sparse Matrix* nicht mitgespeichert werden müssen und sich somit der Speicherverbrauch deutlich reduziert. Wenn innerhalb einer *Digital Library* kein *Text Re-use* ϕ_Θ gefunden werden kann, so ist die Menge E leer, während in der Matrixschreibweise eine 0-Matrix generiert wird. Die Menge an *Edges* eines *Re-use Graph* ist durch dessen *Re-use Capacity* $\mathcal{K}$ wie in Formel 2.7 mit

$$\mathcal{K} = |(V_C \times V_C) \bigcap \cup(s_i, s_i)| \tag{2.7}$$

nach oben beschränkt. Dies entspricht einem voll vernetzten Graphen, in welchem jede *Re-use Unit* mit jeder anderen *Re-use Unit* verbunden ist. Zwischen diesen beiden Extrema quantifiziert die *Re-use Density* $\mathcal{D}(G_{D_S,\phi_\Theta})$ die Menge der *Edges* eines *Re-use Graph* G_{D_S,ϕ_Θ}.

Definition 5 (Re-use Density)**.** *Sei ein Re-use Graph* $G_{D_S,\phi_\Theta} = (V_{D_S}, E_{D_S,\phi_\Theta})$ *aus Definition 4 gegeben. Eine Re-use Density* $\mathcal{D}(G_{D_S,\phi_\Theta})$ *eines Re-use Graph* G_{D_S,ϕ_Θ} *entspricht dem relativen Verhältnis der gefundenen Edges in* E_{D_S,ϕ_Θ} *in Bezug auf die Re-use Capacity* $\mathcal{K}$*.*

Die Bestimmung der Dichte $\mathcal{D}(G_{D_S,\phi_\Theta})$ eines *Re- use Graph* $G_{D_S,\phi_\Theta} = (V_{D_S}, E_{D_S,\phi_\Theta})$ kann wie in Formel 2.8 bestimmt werden.

$$\mathcal{D}(G_{D_S,\phi_\Theta}) = \frac{|E_{D_S,\phi_\Theta}|}{|V_{D_S}|^2 - |V_{D_S}|} = \frac{|E_{D_S,\phi_\Theta}|}{\mathcal{K}(G_{D_S,\phi_\Theta})} \tag{2.8}$$

2.6 Systematisierung des geisteswissenschaftlichen und informationstechnischen *Text Re-use*

Text Re-use ist ein vielschichtig diskutiertes Thema sowohl in den *Humanities* als auch im Bereich der *Computer Science*. Während der Begriff *Text Re-use* aus der *Community* der *Computer Science* in die *eHumanities* eingebracht worden ist, sind Terminologien, wie *Intertextuality* (vgl. [Allen 2011]) oder *Hypertextuality* (vgl. [Riffaterre 1994]), in den *Humanities* sehr gebräuchlich und oftmals synonym benutzt, auch wenn speziell in den Literaturwissenschaften die *Intertextuality* als ein Oberbegriff von *Hypertextuality* aufgefasst wird. Gemeinsam haben jedoch alle drei Begriffe, dass wechselseitige Bezüge zwischen Texten bzw. Textstellen, den *Re-use Units*, beschrieben werden. Während in der *Computer Science* sowohl das Sprachmodell von Interesse ist, welches die besten Ergebnisse im Sinne von Evaluierungsmetriken, wie *Precision* und *Recall* (beide vgl. Abschnitt 4.3), liefert, als auch das Arbeiten mit einem *Re-use Graph* G_{D_S,ϕ_Θ} selbst, ist der Fokus der fachwissenschaftlichen *Humanities* deutlich vielschichtiger. Neben bereits erwähnten Anwendungen, wie der *Textkritik* oder dem Arbeiten mit *Parallelstellen* (vgl. Abschnitt 2.2), steht in den *Humanities* auch die Frage nach dem *Node Type* und *Edge Type* im Vordergrund. Für die

eHumanities wäre es wünschenswert, genau diese Typisierungsaufgaben automatisch bearbeiten zu können. Auch wenn das derzeit technisch noch nicht ausgereift ist, so helfen die in diesem Abschnitt genannten *Node Types* und *Edge Types*, die *Diversity* fachwissenschaftlicher und historischer Texte zu verstehen (vgl. *ACID for the eHumanities* Paradigma aus Abschnitt 1.5).

Ein *Re-use Graph* $G_{D_S,\phi_\Theta} = (V_{D_S}, E_{D_S,\Theta})$ besteht nach dessen automatischer Erstellung aus der ungetypten Menge der *Nodes* V_{D_S} sowie der *Edges* $E_{D_S,\Theta}$. Da die automatische *Typisierung* sich derzeit als schwierig gestaltet, haben die *Computer Science* ein sehr minimalistisches *Type System*. Die Probleme des *Typings* sind eng mit dem Qualitätskriterium der *Collectability* verbunden (vgl. Abschnitt 2.4). Wie kann bspw. entschieden werden, ob ein *Text Chunk* ein *Sprichwort*, eine *Lebensweisheit* oder ein *Witz* ist. Diese Typisierungen sind oftmals mit den kognitiven Fähigkeiten des Menschen verbunden und dementsprechend schwierig durch den Computer zu entscheiden. Ganz im Sinne des Qualitätskriteriums *Collectability* werden in den *Computer Science Node Types* und *Edge Types* gewählt, die auch mit hoher Sicherheit beobachtet und gemessen werden können.

Node Types in den *Computer Science* werden nach dem *Segmentation Criterion* erstellt. Texte können auf Satz-, Absatz-, Seite- oder Dokumentenebene segmentiert werden (vgl. bspw. [Hose 2004, Bendersky 2009]), so dass die Menge der *Node Types* sich auf

$$T_V = \{Sentence, Parapraph, Page, Document\}$$

beschränkt. Die Menge der *Edge Types* ist vergleichbar kurz. Sowohl Bendersky als auch Metzler unterscheiden die *Edge Types* nach der Ähnlichkeit zweier verlinkter *Re-use Units*. Bendersky macht hierbei drei Abstufungen (vgl. [Bendersky 2009]):

- $C3$: *Near-Duplicate*, wenn sich die Texte sehr ähnlich bzw. nahezu identisch sind,
- $C2$: *Text Re-use*, wenn Texte so paraphrasiert sind, dass sich die Wortstellung verändert hat, aber die meisten Wörter gleich geblieben sind[18],
- $C1$: *Topical Similarity*, wenn das Thema noch gleich geblieben ist.

Bei den Klassen $C2$ und $C3$ kann davon ausgegangen werden, dass aufgrund der starken Überlappung ein *Intentional Text Re-use* vorliegt, also ein *Text Re-use*, bei dem einer der beiden *Nodes* Vorlage für den anderen oder ein dritter und ggf. nicht mehr erhaltener *Node* Original für die beiden *Nodes* gewesen ist. Die Klasse $C1$ steht für entweder stark paraphrasierte Dokumente oder Dokumente, die das gleiche Thema behandeln, jedoch die Vorlage eines der beiden Werke nicht gesichert festgestellt werden kann.

Im Gegensatz zu Bendersky's *3-Klassen Edge Type System* führte 2005 Metzler die folgende sechsstufige Klassifizierung der *Edges* ein (vgl. [Metzler 2005]): *5 - Exact match*, *4- Minor revision*, *3 - Major revision*, *2 - Specific topic*, *1 - General topic* sowie *0 - Unrelated*.

Eine erste Überlappung zwischen den *Computer Science* und den *Humanities* kann bereits im Jahr 1976 weit vor den heutigen *eHumanities* ausgemacht werden. In jenem Jahr schrieb der Evolutionsbiologe Dawkins sein revolutionäres und im gleichen Maße umstrittenes Buch "The selfish gene" (vgl. [Dawkins 1976]). Hierbei fasst er die "egoistischen Gene" als in Konkurrenz um Ressourcen, hier die DNA, stehende Individuen auf. Dabei werden *Gene* als Objekte der Selektion verstanden. Er weist weiter darauf hin, dass der "Egoismus" der Gene selbst nicht festgestellt, sondern nur rückblickend durch eine DNA-Analyse

[18] Der Leser sei darauf hingewiesen, dass *Text Re-use* in [Bendersky 2009] anders verstanden wird. Während Bendersky *Text Re-use* nur als einen *Edge Type* versteht, wird unter *Text Re-use* sowohl in dieser Arbeit als auch in weiten Teilen der *Scientific Community* als Disziplin des Erkennens von ähnlichen *Text Chunks* verstanden.

bestimmt werden kann. Hierbei setzen sich bestimmte Gene signifikant messbar dominant und damit selektiv durch.

Auch wenn es auf den ersten Blick keine Überlappung mit dem *Historical Text Re-use* gibt, so ist dies dennoch in zwei Aspekten der Fall. Einerseits kann im Kontext des *Historical Text Re-use* ebenfalls zurückgeblickt und festgestellt werden, ob bestimmte *Re-use Chunks* deutlich erfolgreicher und selektiver bzw. im Kontext von Dawkins "egoistischer" sind als andere. Andererseits hat Dawkins einen grundlegenden Meilenstein mit der Terminologie eines *Meme* (vgl. Definition 6) auch für den kulturellen Bereich gelegt.

Definition 6 (Meme)**.** *Ein Meme ist die kleinstmögliche kulturelle Einheiten, wie ein Gedanke oder ein Symbol, die durch Kommunikation einer natürlichen oder formalen Sprache sowohl weitergegeben als auch verbreitet wird.*

Auf ein *Meme* wirken genau wie auf ein *Gen* sowohl *Mutation* als auch *Selection*. Die *Mutation* sind kleinere und größere Veränderung des *Meme*. Im Rahmen dieser Arbeit ist die *Mutation* in *Shannon's Noisy Channel* (vgl. Abb. 1.6 in Abschnitt 1.7 sowie im Abschnitt 2.8) durch die *Conditional Kolmogorov Complexity* nicht nur beschrieben, sondern auch nachträglich noch messbar. Im Sinne des *Historical Text Re-use* ist die Frage nach der *Selection* von *Meme* im Laufe der Zeit ungleich schwieriger. Was macht ein *Historical Meme* erfolgreicher bzw. in Dawkins' Sinn "egoistischer" als andere?

Es gibt verschiedene Gründe, warum *Meme* im Kontext des *Historical Text Re-use* erfolgreich weitergegeben werden. Wichtig ist hierbei, den Grund des *Text Re-use* zu verstehen. So sind in den *Humanities* bereits zahlreiche *Meme* wie *Koan*, *Mantra*, *Gnome*, *Maxim*, *Pangram* oder *Anagram* wohldefiniert. *Meme*, wie *Koan* oder *Mantra*, haben einen religiösen Hintergrund. Das *Maxim*, eine andere Spezialisierung eines *Meme*, beschreibt meist eine erfolgreiche persönliche Lebenslage. Das *Gnome* beinhaltet eine Weisheit bzw. eine historische Wertung. Ein *Pangram* ist ein spezielles *Meme*, welches seine Besonderheit dadurch erhält, dass es jeden Buchstaben des Alphabets beinhaltet. Weiterhin erhält ein *Anagram* seinen Mehrwert dadurch, dass die gleichen Buchstaben neu zu Worten arrangiert werden und somit die ursprüngliche Nachricht chiffriert wird (vgl. [Thomas 2000]). Ein *Palindrom* hingegen wird aufgrund seiner Eigenschaft verwendet, dass sich sowohl von links nach rechts als auch von rechts nach links gelesen das gleiche *Meme* ergibt. Bei einem *Template* wird eine Schablone eines fertigen Textes genutzt, um nur noch die Inhalte einzusetzen[19] (vgl. Abschnitt 2.2).

Jede dieser Spezialisierungen eines *Meme* bringt nicht nur die *Intentional Diversity* mit sich, sondern auch diverse Unterschiede bei messbaren Charakteristika (vgl. Abschnitt 2.4) wie

- *Verbreitungsart*: *Meme* können sowohl *syntaktisch fest*, d. h. wortwörtlich oder zumindest sehr nahe am Original, aber auch nur *semantisch*, d. h. mit der gleichen Trägerinformation, verwendet werden.
- *Absicht*: *Text Re-use* kann sowohl absichtlich, wie bei einem *Proverb*, oder nicht absichtlich, wie bei einem *Idiom*, wiedergegeben werden.
- *Länge des Text Re-use*: Es gibt kurzen *Re-use*, wie bspw. bei einem *Battle Cry*, von meist nicht mehr als vier Wörtern, aber auch größeren *Text Re-use*, wie bei den *Meme Template* oder *Edition*.
- *Literarische Klassifikation*: Die meisten *Intentional Text Re-use Types*, wie *Mantra*, *Proverb* oder *Battle Cry*, beschränken sich oftmals auf sehr spezifische literarische Bereiche wie *Philosophie*, *Religion* oder *Militär*, in denen sie eingesetzt werden.

[19] Dies ist einem Serienbrief ähnlich.

Node Type	**Verbreitungsart**	**Absicht vs. Allgemeinsprache**	**Länge des *Text Re-use***	**Literarische Klassifikation**	**Grad der Verbreitung**	**synchrone vs. asynchrone Kommunikation**	**geschriebene vs. gesprochene Sprache**
Adage	meist syntaktisch	nicht absichtlich, allgemeinsprachlich	kurz, meist kaum mehr als 10 Wörter	Philosophie, allgemeiner Wortschatz	meist liegt eine starke Verbreitung vor	asynchrone Kommunikation	sowohl geschriebene als auch gesprochene Sprache
Abstract	semantisch	absichtlich	meist mehrere 100 Wörter	keine Einschränkungen	meist liegt eine sehr geringe Verbreitung vor	asynchrone Kommunikation	meist geschriebene Sprache
Anagram	weder noch, Chiffrierung	absichtlich	variable Länge	keine Einschränkungen	meist liegt eine sehr geringe Verbreitung vor	asynchrone Kommunikation	geschriebene Sprache
Aphorism	meist syntaktisch	absichtlich	meist sehr kurz, nicht mehr als 7 Wörter	Philosophie	sowohl geringe als auch stärkere Verbreitung möglich	asynchrone Kommunikation	meist geschriebene Sprache
Apophthegm	syntaktisch	absichtlich	meist kurz, nicht mehr als 10 Wörter	meist Philosophie aber auch allgemeiner Wortschatz	sowohl geringe als auch stärkere Verbreitung möglich	asynchrone Kommunikation	meist geschriebene Sprache
Battle Cry	syntaktisch	absichtlich	sehr kurz, meist nicht mehr als 4 Wörter	Militär bzw. Historiographie	meist stärker verbreitet	synchrone aber auch asynchrone Kommunikation	gesprochene aber auch geschriebene Sprache
Bonmot	semantisch	absichtlich	meist kurz, nicht mehr als 20 Wörter	Alltag	meist nur geringe Verbreitung	meist synchrone Kommunikation	meist gesprochene Sprache
Cliché	semantisch	absichtlich aber auch allgemeinsprachlich	meist kurz, weniger als 20 Wörter	Alltag	meist starke Verbreitung	sowohl synchrone als auch asynchrone Kommunikation	meist gesprochene aber auch geschriebene Sprache
Definition	syntaktisch	absichtlich	meist nicht mehr als 15 Wörter	Mathematik, Physik, Astronomie, Philosophie	meist eine starke Verbreitung	asynchrone Kommunikation	geschriebene Sprache

Tabelle 2.1: *Node Types* für ausgewählte *Meme*: *Adage*, *Abstract*, *Anagram*, *Aphorism*, *Apophthegm*, *Battle Cry*, *Bonmot*, *Cliché*, *Definition*

Node Type	**Verbreitungsart**	**Absicht vs. Allgemeinsprache**	**Länge des *Text Re-use***	**Literarische Klassifikation**	**Grad der Verbreitung**	**synchrone vs. asynchrone Kommunikation**	**geschriebene vs. gesprochene Sprache**
Edition	syntaktisch	absichtlich	ganzes Dokument	keine Einschränkungen	meist eine schwache Verbreitung	asynchrone Kommunikation	geschriebene Sprache
Epigram	syntaktisch	absichtlich	meist kurz, weniger als 10 Wörter	Poesie, Alltag	sowohl geringe als auch stärkere Verbreitung möglich	asynchrone Kommunikation	geschriebene Sprache
Epithet	meist syntaktisch	absichtlich	meist sehr kurz, nicht mehr als 3 Wörter	Alltag, Philosophie, Historiographie	sowohl geringe als auch stärkere Verbreitung möglich	asynchrone Kommunikation	sowohl geschriebene als auch gesprochene Sprache
Epitome	semantisch	absichtlich	können bis zu mehrere 100 Wörter lang sein	keine Einschränkungen der lit. Klassifikation	meist sehr niederfrequent verbreitet	asynchrone Kommunikation	geschriebene Sprache
Equation	syntakisch	absichtlich	meist nicht mehr als 10 Wörter und Formelelemente	Mathematik	oftmals sehr stark verbreitet	asynchrone Kommunikation	geschriebene Sprache
Fact	semantisch	absichtlich	Länge kann variieren	meist Historiographie	kann teilweise sehr stark verbreitet sein	asynchrone Kommunikation	gesprochene als auch geschriebene Sprache
Flowery Phrase	syntaktisch	allgemeinsprachlich	meist sehr kurz, nicht mehr als 7 Wörter	Alltag	stark verbreitet	synchrone und asynchrone Kommunikation	gesprochene aber auch geschriebene Sprache
Gnome	semantisch	absichtlich	meist kurz, nicht mehr als 20 Wörter	Philosophie, Alltag	teilweise sehr starke Verbreitung	asynchrone Kommunikation	meist geschriebene aber auch gesprochene Sprache
Idiom	syntaktisch	meist allgemeinsprachlich	oftmals sehr kurz, nicht mehr als 5 Wörter	Alltag	stark verbreitet	asynchrone Kommunikation	gesprochene und geschriebene Sprache

Tabelle 2.2: *Node Types* für ausgewählte *Meme*: *Edition*, *Epigram*, *Epithet*, *Epitome*, *Fact*, *Flowery Phrase*, *Gnome*, *Idiom*

Node Type	**Verbreitungsart**	**Absicht vs. Allgemeinsprache**	**Länge des *Text Re-use***	**Literarische Klassifikation**	**Grad der Verbreitung**	**synchrone vs. asynchrone Kommunikation**	**geschriebene vs. gesprochene Sprache**
Joke	syntaktisch	absichtlich	meist kurz, nicht mehr als 20 Wörter	Alltag	meist starke Verbreitung	asynchrone Kommunikation	gesprochene und geschriebene Sprache
Koan	syntaktisch	absichtlich	meist kurz, nicht mehr als 20 Wörter	chinesische Philosophie, Mahayana-Buddhismus	sowohl geringe als auch stärkere Verbreitung möglich	meist synchrone Kommunikation	gesprochene und geschriebene Sprache
Law	syntaktisch	absichtlich	Paragraph, meist nicht mehr als 20 Wörter	Rechtssprechung	meist stark verbreitet	asynchrone Kommunikation	geschriebene Sprache
Legend	semantisch	absichtlich	meist mehr als 50 Wörter	meist Mythologie	meist nur lokale Verbreitung	asynchrone und meistens direkte Kommunikation	meist gesprochene Sprache
Loanword	semantisch	absichtlich	meist nur 1 oder 2 Worte	keine Einschränkungen	stark verbreitet	synchrone und asynchrone Kommunikation	meist gesprochene aber auch geschriebene Sprache
Mantra	syntaktisch, wiederholend, oft metrisch	absichtlich	oftmals auch mehr als 20 Wörter	Religion, Hinduismus, Buddhismus, Yoga	innerhalb der Klassifikation sehr stark verbreitet	asynchrone Kommunikation	gesprochene Sprache
Maxim	syntaktisch	meistens allgemeinsprachlich	meist nicht mehr als 10 Wörter	Alltag, Lebenskunde	stark verbreitet	asynchrone Kommunikation	gesprochene aber auch geschriebene Sprache
Meme	syntaktisch und semantisch	absichtlich oder allgemeinsprachlich	variabel	keine Einschränkungen	variabel	synchrone und asynchrone Kommunikation möglich	gesprochene und geschriebene Sprache
Metaphor	semantisch	absichtlich	meist kurz, nicht mehr als 30 Wörter	Philosophie, Lebenskunde	variabel	asynchrone Kommunikation	meist geschriebene Sprache

Tabelle 2.3: *Node Types* für ausgewählte *Meme*: *Joke*, *Koan*, *Law*, *Legend*, *Loanword*, *Mantra*, *Maxim*, *Meme*, *Metaphor*, *Motto*

Node Type	**Verbreitungsart**	**Absicht vs. Allgemeinsprache**	**Länge des *Text Re-use***	**Literarische Klassifikation**	**Grad der Verbreitung**	**synchrone vs. asynchrone Kommunikation**	**geschriebene vs. gesprochene Sprache**
Motto	syntaktisch	meist absichtlich	meist nicht mehr als 10 Wörter	keine Beschränkungen	oftmals sehr stark verbreitet	asynchrone Kommunikation	gesprochene und geschriebene Sprache
Palindrom	syntaktisch	absichtlich	meist zwischen 5 und 10 Wörter	unabhängig von der lit. Klasse	aufgrund ihrer Seltenheit meist stark verbreitet	asynchrone Kommunikation	geschriebene Sprache
Pangram	syntaktisch	absichtlich	meist zwischen 5 und 10 Wörter	unabhängig von der lit. Klasse	schwach verbreitet	asynchrone Kommunikation	geschriebene Sprache
Parable	semantisch	absichtlich	mehrere 100 Wörter möglich	Moral	schwach verbreitet	asynchrone Kommunikation	geschriebene aber auch gesprochene Sprache
Paroimia	syntaktisch	absichtlich	meist nicht mehr als 10 Wörter	antike Sprichwörter, antike Weisheiten	schwache Verbreitung	asynchrone Kommunikation	gesprochene und geschriebene Sprache
Phraseme	syntaktisch	oftmals allgemeinsprachlich	meist nicht mehr als 5 Wörter	keine Beschränkungen	oftmals stark verbreitet	asynchrone Kommunikation	gesprochene und geschriebene Sprache
Platitude	syntaktisch	absichtlich	meist kurz, nicht mehr als 20 Wörter	keine Einschränkungen in der literarischen Klassifikation	meist nicht sehr verbreitet	asynchrone Kommunikation	sowohl geschriebene als auch gesprochene Sprache
Proverb	syntaktisch	meistens absichtlich	oftmals nicht mehr als 10 Wörter	Lebensregel, Weisheit	oftmals sehr stark verbreitet	asynchrone Kommunikation	meist geschriebene aber auch gesprochene Sprache
Punch Line	semantisch, schwierige semant. Überlieferung	absichtlich	meist nicht mehr als 20 Wörter	keine Beschränkungen	schwach verbreitet	asynchrone Kommunikation	meist geschriebene Sprache

Tabelle 2.4: *Node Types* für ausgewählte *Meme*: *Motto*, *Palindrom*, *Pangram*, *Parable*, *Paroimia*, *Phraseme*, *Platitude*, *Proverb*, *Punch Line*

Node Type	**Verbreitungsart**	**Absicht vs. Allgemeinsprache**	**Länge des *Text Re-use***	**Literarische Klassifikation**	**Grad der Verbreitung**	**synchrone vs. asynchrone Kommunikation**	**geschriebene vs. gesprochene Sprache**
Quip	semantisch	absichtlich	meist nicht mehr als 7 Wörter	keine Beschränkungen	meist schwach verbreitet	asynchrone Kommunikation	meist gesprochene aber auch geschriebene Sprache
Rant	semantisch	allgemeinsprachlich	nicht selten mehrere 100 Wörter	Moral, Philosophie, Fokus auf der Antike	meist schwach verbreitet	meist synchrone Kommunikation	gesprochene Sprache
Saw	syntaktisch	meist absichtlich	oftmals nicht mehr als 10 Wörter	allgemein gebräuchliche Weisheit	teilweise starke Verbreitung	asynchrone Kommunikation	meist geschriebene aber auch gesprochene Sprache
Sententiae	syntaktisch	meist absichtlich	oftmals nicht mehr als 10 Wörter	Weisheit	schwache bis starke Verbreitung	asynchrone Kommunikation	geschriebene aber auch gesprochene Sprache
Simile	semantisch	absichtlich	meist nicht mehr als 10 Wörter	keine Beschränkungen	schwache Verbreitung	meistens asynchrone Kommunikation	geschriebene Sprache
Slogan	syntakisch	absichtlich	meistens nicht mehr als 10 Wörter	Militär, Erkennungszeichen	oftmals stark verbreitet	synchrone und asynchrone Kommunikation	meist gesprochene aber auch geschriebene Sprache
Template	syntaktisch	absichtlich	meistens kürzere Dokumente	oft in der Rechtsprechung, Urkunden, Verträge	teilweise stark verbreitet	asynchrone Kommunikation	geschriebene Sprache
Truism	syntaktisch	absichtlich	meistens nicht mehr als 10 Wörter	Alltag	teilweise stark verbreitet	asynchrone und synchrone Kommunikation	gesprochene und geschriebene Sprache
Wit	semantisch	absichtlich	meistens nicht mehr als 20 Wörter	Alltag	schwach verbreitet	synchrone Kommunikation	meistens gesprochene aber auch geschriebene Sprache

Tabelle 2.5: *Node Types* für ausgewählte *Meme*: *Quip*, *Rant*, *Saw*, *Sententiae*, *Simile*, *Slogan*, *Template*, *Truism*, *Wit*

- *Grad der Verbreitung*: Während die meisten *Intentional Text Re-use Types* oftmals relativ selten wiederverwendet werden, sind speziell *Idiom* und *Winged Words* sehr stark verbreitet. Der Grad der Verbreitung entspricht hierbei dem *Grad d* eines *Nodes* eines *Re-use Graphen* G_{D_S,ϕ_Θ}.

- *synchrone vs. asynchrone Kommunikation*: *Text Re-use* kann sowohl aus einer *synchronen Kommunikation*, wie beim *Meme Parole*, hervorgehen, bei welcher mindestens zwei Personen interagieren. Andererseits liegt eine *asynchrone Kommunikation* vor, wenn die Kommunikation unidirektional ist.

- *geschriebene vs. gesprochene Sprache*: Insbesondere im historischen Kontext mit hohen Kosten, um Texte aufzuschreiben, haben sich verschiedene *Meme* evolutionär herauskristallisiert, die meist gesprochen werden, wie bspw. das *Mantra*. Andere *Meme*, wie das *Proverb*, werden oftmals geschrieben.

Auch wenn noch weitere Charakteristika aufgestellt werden können, so ist bereits die *Diversity* der Daten offensichtlich. Selbst unter Annahme einer rein binären Unterscheidung zwischen den sieben genannten Charakteristika ergeben sich bereits 128 verschiedene Permutationen. In den Tabellen 2.1 bis 2.5 sind 45 der wichtigsten und dominantesten *Meme* des *Historical Text Re-use* aufgelistet und nach den eben eingeführten *Charakteristika* klassifiziert[20].

Während Bendersky und Metzler (vgl. [Bendersky 2009, Metzler 2005]) in der *Computer Science* eine drei- bzw. sechsstufige Klassifizierung der *Edge Types* vorschlagen, ist genau wie bei den *Node Types* auch die *Edge Type Diversity* in den Humanities deutlich größer. Das geisteswissenschaftliche Projekt *Sharing Ancient Wisdoms* (vgl. [Roueché 2010, Jordanous 2012]) definiert bspw. im Sinne des *Semantic Webs* Relationen, wie *isShorterVersionOf*, *isTranslationOf* oder *isVerbatimOf*. Diese Form eines *Edge Type Systems* ist noch relativ nah an den Ansätzen der Informatik.

In den fachwissenschaftlichen *Humanities* hingegen existieren weit komplexere Zusammenhänge zwischen zwei paarweise verlinkten *Re-use Units* (vgl. Abb. 2.3). Ausgangslage ist der *Edge Type Parallel Text*, der für mindestens zwei *Re-use Unit* und ganz im Sinne des *Historical Text Re-use* eine inhaltliche Mindestähnlichkeit beschreibt. Das in Abbildung 2.3 dargestellte *Edge Typing System* folgt dem Vorbild eines *Entscheidungsbaumes* mit den folgenden drei Entscheidungen:

- *Arten des Text Re-use* (*Types*): Es kann im Sinne des *Noisy Channels* sowie der *Kolmogorov Complexity* zwischen drei Arten des *Text Re-use* unterschieden werden. Einerseits kann ein eher wortwörtlicher *Re-use Style*, dem *Syntactic Text Re-use*, vorliegen. Andererseits gibt es den eher semantisch ähnlichen *Re-use Style*, dem *Semantic Text Re-use*. Ist die *Source* eines *Text Re-use* nicht Teil der *Digital Library*, dann spricht man vom *Incomplete Text Re-use*.

- *Größe des Text Re-use* (*Size*): Die Größe des *Text Re-use* hat zwei wesentliche Aspekte. Auf der einen Seite ist sie abhängig von der Größe der *Re-use Unit*. Auf der anderen Seite beschreibt sie auch den Größenunterschied zwischen der wiederverwendeten und der wiederverwendenden *Re-use Unit*, wie dies bspw. beim *Edge Type Summarizing* der Fall ist.

[20] Die Aufstellung und Klassifikation ist im Rahmen dieser Arbeit entstanden. Vielmehr wurde durch sie erst deutlich, dass es für einen *Text Re-use* nicht nur oftmals bereits einen Namen gibt, sondern vor allem auch wie divers das Thema des *Historical Text Re-use* ist.

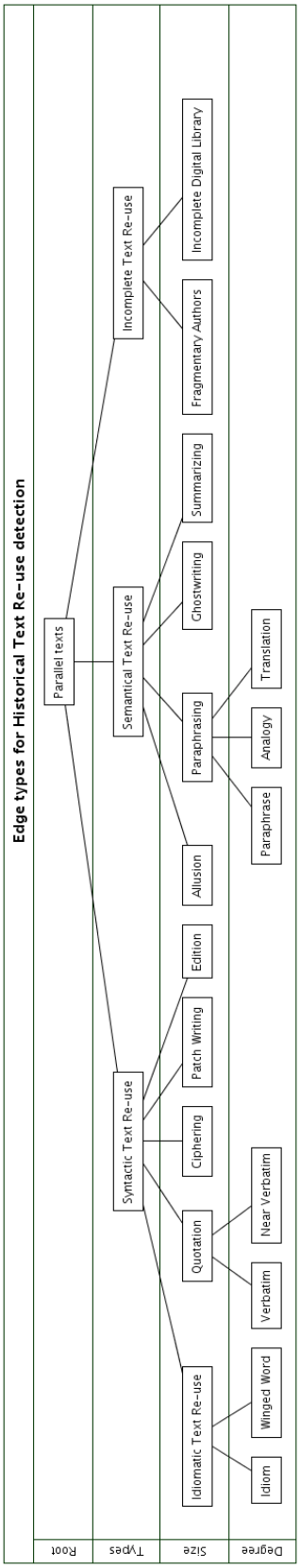


Abbildung 2.3: Ein fachwissenschaftliches *Edge Type System* auf Basis eines dreistufigen Entscheidungsbaumes. Die Wurzel ist *Parallel Text*. Die erste Unterscheidung wird nach dem *Type* des *Text Re-use* gemacht. Die zweite Unterscheidung richtet sich nach der Größe (*Size*). Die dritte Entscheidung ist vom Grad der Veränderung zwischen beiden *Re-use Units* bestimmt.

- *Grad der Veränderung* (*Degree*): Die letzte Entscheidung geht der Frage nach dem Grad der Veränderungen nach. *Edge Types* wie *Verbatim* oder *Near Verbatim* bleiben sehr nah am Original. Andere *Edge Types*, wie die *Paraphrase* oder die *Analogy* sind nur mit wesentlich mehr kognitiver Leistung erkennbar. Der *Edge Type Translation* hingegen paraphrasiert ein *Meme* in eine andere Sprache.

Sowohl die in diesem Abschnitt vorgestellten *Node Types* als auch *Edge Types* stellen die *Diversity* des *Text Re-use* auf historischen Dokumenten dar (vgl. *ACID for the eHumanities* Paradigma aus Abschnitt 1.5). Insbesondere für Infrastrukturprojekte wie *CLARIN*, *DARIAH* und *Bamboo* ist diese Vielfalt eine Herausforderung. Einerseits können automatische Methoden helfen, *Text Re-use* zu identifizieren. Andererseits ist eine automatische Analyse für stärker paraphrasierten *Text Re-use* als auch die *Typisierung* von *Nodes* und *Types* schwierig. In diesem Sinne ergänzen sich automatische Methoden und Systeme des *Philological Crowd Sourcing*. Während die automatischen Methoden den fachwissenschaftlichen *Serendipity Effect* (vgl. [Büchler 2013b]) unterstützen, bringt ein *Philological Crowd Sourcing* den Vorteil, die *Nodes* und *Types* zu typisieren.

Anhand dieser Typisierungen kann dann wiederum analysiert werden, welche Bestandteile eines *Meme* für welchen *Node Type* (vgl. *Signature* in Abschnitt 3.5) essenziell sind.

Die bereits erwähnte *Diversity* bedingt durch die *Node Types* und *Edge Types* sei noch einmal numerisch belegt. Allein die hier vorgestellten 45 *Node Types* können durch jedes der 15 Blätter des *Edge Type*-Baumes modifiziert werden. Damit ergeben sich allein in diesem Szenario bis zu 750 verschiedene *Meme-Re-use-Style*-Paare, die in einer *Digital Library* theoretisch beobachtet werden können. Daraus resultiert unmittelbar, dass in einem ersten Schritt zum Bestimmen von *Text Re-use* die enthaltenen *Meme* sowie die damit verbundenen *Re-use Styles* zu identifizieren sind (vgl. Abschnitt *Herausforderungen des textuellen Wissenstransfers auf geisteswissenschaftlichen Texten* in Abschnitt 1.6). Der Herausforderung durch die *Diversity* in den *historischen Daten* kann nahezu in keinem Fall mit nur einem Algorithmus entgegengetreten werden (vgl. hierzu Kapitel 3). Daher muss der geisteswissenschaftliche *Text Re-use* ϕ_Θ in n verschiedene *Re-use Styles* für unterschiedliche *Meme* zerlegt, separat analysiert und anschließend wieder zu einem *Hybrid Text Re-use* ϕ_Θ^H, wie in Formel 2.9, zusammengesetzt werden.

$$\phi_\Theta^H = \phi_\Theta^n = \bigcup_{i=1}^{i \leq n} \phi_\Theta^i \tag{2.9}$$

Definition 7 (Hybrid Text Re-use)**.** *Sei eine Digital Library D_S einer Sprache S sowie n verschiedene Text Re-use ϕ_Θ^1, ϕ_Θ^2, $\cdots$, ϕ_Θ^n gegeben. Ein Hybrid Text Re-use ϕ_Θ^H ist die Aggregation der Ergebnisse von mehreren Text Re-use nach Formel 2.9.*

Für den *Re-use Graph* $G^H_{D_S,\phi_\Theta}$, der aus dem *Hybrid Text Re-use* ϕ_Θ^H folgt, müssen noch die Mengen der *Re-use Units* $V^i_{D_S}$ sowie die Mengen der *Edges* $E^i_{D_S,\phi}$ betrachtet werden. Da der *Hybrid Re-use Graph* $G^H_{D_S,\phi_\Theta} = (V^H_{D_S}, E^H_{D_S,\phi_\Theta})$ in mehreren Analysen auf der gleichen *Digital Library* D_S bestimmt wird, gilt dementsprechend die Formel 2.10.

$$V^H_{D_S} = V^1_{D_S} = V^2_{D_S} = \cdots = V^n_{D_S} \tag{2.10}$$

Daher wird $V^H_{D_S}$ sowie $V^i_{D_S}$ auch einfach mit V_{D_S} abgekürzt. Für die Betrachtung der *Edges* $E^H_{D_S,\phi}$ aus einem *Hybrid Text Re-use* gilt

$$E^H_{D_S,\phi_\Theta} = \bigcup_{i=1}^{i \leq n} E^i_{D_S,\phi_\Theta} \tag{2.11}$$

mit der kleinen Kardinalitätseigenschaft

$$\forall i : |E^H_{D_S,\phi_\Theta}| \geq |E^i_{D_S,\phi_\Theta}| \tag{2.12}$$

sowie der großen Kardinalitätseigenschaft

$$|E^H_{D_S,\phi_\Theta}| \leq \sum_{i=1}^{i \leq n} |E^i_{D_S,\phi_\Theta}|. \tag{2.13}$$

Eine Analyse des *Text Re-use* auf einer *Digital Library* unter Berücksichtigung von Fragestellungen aus den *Humanities* unterliegt meistens einem komplexen *Hybrid Text Re-use*, der verschiedene *Meme* sowie *Re-use Styles* berücksichtigt. Daraus leitet sich die Terminologie *Text Re-use Analysis* ab.

Definition 8 (Text Re-use Analysis)**.** *Sei eine Digital Library D_S einer Sprache S sowie ein Hybrid Text Re-use ϕ^H_Θ gegeben. Eine Text Re-use Analysis ist der komplexe Vorgang eines Hybrid Text Re-use ϕ^H_Θ, welcher die in einer Digital Library existierenden Meme und Re-use Styles berücksichtigt.*

Eine *Text Re-use Analysis* setzt dementsprechend eine fundierte Kenntnis über die Daten voraus bzw. bedeutet, dass über eine Vielzahl von Experimenten die relevanten und dominanten *Node Types* sowie *Edge Types* aufwendig bestimmt werden müssen.

Ein abschließender Vergleich mit der *Biometrie* soll helfen, die Komplexität und Qualität einer *Text Re-use Analysis* im Vergleich zu einer *Biometric Analysis* zu verstehen. Bei bspw. einer *Biometric Human Fingerprint Analysis* kann mit modernen Methoden eine *Precision* von 97%-98% erreicht werden. Dazu kommen etwa 2% der Menschen, die meisten davon sind Chinesen, für die keine *Human Fingerprint Analysis* durchgeführt werden kann. Dies liegt einfach darin begründet, dass die *Biometric Features* aufgrund von schwachen Merkmalen nicht abgelesen werden können. Im Detail bedeutet dies, dass eine *Biometric Human Fingerprint Analysis* eine Fehlerquote von etwa 4%-5% hat. Ferner wird dadurch beschrieben, dass ein Fehler in etwa 20-25 Vergleichen auftritt, was für tyische Anwendungen inakzeptabel ist. Im Gegensatz zu den komplexen *Node Types* und *Edge Types* des *Historical Text Re-use*, sind die biometrischen Daten für einen Vergleich eines Fingerabdruckes relativ einfach strukturiert. So gibt es nur einen *Node Type*, wie etwa *Human Fingerprint*. Auch bei den *Edge Types* eines biometrischen Vergleiches ist eine derartige Analyse am ehesten mit *Verbatim* oder durch Veränderungen, wie Schnittwunden, mit *Near Verbatim* zu vergleichen. Aufgrund der deutlich größeren Komplexität des *Historical Text Re-use* sollte auch die realistische Erwartung im Vergleich zur Qualität von biometrischen Analysen richtig gewählt sein.

Das Typisieren, egal ob maschinell oder manuell durch bspw. ein *Philological Crowd Sourcing*, stellt bereits einen *Text Re-use Task* aus dem folgenden Abschnitt dar.

2.7 Text Re-use Tasks

Im vorherigen Abschnitt 2.6 lag der Fokus auf der *Diversity* des *ACID for the eHumanities* Paradigmas. Hierbei wurden bereits zwei *Text Re-use Tasks* erklärt: das Typisieren der *Nodes* und *Types*. Ferner sind *Text Re-use Tasks* bestimmte Aufgaben, die weit über den Horizont einer *Text Re-use Analysis* sowie fachspezifischen Anwendungen hinaus gehen. Ziel dieses Abschnittes ist es, die mit dem *Text Re-use* oftmals sowohl verbundenen als auch unsichtbaren Aufgaben zu definieren sowie deren Bedeutung im Kontext der *Complexity* des *ACID for the eHumanities* Paradigmas (vgl. Abschnitt 1.5) darzustellen.

Text Re-use Tasks können, wie in Abbildung 2.4 dargestellt, nach den vier Clustern *Algorithmic Re-use Detection*, *Graph Based Tasks*, *Semantic Detection* und *Noisy Channel Mining* gruppiert werden.

Das Cluster des *Algorithmic Re-use Detection* beschreibt die Menge der Aufgaben, die nötig sind, um eine *Text Re-use Analysis* durchführen zu können. Im Wesentlichen enthält dieses Cluster die beiden Aufgaben *Text Sort Detection* sowie *Re-use Style Detection*. Die erste Aufgabe ist nötig, um die Techniken gemäß den Gewohnheiten innerhalb einer Textsorte anzupassen. Die zweite Aufgabe hingegen hängt sehr stark von den im Abschnitt 2.6 bereits dargestellten *Node Types* und *Edge Types* ab. Die Umsetzung eines *Frameworks* für verschiedene *Re-use Styles* ist Gegenstand des Kapitels 3.

Ein *Hybrid Re-use Graph* ϕ_{Θ}^{H} besteht aus heterogenen *Re-use Data*. Das Cluster der *Graph Based Tasks* umfasst alle Aufgaben, die ϕ_{Θ}^{H} typisieren. Dieses Cluster kann in zwei verschiedene Subcluster gesplittet werden: den *Typing Tasks* sowie *Graph Properties*.

Typing Tasks haben zum Ziel, einen ϕ_{Θ}^{H} zu typisieren. Im Detail können diesem Subcluster die folgenden Aufgaben zugeordnet werden:

- *Re-use Boundary Detection*: Dieser *Task* beschreibt das Auffinden des Anfangs und des Endes eines *Text Re-use* (vgl. Abb. 2.2 aus Kapitel 2.2, Seite 57).

- *Node Type Detection*: Dieser *Task* typisiert die *Meme* nach ihren spezifischen *Node Types*, wie *Mantra*, *Pangram*, *Battle Cry* oder *Law* (vgl. Tabellen 2.1 bis 2.5 auf den Seiten 71 bis 75).

- *Edge Type Detection*: Dieser *Task* typisiert die Beziehung zwischen zwei *Meme*. Entsprechende *Edge Types* sind in Abbildung 2.3 auf Seite 77 in Kapitel 2.6 abgebildet.

- *Direction Detection*: Ein *Hybrid Re-use Graph* ϕ_{Θ}^{H} ist immer ungerichtet. Insbesondere bei einem *Intentional Text Re-use* ist jedoch eine Absicht beim *Text Re-use* gegeben, wodurch eine Richtung des *Re-use* abgeleitet werden kann. Hierfür können bspw. *Dating Information* helfen, jedoch reichen diese meistens nicht aus. Dies sei an einem kleinen Beispiel gezeigt: Es seien drei *Meme* A, A' und A'' gegeben, die alle drei bzgl. *Text Re-use* ähnlich sind. So können *Dating Information* helfen, dass A nicht von A' oder A'' abstammen kann. Dennoch reichen *Dating Information* nicht aus, um die genaue *Transmission Line* zu bestimmen. Im Beispiel würde dies bedeuten, dass *Dating Information* nicht dabei helfen, die Abstammung von A'' zu klären, da dies sowohl A als auch A' sein kann.

- *Intention Detection*: Dieser *Task* hat zwei Aspekte. Einerseits gilt es, die *Absicht* bzw. *Allgemeinsprachlichkeit* festzustellen. Auf der anderen Seite beschäftigt sich die *Intention Detection* mit der Frage, warum ein *Meme* wiederverwendet wird und andere nicht. Damit ein *Meme* wiederverwendet wird, hat es oftmals eine spezielle und neuartige Information. Dies können u. a. *Events*, wie Kriege, mathematische Ausdrücke, besondere Eigenschaften, wie die eines *Palindroms*, eine bestimmte *Weisheit* oder auch eine *moralische Folgerung*, sein[21].

- *Archetype Detection*: In Abschnitt 1.6 wurde bereits das Beispiel *comparing apples with oranges* angeführt, welches nur exemplarisch für die Interaktion aus gesprochener und geschriebener Sprache steht, wodurch das Bestimmen des *Archetyp* deutlich erschwert wird. Den Ursprung eines *Meme* zurückzuverfolgen, ist insbesondere auf Grund der sehr fragmentarisch erhaltenen Texte nicht nur herausfordernd, sondern mit konventionellen Methoden oftmals nicht mehr möglich.

[21] Auch wenn der Autor in diesem Bereich noch keine Untersuchungen angestellt hat, so geht er davon aus, dass es wahrscheinlich nicht mehr als 20 dieser *Intention Types* gibt.

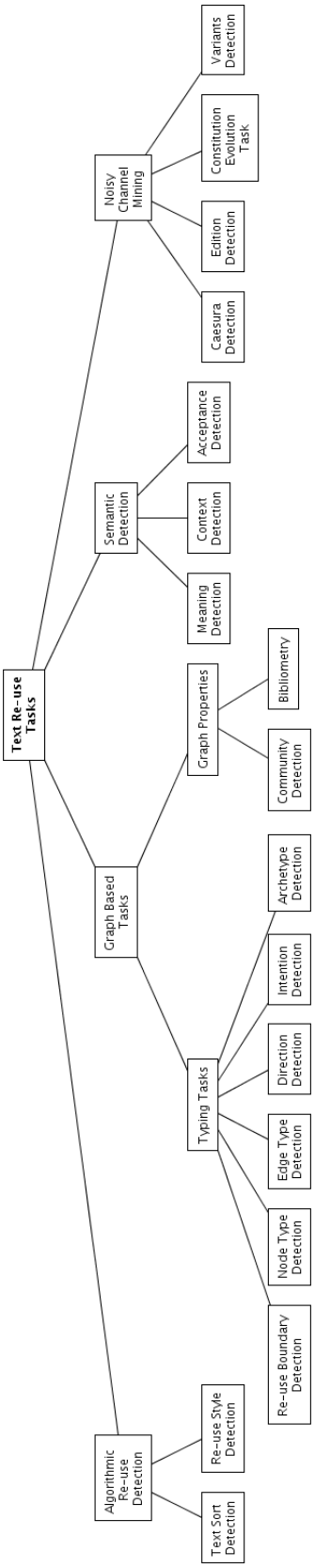


Abbildung 2.4: Ein Auflistung der wichtigsten *Text Re-use Tasks*. *Text Re-use Tasks* können nach *Algorithmic Re-use Detection*, *Graph Based Tasks*, *Semantic Detection* und *Noisy Channel Mining* geclustert werden.

Graph Properties wiederum haben zum Ziel, Eigenschaften eines *Hybrid Re-use Graph* ϕ_{Θ}^{H} zu bestimmen. Im Detail bedeutet dies, dass einerseits *Re-use Cluster* im *Community Detection* bestimmt werden können. *Re-use Cluster* können bspw. die Verweise auf Platon's Timaeus im *Mittel-* und Neuplatonismus in Abb. 1.4 sein. Auf der anderen Seite dient ein *Hybrid Re-use Graph* ϕ_{Θ}^{H} auch für die *Bibliometry*, welche die Zitationsabhängigkeiten auflöst, um einen *historischen H-Index* zu berechnen.

Das dritte Cluster, das *Semantic Detection*, umfasst Aufgaben, welche die semantische Bedeutung eines *Meme* bestimmen. Die Bedeutung eines *Text Re-use*, die *Meaning Detection*, ist oftmals, wie im Beispiel "jemanden ein Ohr abkauen", nicht direkt aus den enthaltenen Wörtern ableitbar. Ergänzend hierzu ist der *Task* der *Acceptance Detection* darauf orientiert, die positive oder ablehnende Benutzung eines *Text Re-use* zu messen. Dies geht mit der Annahme einher, dass Text nicht zufällig wiederverwendet wird, sondern damit entweder die eigene Arbeit unterstützt oder eine andere Aussage wiederholt werden soll, um dann widerlegt zu werden. Für die meisten *Meme*, wie *gleich und gleich gesellt sich gern*, ist es nötig, den Kontext zu kennen, um die Bedeutung verstehen zu können. Der *Task* des *Context Detection* beschäftigt sich daher mit der näheren Umgebung eines *Meme*.

Das vierte Cluster im Kontext eines *Historical Text Re-use* ist *Noisy Channel Mining*. Bei Aufgaben dieses Clusters werden Veränderungen durch die zeitliche Überlieferung analysiert. *Caesura Detection* beschäftigt sich mit allen Fragen, um systematische Veränderungen in der Schreibweise von Wörtern, wie durch eine *Spelling Reform* oder dem systematischen Wechsel *eines Re-use Styles*, festzustellen. So war es in bestimmten Zeiten nicht nur üblich, sondern auch guter rhetorischer Stil, dass ein *Meme* besonders genau und akkurat wiedergegeben wurde. In anderen Zeit war die Praxis des *Text Re-use* weniger präzise. Der *Variants Detection Task* wird mit den Arbeiten an der *Conditional Kolmogorov Complexity* abgedeckt. Historische Texte sind mit vielen historischen Varianten versehen, mit denen während einer *Text Re-use Analysis* umgegangen werden muss. Der *Constitution Evolution Task* fokussiert sich auf konstitutionelle Werke, wie einer Verfassung oder der Bibel. Je nach Version einer Verfassung und dem entsprechenden Zeitgeist verändert sich nicht nur die Anordnung der Paragraphen, sondern auch die Inhalte der Paragraphen selbst. An dieser Stelle sei auf das Beispiel mit dem Paragraphen 13 der deutschen Verfassung aus Abschnitt 1.6 verwiesen. Jener Paragraph lautet heutzutage "*Die Wohnung ist unantastbar.*", welcher bspw. während des *nationalsozialistischen Führerstaat* undenkbar gewesen wäre. Der *Constitution Evolution Task*, egal ob ein Verfassungs- oder religiöser Text, beschäftigt sich dementsprechend stark mit den Wertvorstellungen zu bestimmten Zeiten. Der *Constitution Evolution Task* ist ein beliebtes Forschungsthema speziell in den Politikwissenschaften.

Der letzte Task des *Noisy Channel Mining* ist das *Edition Mining*. Im Sinne der *Textkritik* bzw. der *Massendigitalisierung* können verschiedene Versionen ein- und desselben Textes in einer *Digital Library* enthalten sein. Diese zu identifizieren, ist nicht nur für eine *Text Re-use Analysis* von Bedeutung, sondern auch für jede Art der weiteren Verwendung einer *Digital Library* für maschinelle Lernverfahren, da sonst statistisch aus Dubletten gelernt werden würde.

Alle hier genannten und in Abb. 2.4 dargestellten Tasks repräsentieren die *Complexity* des *Historical Text Re-use* (vgl. *ACID for the eHumanities* Paradigma im Abschnitt 1.5). Die meisten dieser *Tasks* sind in der Forschung und von der *Scientific Community* bisher unberücksichtigt geblieben. Andere Tasks, wie die *Typing Tasks*, erfordern hingegen ein hohes Maß an kognitiver Leistung, so dass diese Aufgaben nur schwierig automatisch umzusetzen sind. Vielmehr wird durch die hier genannten Aufgaben der Forschungsaufwand deutlich, um den für alle Seiten bestmöglichen Kompromiss auf die Fragen zu finden, welche dieser Aufgaben manuell von Fachwissenschaftlern im Sinne eines *Philological Crowd Sourcing* gesammelt und welche Aufgaben durch einen Algorithmus abgebildet werden können.

2.8 Noisy Channel Theorem und Conditional Kolmogorov Complexity

In den bisherigen Abschnitten dieses Kapitels wurde bereits auf einige Aspekte der *Complexity* und der *Diversity* des *ACID for the eHumanities* (vgl. Abschnitt 1.5) eingegangen. Insbesondere die *Diversity* der *Node Types* und *Edge Types* wurde in Abschnitt 2.6 dargestellt. Diese Form der *Diversity* natürlichsprachlicher und vor allem auch historischer Texte stellt eine Herausforderung an eine *Text Re-use Analysis* dar.

Für den *Task* des *Re-use Style Detection* (vgl. Abb. 2.4 aus Abschnitt 2.7) ist eine sprachliche Stabilität nötig. So verringern etwaige Rechtschreibfehler, editorische sowie dialektische Veränderungen, aber auch Sprachevolution sowie die Anpassung an ein anderes Zielpublikum automatisch das Ergebnis einer *Text Re-use Analysis*[22]. In Abschnitt 1.7 wurde bereits die Forschung des *Historical Text Re-use* in Shannon's *Noisy Channel* $\mathcal{S} \oplus \mathcal{N}$ mit einem Signal $\mathcal{S}$ sowie der nicht näher bekannten Zusammensetzung von *Veränderungen* $\mathcal{N}$, dem *Noise*, eingebettet. Ziel muss es also sein, das *systematische* bzw. *nicht zufällige Rauschen* von einem Grundrauschen einer *Text Re-use Analysis* zu unterscheiden. Die Motivation hierfür kann aus dem *Birthday Paradox* ([Bloom 1973]) geschlossen werden. Das *Birthday Paradox* zeigt auf, dass um eine Kollision, d. h. zwei oder mehr Personen haben am gleichen Tag Geburtstag, zu erzeugen, bereits sehr wenige Daten ausreichen. Im Sinne des *Noisy Channel* sind Kollisionen jedoch nichts Negatives, sondern das wiederholte Auftreten eines *systematischen Rauschens* auf *Phonem*-, *Morphem*-, *Wort*- bzw. *Phrasen*-Ebene eines

- *Ancient Author*, der Wörter nicht richtig verstanden hat und dadurch einen semantischen Fehler in den *Text Re-use* ϕ_{Θ}^{H} einfügt,
- *Copyist*, der bis zum Buchdruck die Aufgabe hatte, historische Werke zu vervielfältigen und der je nach Qualifikation und verschiedenen Interessen in historischen Texten nur kleinere Abschreibfehler hinzufügte aber auch größere Veränderungen vornahm sowie
- *Editor*, der die historischen Texte insbesondere nach dem Buchdruck sowohl in Printmedien als auch digital aufbereitet. Hierbei können eingebaute Veränderungen auch nur auf technischen Einschränkungen basieren.

Um das *systematische Rauschen* von einem *Grundrauschen* einer *Text Re-use Analysis* (vgl. Evaluierung im *Noisy Channel* in Abschnitt 4.4) auch terminologisch deutlicher zu unterscheiden, sei die folgende Definition gemacht.

Definition 9 (Re-use Variant)**.** *Sei ein Hybrid Re-use Text* ϕ_{Θ}^{H} *mit einem Hybrid Re-use Graph* $G^{H}_{D_S,\phi_{\Theta}^{H}} = (V_{D_S}, E^{H}_{D_S,\phi_{\Theta}})$ *gegeben. Ein Re-use Variant* R *ist eine Insertion-, Substitution- oder Deletion-Regel, die zwei paarweise verlinkte Re-use Units* s_i *und* s_j *mit der Eigenschaft* $i \neq j$ *durch Anwendung ähnlicher macht.*

Die Systematisierung von Veränderungen im *Noisy Channel* kann aus drei verschiedenen Perspektiven betrachtet werden. Seitens der *Humanities* werden unter anderem in [Ernst-Gerlach 2008] *Re-use Variants* wie *Omission*, *Insertion*, *Substitution* sowie *Regular* und *Irregular text differences* als Textoperationen eingeführt. Aus der Sicht der *Computer Science* definieren Damerau und Levenshtein ebenfalls die Textoperationen *Omission*, *Insertion*, *Substitution* in den Metriken der *Damerau–Levenshtein Distance* [Damerau 1964] sowie der bekannteren *Levenshtein Distance* [Levenshtein 1966], um Sprachvarianten bzw.

[22] Der Umgang mit *Noisy Data* ist auch in der *Biometry* ein ganz spezieller Forschungsbereich (vgl. *Security with Noisy Data* in [Tuyls 2007]).

Rechtschreibfehler zu korrigieren. In der Psycholinguistik[23] wird wesentlich diffiziler aus der Sicht der Sprachproduktion unterschieden. So gibt es bspw. nicht nur *Substitutions*, sondern auch die Unterscheidung nach *Spoonerism* und *Malapropism*. Während im ersten Fall ähnlich klingende Wörter vertauscht werden, wird im zweiten Fall eine *Substitution* auf ein Wort oder eine Phrase gemacht, deren Bedeutung im Ursprungskontext keinen Sinn macht. Hierbei hat der verändernde Mensch eine falsche Bedeutung eines Wortes im Sinn. Dies ist insbesondere aufgrund der Zeitspanne von teilweise mehreren Jahrtausenden beim *Historical Text Re-use* nicht zu vernachlässigen.

Abb. 2.5 zeigt die Systematik der *Re-use Variants* unter Einflussnahme der *Humanities*, der *Psycholinguistik* als auch der *Computer Science*[24]. Sie stellt damit die *Diversity* des *Noisy Channels* im *Historical Text Re-use*.

Aus der Sicht der *eHumanities* stellt sich schließlich im Sinne der *Textkritik* die Frage, wie die Veränderungen eines Signals $\mathcal{S}$ nach $\mathcal{S}'$ systematisch aufgezeichnet und analysiert werden können. Die Anwendungsszenarien für eine Analyse der systematischen Veränderungen sind sehr vielfältig. Neben dem Erkennen von Sprachevolution, können so auch Zäsuren im Sinne einer Rechtschreibreform aber auch Änderungen eines Editors erkannt und verfolgt werden.

Die Metriken der *Damerau–Levenshtein Distance* [Damerau 1964] und *Levenshtein Distance* [Levenshtein 1966] können einerseits zwar die unterschiedlichen Operationen *Insertion*, *Substitution* und *Deletion* erkennen und dementsprechend verschieden gewichten. Jedoch ist sowohl eine genauere Unterscheidung als auch die mengenmäßige Beschreibung durchgeführter Textoperationen nicht bzw. schwer möglich. Die *Kolmogorov Complexity* kann als eine Generalisierung der *Levenshtein Distance* verstanden werden. Die *Kolmogorov Complexity* $K(\mathcal{S})$ ist eine algorithmische Beschreibung des minimalsten Programmes $\mathcal{P}_{min}$, welches ein Signal $\mathcal{S}$ beschreibt (vgl. [Li 1989, Fortnow 2001]). Die *Conditional Kolmogorov Complexity* $K_U(\mathcal{S}'|\mathcal{S})$ (vgl. Formel 2.14) beschreibt vielmehr die algorithmische Komplexität $\mathcal{P}_{min}$, welche Textoperationen nötig sind, um ein Signal $\mathcal{S}$ in ein Signal $\mathcal{S}'$ zu transformieren (vgl. [Faloutsos 2007]). $K_U(\mathcal{S}'|\mathcal{S})$ ist hierbei definiert durch

$$K_U(\mathcal{S}'|\mathcal{S}) = \min\{|\mathcal{P}| : \mathcal{P} \in 0,1^* \textit{ und } U(P,\mathcal{S}') = \mathcal{S}\} \tag{2.14}$$

mit den beiden Signalen $\mathcal{S}$ und $\mathcal{S}'$, der Menge aller Programme $\mathcal{P}$ sowie der *Universal Turing Machine* U, welches das Signal $\mathcal{S}$ in das Signal $\mathcal{S}'$ durch das binäre Programm $\mathcal{P}$ transformiert.

Auch wenn die Wahl der Programmiersprache für $\mathcal{P}$ im Rahmen der *Conditional Kolmogorov Complexity* theoretisch frei wählbar ist, wird im Rahmen des *Historical Text Re-use* eine künstliche Sprache eingeführt, die Ersetzungsregeln $R(s,s')$ nach den *Re-use Variants* aus Abb. 2.5 umsetzt. Eine Regel $R(s,s')$ ersetzt die in $\mathcal{S}$ enthaltene Sequenz s durch die in $\mathcal{S}'$ enthaltene Sequenz s' mit $s \neq s'$. Die geordnete Menge von Ersetzungsregeln $R(s,s')$ bilden ein Programm $\mathcal{P}$. Das binär kürzeste Programm ist $\mathcal{P}_{min}$. Bei der binären Übersetzung des Ersetzungstriples $R(s,s')$ kann die Länge für verschiedene R aus Definition 9 variieren. So können z. B. im Sinne eines *Huffman Codes* oder der *Entropy* frequentere Ersetzungsregeln mit einer geringeren Menge an *Bits* übersetzt werden, während seltenere Regeln dementsprechend mehr Bits benötigen, wodurch eine Gewichtung der Ersetzungstriples der verschiedenen *Re-use Variants* induziert wird.

[23] vgl. `http://en.wikipedia.org/wiki/Speech_error`

[24] Die Systematik beruht auf `http://en.wikipedia.org/wiki/Speech_error`, [Ernst-Gerlach 2008] sowie eigenen Ergänzungen.

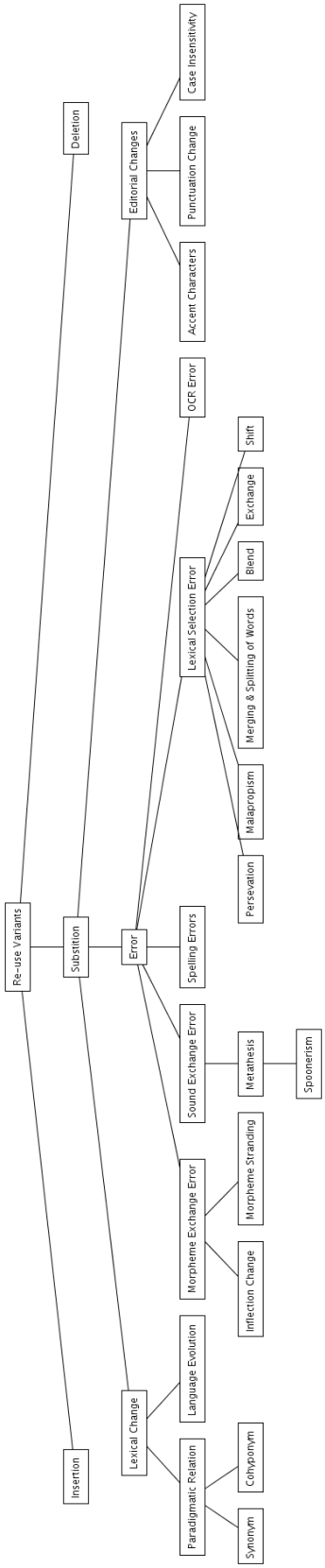


Abbildung 2.5: Systematisierung der *Re-use Variants*. In Anlehnung an die drei Textoperationen *Insertion*, *Substitution* sowie *Deletion* der *Levenshtein Distance* Metrik zeigt die Systematisierung diffizilere Textoperationen, die über ein *Philological Crowd Sourcing* gesammelt werden können.

Kapitel 3

Historical Text Re-use Detection

Contents

3.1 Einführung **88**
3.2 Level 1: Segmentation **91**
3.3 Level 2: Preprocessing **93**
3.4 Level 3: Featuring **100**
3.5 Level 4: Selection **104**
3.6 Level 5: Linking **113**
3.7 Level 6: Scoring **116**
3.8 Level 7: Postprocessing **120**
3.9 Wechselwirkungen zwischen den einzelnen Level **123**
3.10 Text Re-use Compression **124**

Think Big, Start Small.

Daniel Burnham, (1846-1912)

Dieses Kapitel führt in die technische Umsetzung des *Historical Text Re-use* mit der entsprechenden sprachlichen aber auch semantischen Vielfalt historischer Dokumente ein. Um dieser Vielfalt Herr zu werden, wird in diesem Abschnitt die *7-Level-Architektur* des *TRACER*-Tools vorgestellt. Ziel dieser Architektur ist es, auf die Vielfalt des *Historical Text Re-use* auch ein technisches *Framework* anzubieten, welches auf verschiedenste sprachliche Varianten, *Meme* und *Re-use Styles* entsprechend eingestellt werden kann. In diesem Kapitel wird jedes der Level in einem eigenen Unterkapitel aufbereitet, gefolgt von einem Abschnitt, der die wichtigsten Abhängigkeiten zwischen den *Level* zusammenfasst. Während der Fokus in den Abschnitten zu den einzelnen Level auf der fachwissenschaftlichen Umsetzung einer *Historical Text Re-use Detection* liegt, ist die Aufmerksamkeit im zusammenfassenden Abschnitt auf das Laufzeitverhalten gerichtet. Ziel hierbei ist es, Lösungen für den Kompromiss zwischen einer quantitativen *Text Re-use Analysis* und dem Laufzeitverhalten anzubieten.

In diesem Kapitel wird *TRACER* und seine wissenschaftlichen Techniken in einer Version vorgestellt, in welcher durch Kombination bereits über eine Million Permutationen von Techniken der *7-Level-Architektur* möglich sind.

Dieses Kapitel wird mit einem Abschnitt abgeschlossen, welcher die durch eine *Text Re-use Analysis* aufgedeckten *Redundanzen* dazu nutzt, eine *Text Re-use Compression* zu entwickeln. Ziel der *Text Re-use Compression* ist jedoch nicht die effizientere Speicherung von Daten, sondern mit dem eingeführten *Compression Ratio* eine quantitative Evaluierungsmöglichkeit zu schaffen, welche keinen qualitativen *Gold Standard* voraussetzt, da aufgrund der *Diversity* bestenfalls eine fragmentarische Richtigkeit dargestellt werden kann.

3.1 Einführung

Why do we quote? ist nicht nur Titel von [Finnegan 2011], sondern gibt auch bereits Aufschlüsse darüber, dass bestimmte *Re-use Units* bevorzugter wiederverwendet werden als andere. Die Frage nach einer entsprechenden Systematisierung dieser Warum-Frage gestaltet sich als ungleich schwieriger, da wie in Abschnitt 2.6 bereits aufgezeigt, eine Vielzahl verschiedener *Meme*, verschiedener *Re-use Styles* aber auch verschiedener Motivationen bzw. Intentionen oder den Wechsel des Zielpublikums für einen *Text Re-use* gegeben sind, die von Person zu Person, bzw. Textsorte zu Textsorte variieren können. Ziel dieses Kapitels ist es, das *TRACER*-Framework für *Historical Text Re-use Detection* vorzustellen. *TRACER* wurde mit den Erfahrungen aus den *eAQUA*- und *eTRACES*-Projekten entwickelt, um mit der genannte *Diversity* umgehen zu können.

Trotz der bereits vielfach genannten *Diversity* von Sprache und den daraus resultierenden natürlichsprachlichen Texten und Korpora gibt es auch einige grundlegende und gemeinsame Fragestellungen über die Grenzen der *Diversity* hinaus, die mit *Text Re-use* im Allgemeinen und *Historical Text Re-use* im Speziellen einhergehen. Grundlegend geht dies damit einher, dass danach gefragt werden muss, was ein gutes *Feature* ist (vgl. auch *Feature Engeneering* in [Turner 1998]).

Erstens jeder *Text Re-use*, das Wiederverwenden von etwas Erfolgreichem, Neuem bzw. für den Wiederverwendenden Interessantem, entspricht der Duplizierung einer Information, die die *Redundancy* in einer *Digital Library* erhöht. Die durch *Text Re-use* erzeugte *Redundancy* kann durch die *Text Re-use Compression* (kurz TRC) gemessen werden (vgl. Abschnitt 3.10). Je größer der *Text Re-use* in einer *Digital Library* ist, desto größer ist auch die *TRC*. Daher kann die *TRC* als ein quantitatives Maß für den *Text Re-use* verstanden werden (vgl. Abschnitte 3.10 und 5.3). Jedoch gibt es auch einige deutliche Einschränkungen. Je geringer die Schwellwerte eines *Mining*-Verfahrens gewählt werden, desto größer ist die *Text Re-use Compression*. Auf der anderen Seite gilt auch, je stärker eine *Digital Library* normalisiert wird, desto größer wird auch die *Text Re-use Compression*. Dies sei an einem einfachen Beispiel verdeutlicht. Es sei eine *Digital Library* mit n Tokens gegeben. Eine Normalisierung jedes Tokens durch eine Zeichenkette, wie *abcde*, erzeugt eine im Sinne der *Kolmogorov Complexity* minimale Beschreibung bzw. ein Programm $\mathcal{P}_{min} = (abcde)^n$, wodurch die *Text Re-use Compression* bei größer werdendem n gegen 1.0 konvergiert. Die *Text Re-use Compression* kann daher niemals eingesetzt werden, um zwei verschiedene *Mining*-Verfahren miteinander zu vergleichen, sondern lediglich zwei vergleichbare Teilschritte, wie bspw. der Einsatz einer *Lemmatisierung* im Vergleich zu *keiner Lemmatisierung* oder der Unterschied beim Einsatz eines *Bigram Shingling* im Vergleich zu *Trigram Shingling* (vgl. Abschnitte 3.4 und 5.3.1) .

Zweitens stellt sich die Frage danach, wie gut ein *ausgewähltes Feature* bzw. die daraus resultierende *Signature* einer *Re-use Unit* ist. Die *Signature* muss aus technischer Sicht sowohl alles beschreibend sein, aber zeitgleich auch so kompakt wie möglich. Jedes *Feature* einer *Signature* hat zeitintensive *Linking*-Kosten, um alle *Re-use Units*, die dieses *Feature* enthalten, miteinander zu verlinken. Aufgrund der quadratischen Komplexität $O(n^2)$ verursachen insbesondere frequentere *Features* hohe Laufzeitkosten. Speziell auf historischen Texten zeigt sich jedoch ein weiteres inhaltliches Problem. Bedingt durch die teilweise sehr großen Zeitfenster von mehreren Jahrhunderten und den damit verbundenen verschiedenen Überlieferungslinien ergeben sich nicht immer natürliche Veränderungen des *Text Re-use*. So lautet die englische Version des Spruches *Äpfel mit Birnen vergleichen* wie folgt: *Comparing apples to oranges*. Hierbei sind Birnen durch Orangen ausgetauscht worden, obwohl sie nicht einmal in einer synonymen semantischen Verwandtschaft stehen. Was wäre die Signatur einer solchen *Redewendung*? Ein weiteres Beispiel ist durch *Was kümmert/juckt es die*

(stolze/deutsche) Eiche, wenn sich ein/eine/der/die/das Borstenvieh/Eber/Sau/Wildsau dran/daran/an ihr wetzt/reibt?[1] gegeben. Hierbei stellen durch den Schrägstrich getrennte Wörter Alternativen an der jeweiligen Position dar. Bedingt durch die sprachliche Vielfalt, den *Re-use Variants* dieses Spruches, ist nach einer Stoppwortentfernung lediglich das Wort *Eiche* als sicher konstant anzunehmen. Alle anderen möglichen Wörter unterliegen der Varianz. Die Frage nach der *Signature* einer solchen *Text Re-use Unit* zeigt genau wie im vorangegangenen Beispiel die Komplexität und die Schwierigkeit dieses Prozesses mit historischen Daten auf.

Drittens sollte aus technischer Sicht so wenig wie möglich Vorwissen nötig sein. Vorwissen könnte bspw. bereits Worthäufigkeiten sein, um Stoppwörter gezielt zu entfernen, so dass sich die Geschwindigkeit der *Text Re-use Analysis* reduziert. Auf der anderen Seite muss es das Ziel sein, eine *Text Re-use Analysis* möglichst *streamingfähig* zu gestalten, so dass neben einer unabhängigen Parallelisierbarkeit, insbesondere beim Hinzufügen von neuen Dokumenten, der *Text Re-use* einer *Digital Library* nicht vollständig neu berechnet werden muss, sondern lediglich der *Text Re-use* zwischen den neuen Dokumenten mit den bereits indexierten Texten der *Digital Library*.

Viertens gibt es unterschiedliche Ansichten über die Qualität und Quantität einer *Text Re-use Analysis*. Was ist das zu erwartende Ergebnis? Während die Informatik die Quantität im Sinne des *Recalls* bedient und dabei bspw. eingangs genannte *Re-use Variants* nicht erkannt werden, sind geisteswissenschaftliche Arbeiten auf ein hohes Maß an *Precision* ausgerichtet. In diesem Kontext ist immer die Frage nach der *Re-use Completeness* zu stellen, die für die fachwissenschaftliche Arbeit eine grundlegende Annahme ist. Sowohl das qualitative und tiefenanalytische Sammeln von Belegstellen in geringen Mengen als auch die quantitative Breitenanalyse eines *Mining*-Verfahrens können die notwendige *Completeness* nicht per se bedienen. Wenn auch widersprüchlich, so muss es das Ziel sein, sowohl auf *Precision* als auch *Recall* zu optimieren (vgl. [Büchler 2012c]).

Fünftens muss ein *Feature* aus Infrastruktursicht, wie *CLARIN*, *DARIAH*, *Bamboo* oder *Europeana*, effizient sein. Insbesondere bei verteilten Anwendungen, wie dem *Distributed Text Re-use*, sind anfallende *Linking*-Kosten durch Latenzzeiten des Netzwerkes ausschlaggebend für die Gesamtgeschwindigkeit, so dass ein wenig oder gar kein *Text Re-use* aufdeckendes *Feature* als ineffizient angesehen werden muss. Ziel sollte es daher sein, *Features* so auszuwählen, dass eine hohe Wahrscheinlichkeit besteht, durch sie *Text Re-use* aufzudecken.

Sechstens ist eine *Text Re-use Analysis* im Rahmen der *eHumanities* nur dann sinnvoll, wenn sowohl die Verfahren als auch die Ergebnisse von den Fachwissenschaften akzeptiert werden. Das gestaltet sich oftmals bereits insofern als schwierig, als dass die Fachwissenschaften nicht wissen, was in der *Black Box* Text Mining im Detail geschieht (vgl. *ACID for the eHumanities* Paradigma aus Abschnitt 1.5).

All diese sechs Aspekte sind in das eingangs bereits erwähnte *TRACER*-Tool eingeflossen. Entgegen anderen monolithischen *Text Mining* Werkzeugen, wurde eine modulare, siebenschichtige Architektur zugrunde gelegt, wobei jede Schicht bzw. jedes Level eine der sieben Subaufgaben des *Text Re-use Detection* entspricht:

- Segmentierung (vgl. *Level 1: Segmentierung* in Abschnitt 3.2)
- Preprocessing (vgl. *Level 2: Preprocessing* in Abschnitt 3.3)
- Featuring (vgl. *Level 3: Featuring* in Abschnitt 3.4)
- Selection (vgl. *Level 4: Selection* in Abschnitt 3.5)

[1] Dieses Beispiel wurde freundlicherweise von Cerstin Mahlow im Rahmen ihres Vortrags *Exploring diachronic German corpora with respect to phraseologic information* im *Leipzig eHumanities Seminar* aus dem *OldPhrase*-Projekt (vgl. `http://oldphras.unibas.ch/`) bereitgestellt.

- Linking (vgl. *Level 5: Linking* in Abschnitt 3.6)
- Scoring (vgl. *Level 6: Scoring* in Abschnitt 3.7)
- Postprocessing (vgl. *Level 7: Postprocessing* in Abschnitt 3.8)

Durch diese Modularisierung kann nicht nur ein neues *Software Re-use Paradigma* sondern auch eine neue Stufe der *Acceptance* erreicht werden. Die Aufsplittung einer monolithischen Software in diese sieben Level ermöglicht das Verwenden einer *Level Chain* der ersten vier Level, um bspw. beschädigte Inschriften und Papyri semi-automatisch zu vervollständigen [Kruse 2009, Büchler 2012d] oder auch durch Austauschen von Daten bzw. Implementierungen auf den ersten beiden Level Wortähnlichkeiten sowohl auf *String*- als auch semantischer Ebene berechnen zu lassen. Die *Acceptance* der Fachwissenschaften kann insofern verbessert bzw. erreicht werden, wenn durch das Abspeichern aller Zwischenergebnisse der einzelnen Level ein *Text Re-use Debugger* nach dem Vorbild eines Debuggers einer Softwareentwicklungsumgebung wie *Netbeans*[2] geschaffen wird. Hierbei ist es möglich, sich sowohl den *Input* als auch den *Output* einer Implementierung auf jedem Level anzeigen zu lassen[3].

Ganz im Sinne dieser Infrastrukturparadigmen zeigt Tabelle 3.1 die *Input*- und *Output*-Parameter bzw. die entsprechenden Terminologien der einzelnen Level. Hierbei werden die Dokumente d einer *Digital Library* D_S sukzessive in einzelne *Re-use Units* zerlegt (Segmentation), aufbereitet, lemmatisiert oder normalisiert (*Preprocessing*), der digitale *Fingerprint* einer *Re-use Unit* bestimmt, aus ebendiesem *Fingerprint* die charakteristischsten Merkmale, die *Signature*, entnommen (*Selection*) sowie abschließend die Ähnlichkeiten zwischen zwei paarweise verglichenen *Re-use Units* berechnet (*Linking* und *Scoring*).

Level	**Formel-zeichen**	**Input**	**Output**
Segmentation	ξ_Θ	Digital Library D_S	Re-use Units V_{D_S}
Preprocessing	ψ_Θ	Re-use Units V_{D_S}	Cleaned Re-use Units
Featuring	μ_Θ	Cleaned Re-use Units	Digital Fingerprint
Selection	σ_Θ	Digital Fingerprint	Signature
Linking	λ_Θ	Signature	Candidate List
Scoring	θ_Θ	Candidate List	Result List $E^H_{D_S,\phi_\Theta}$
Postprocessing	π_Θ	Result List $E^H_{D_S,\phi_\Theta}$	Reduced Result List

Tabelle 3.1: *7-Level-Architektur* des *Historical Text Re-use*. Die Tabelle bildet für die sieben Level *Segmentation*, *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring*, *Postprocessing* das benutzte Formelzeichen sowie den jeweiligen *Input* und *Output* der einzelnen Level ab. Bereits eingeführte Formelzeichen sind an den jeweiligen Stellen genannt.

Der *Text Re-use* ϕ_Θ einer *Digital Library* D_S (vgl. Definition 2 auf Seite 65) kann somit als die in Formel 3.1 dargestellte Verkettung der in Tabelle 3.1 eingeführten Level verstanden werden.

$$\phi_\Theta = \pi_\Theta(\theta_\Theta(\lambda_\Theta(\sigma_\Theta(\mu_\Theta(\psi_\Theta(\xi_\Theta(D_S)))))))) \tag{3.1}$$

Dieses Kapitel wird nachfolgend auf die sieben Level mit jeweils einem Abschnitt im Detail eingehen. Dem folgt ein Abschnitt, welcher Zusammenhänge bzw. Wechselwirkungen zwischen den sieben Level beschreibt sowie einem Abschnitt, der die *Text Re-use Compression* einführt.

[2] vgl. `http://netbeans.org/`
[3] vgl. `http://roedel.e-humanities.net:8080/webdebugger/webdebugger/word_input_form`

3.2 Level 1: Segmentation

Die Segmentierung $\xi_\Theta(D_S)$ einer *Digital Library* D_S in einzelne *Re-use Units* ist ein komplexer Prozess, der vom Forschungsinteresse und noch mehr von den enthaltenen *Meme* bestimmt wird (vgl. Tabelle 2.1 bis 2.5 in Abschnitt 2.6). Die Segmentierung kann daher aus zweierlei Sicht betrachtet werden. Einerseits, ob eine *Digital Library* überlappend oder disjunkt segmentiert werden soll. Andererseits muss die Größe des *Re-use Unit* bestimmt werden.

Eine überlappende oder disjunkte Segmentierung hängt im Wesentlichen von der Größe der zu bestimmenden *Meme* ab. Bei kurzen *Meme*, wie größtenteils innerhalb der *Perseus Digital Library* beobachtbar, liefert eine überlappende Segmentierung in Form eines *Moving Windows* die besten Ergebnisse (vgl. [Büchler 2012c]). Ist der *Text Re-use* länger, dann wird in der *Scientific Community* oftmals disjunkt auf Satzebene segmentiert. Während bei einer überlappenden Segmentierung durch ein *Moving Window* der Länge w jedes Token in genau w *Re-use Units* enthalten ist[4], ist jedes Token bei der disjunkten Segmentierung genau einer *Re-use Unit* zugeordnet.

Definition 10 (Disjoint Re- use Unit)**.** *Sei eine Digital Library D_S einer Sprache S gegeben. Eine Disjoint Re-use Unit s_i ist die geordnete Zerlegung einer Digital Library $D_S = s_1 s_2 s_3 ... s_{n-1} s_n$, bei welcher sich paarweise benachbarte Re-use Units s_i s_{i+1} in den Tokens nicht überlappen.*

Definition 11 (Overlapping Re- use Unit)**.** *Sei eine Digital Library D_S einer Sprache S gegeben. Eine Overlapping Re-use Unit s_i ist die überlappende Zerlegung einer Digital Library $D_S = s_1 s_2 s_3 ... s_{n-1} s_n$, bei welcher sich paarweise benachbarte Re-use Units s_i s_{i+1} in $\min(|s_i|, |s_{i+1}|) - 1$ Tokens überlappen.*

Auf historischen und im Speziellen antiken Texten wird diese Entscheidung dadurch erschwert, dass in den Originaltexten keine Leer- und Satzzeichen enthalten waren. Antike Texte ähneln daher sehr einer Folge von Zeichen wie die einer DNA-Sequenz. Etwaige Wort- und Satzmarkierungen sind nachträglich durch den Editor eines Textes hinzugefügt worden und somit keine zwingend verlässliche Informationsquelle. So ist es nicht selten beobachtbar, dass zwei Editionen des gleichen Textes existieren, welche Satzzeichen enthalten, die in der jeweils anderen Edition nicht gesetzt worden sind (vgl. *Textkritik* in [Dover 1997]). Im Kontext der in Abschnitt 2.4 eingeführten Qualitätskriterien werden auf historischen Dokumenten gleich sechs der acht aufgestellten Kriterien bei einer satzweisen Segmentierung verletzt:

- *Acceptance*: Ergebnisse einer satzweisen Segmentierung werden nur schwer in den Fachwissenschaften auf Akzeptanz treffen (vgl. *ACID for the eHumanities* Paradigma aus Abschnitt 1.5).

- *Circumvention*: Im Kontext des in Abschnitt 1.7 bereits eingeführten *Counter Text Re-use* stellt diese einfache und nicht absichtliche Varianz der Markierung des Satzendes bereits eine *Circumvention* während einer *Text Re-use Analysis* dar.

- *Collectability*: Das Bestimmen von Satzenden ist an Satzendezeichen gebunden. Sind diese verschoben oder nicht existent, können Satzenden nicht bestimmt werden.

[4] Für Randwörter am Anfang und Ende eines Dokumentes gilt dies nicht, sondern ist je nach Position zur Dokumentgrenze zwischen 1 und $w - 1$ *Re-use Units* zugeordnet.

- *Performance*: Die Qualität des gefundenen *Text Re-use* kann durch eine Satzsegmentierung insofern negativ beeinflusst werden, als dass er sich auf zwei benachbarte *Re-use Units* aufteilt und in keiner von beiden hinreichend signifikant mit einem Original ist.

- *Permanence*: Die Satzzeichen wurden erst nachträglich in den jeweiligen Editionen hinzugefügt.

- *Universality*: Es kann nicht davon ausgegangen werden, dass sowohl in jeder Edition eines Werkes als auch jedem Werk, welches einen relevanten *Text Re-use* enthält, die gleichen Regeln für die Satzsegmentierung angewendet wurden.

Entgegen der Segmentierung zu *Overlapping Re-use Units* auf antiken Texten, wie der *Perseus Digital Library* (vgl. [Crane 1985, Büchler 2012c]), mit den genannten Eigenschaften ist der *Re-use* in der Bibel, wie zwischen zwei Büchern (vgl. [Lee 2007] oder zwei Editionen (vgl. [Büchler 2011c]), wesentlich größer und auf ganze Verse bezogen, wodurch eine versweise Segmentierung zu *Disjoint Re-use Units* nahe liegt.

Neben der Frage nach einer Segmentierung in *Disjoint* oder *Overlapping Re-use Units* ist die Größe bzw. Länge des *Text Re-use* ein entscheidendes Kriterium für die Fensterwahl. Insbesondere beim *Disjoint Text Re-use* kann die Fenstergröße zwischen Segmentierungen auf *Satz-* (vgl. [Hose 2004, Büchler 2010d]), *Absatz-* (vgl. [Seo 2008]) oder *Dokumentenebene* variieren.

Abb. 3.1 repräsentiert alle möglichen bzw. in der Forschung relevanten Segmentierungen in einer Entscheidungsbaumdarstellung. Hierbei wird zuerst eine Entscheidung über eine *Overlapping* oder *Disjoint Segmentation* getroffen. Im zweiten Schritt wird die Größe der *Re-use Unit* sowie eine statische (vgl. *Moving Window* in Abb. 3.1) oder dynamische Länge der *Re-use Units* (vgl. *Sentence* in Abb. 3.1) auf Basis der zu erwartenden oder zu messenden *Meme* und der damit entsprechenden Länge definiert.

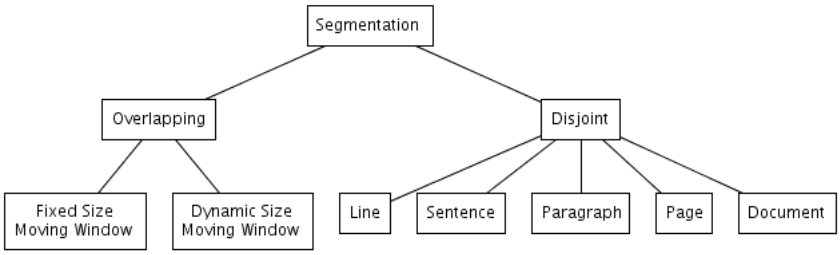


Abbildung 3.1: Taxonomie des Level *Segmentation* für den *Historical Text Re-use*. Im Sinne eines Entscheidungsbaumes werden die beiden dominanten Entscheidungen a) überlappende oder disjunkte *Re-use Units* und b) Größe (inkl. statischer oder dynamischer Fenstergröße) hierarchisch dargestellt.

Auch wenn die Entscheidung über eine geeignete Wahl der *Segmentation* besonders schwer erscheint, so können sogenannte *Document Citation* bzw. Dokumentstrukturen digital verfügbarer Texte (vgl. *Digital Humanities* in Abb. 2.1) genutzt werden. Als *Document Citation* kann die Einteilung eines digitalen Dokumentes in Werk, Buch, Seite, Absatz bzw. Sektion oder Zeile verstanden werden. Auch wenn, wie im *TLG*[5], insgesamt 147 verschie-

[5]TLG oder auch *Thesaurus Linguae Graecae* ist ein kommerzielles Korpus altgriechischer Texte.

dene *Document Citations* beobachtet werden können, so bietet der *Canonical Text Service* (vgl. [Smith 2012]) entsprechende hierarchische *Identifier*[6], wie

urn:cts:greekLit:tlg0012.tlg001:1.10

an, um beliebig große *Re-use Units* auf Basis der *Document Citation* zu adressieren. Hierbei steht *tlg0012* für den Autor, hier *Homer*, *tlg001* für das Werk, hier die *Iliad*, *:1* für das *Buch* und *.10* für die *Line*. Durch Weglassen von bspw. *.10* oder *:1.10* kann eine größere *Re-use Unit* auf *Buch-* bzw. *Werkebene* vom *Canonical Text Service Server* abgefragt und indexiert werden. Der Vorteil dieser Form der *Segmentation*, die mitunter nicht den Regeln der Informatik entspricht, ist, dass auf die eingangs bereits definierten Qualitätskriterien (vgl. Abschnitt 2.4) deutlich besser eingegangen wird. Insbesondere die *Acceptance* (vgl. auch das *ACID for the eHumanities* Paradigma aus Abschnitt 1.5) wird durch eine einschlägig in der fachwissenschaftlichen Community etablierten *Document Citation* deutlich verbessert.

3.3 Level 2: Preprocessing

Sprache wird divers bei der Sprachproduktion benutzt. Neben unterschiedlichen Gewohnheiten eines jeden Menschen spielen auch gesellschaftliche Konventionen, wie die Regel, dass ein Wort innerhalb eines Abschnittes oder Seite möglichst wenig wiederholt wird, eine Rolle. Verschiedene Techniken, wie das Benutzen von Synonymen oder Umschreiben von bereits im Text benutzten Worten, sind nur zwei Techniken, die die Vielfalt von im Text benutzten Konzepten repräsentieren. In den letzten Jahren ist die Bedeutung des *Preprocessings* für das *Text Mining* sukzessive gewachsen (vgl. [Buss 2008]). Auch wird seitens der *International Standard Organisation* versucht, die grundlegenden Terminologien und deren Bedeutung zu definieren (vgl. [Internation Standard Organisation ISO/TC 37 SC 4 2010]). Die Bedeutung des *Preprocessing* geht insbesondere mit der *Power Law* Verteilung von Wörtern oder Mengen von Wörtern, sogenannte *Frequent Itemsets* oder *Bibliograms*, einher.

Wortverteilungen folgen dem *Zipfschen Gesetz* $f \sim r^{-\alpha}$, wobei r der Rang eines Wortes und f die Frequenz mit einem $\alpha = 1 + \varepsilon$ nahe 1 ist (vgl. [Zipf 1949]). Aufgrund des Exponenten $\alpha \sim 1$ kann die relative Anzahl der Wörter $s(f)$ mit der gleichen Wortfrequenz f nach der harmonischen Reihe wie folgt bestimmt werden:

$$s(f) = \frac{1}{f * (f + 1)} \tag{3.2}$$

Bei einer Wortfrequenz von $f = 1$ folgt somit, dass etwa 50%[7] aller *Word Types* nur einmal in einer *Digital Library* beobachtet werden können. Diese Formel zeigt bereits auf, dass die meisten in einer *Digital Library* beobachteten *Word Types* sehr selten auftreten. Das Aufsummieren dieser Glieder wie in

$$S^n(f) = \sum_{f=1}^{n} \frac{1}{f * (f + 1)} \tag{3.3}$$

zeigt auf, dass 75% aller Wörter dreimal und seltener, 83% fünfmal und seltener sowie bereits über 90% aller *Word Types* zehnmal und seltener in einer *Digital Library* beobachtet werden können. Bei Bibliogrammen aus zwei Wörtern wie *Co-occurrences* oder *Bigrams*

[6] vgl. `http://folio.furman.edu/hmt-doc/cite/cts-urn-overview.html`

[7] Dieser Abschätzung wird sich mit steigender Größe einer *Digital Library* von oben genähert. Auf kleineren Textbeständen kann oftmals ein Wert von 60% und mehr beobachtet werden.

(vgl. Abschnitt 3.4) mit einem $\alpha = 2 + \varepsilon$ nahe 2 werden diese mit einer *Frequenz* von 1 in 75% aller Fälle beobachtet. 95% aller Zweiwortbibliogramme können fünfmal und seltener in einer *Digital Library* beobachtet werden.

Eines der Ziele des *Preprocessing* ist es demnach, die *Word Tokens* innerhalb einer *Digital Library* so zu verarbeiten, dass insbesondere die Sprachstatistik, die ein essenzieller Bestandteil des *Text Mining* und damit auch des *Text Re-use Mining* ist, zu verbessern. Auch wenn in Formeln beliebige Zahlen eingesetzt werden können, so benötigen probabilistische Sprachmodelle auch hinreichend oft beobachtete Ereignisse, so dass die statistische Aussagekraft gegeben ist. In [Church 1990] wird für Zweiwortbibliogramme (*Co-occurrences*) angegeben, dass die dort genannte statistische Methode für Frequenzen $f \leq 5$ instabil wird, was jedoch bereits 95% aller Daten ausmacht.

Das Verbessern der Sprachstatistik durch ein adäquates *Preprocessing* ψ_Θ umfasst Techniken des *Tokenisierens*, *Lemmatisierens* sowie der *String-* als auch *semantischen Normalisierung* (vgl. [Buss 2008]) mit dem Ziel, sprachliche Varianzen auf entsprechende Äquivalenzklassen zu reduzieren.

Auf historischen Texten können folgende Äquivalenzklassen beobachtet werden:

- *Encoding*: Die wichtigsten historischen Zeichen sind inzwischen vom *Unicode Consortium*[8] in das *Universal Character Set* aufgenommen. Insbesondere im Umgang mit diakritischen Zeichen ist das *Unicode Consortium* darum bemüht, die Speicherung und Darstellung entsprechender Buchstaben bestmöglich zu vereinheitlichen (vgl. [Davis 2012]). Insbesondere für historische Sprachen sind immer wieder *Unicode Form C* (Combining Diacritics Normalisation)[9] und *Unicode Form D* (Precomposed Normalisation) beobachtbar. Das technisch daraus resultierende Problem ist, dass sich gleiche altgriechische Texte nur durch die unterschiedlichen Unicode-konformen Normalisierungen nahezu komplett unähnlich bzw. ungleich sind[10].

- *Diacritics*: Diakritische Zeichen haben verschiedene Funktionen. Neben dem bereits genannten *Encoding*-Problem werden insbesondere im Altgriechischen Varianten eines Wortes beobachtet, die unterschiedliche diakritische Zeichen bzw. Kombinationen wie in Πλάτωνος und Πλάτωνός enthalten (Beispiel aus [Büchler 2008b] entnommen). Der *Greek Letter Shaver* (vgl. [Pansch 2011]) sowie die *ICU*-Implementierung[11] (vgl. [IBM 2012]) sind zwei Softwarepakete, die einen *String* S mit griechischen und anderen diakritischen Zeichen in einen *String* S' umwandeln können, der keine entsprechende Zeichen mehr enthält.

- *Capitalisation*: Viele antike Texte sind ursprünglich in Großbuchstaben geschrieben worden. Etwaige Anpassungen an die Groß- und Kleinschreibung stammen vom jeweiligen Editor. Prinzipiell werden in modernen Editionen alle Buchstaben kleingeschrieben. Jedoch bei Personennamen wie *Platon* finden sich immer wieder verschiedene und oftmals editorspezifische Schreibweisen wie Πλάτων, ΠΛΑΤΩΝ oder πλάτων (Beispiel aus [Büchler 2008b] entnommen).

[8] vgl. `http://unicode.org/`

[9] vgl. `http://www.unicode.org/charts/PDF/U0300.pdf`

[10] Das ist eine Erfahrung aus dem *eAQUA*-Projekt. Es wurden drei verschiedene Extraktoren (*tlgu*, *Lector* sowie der *Epidoc Transcoder* auf dem *Thesaurus Linguae Graecae*) getestet. Nach der Extraktion und Konvertierung nach UTF-8 wurde die jeweilige Wortliste erstellt. Obwohl immer das gleiche Korpus zugrunde lag, ergab eine Differenzanalyse, dass sich im Maximum 30% der Wörter in paarweise verglichenen Wortlisten überlappen, wobei die Schnittmenge meist aus Stopp- und kurzen Wörtern bestand, die oft keine diakritischen Zeichen enthielten.

[11] ICU=International Components for Unicode

- *Lemmatisation*: Verfahren zur Lemmatisierung von historischen Sprachen, wie dem Altgriechischen und dem Latein durch *Morpheus* (vgl. [Crane 1991]) sowie dem Arabischen durch den *Buckwalter Arabic Morphological Analyzer*[12] (vgl. [Buckwalter 2004a, Buckwalter 2004b]), decken oftmals nur bestimmte Epochen der jeweiligen Sprache ab. Das Ziel von *MorphAdorner* [Burns 2012] ist es, für englischsprachige Texte, wie in [Büchler 2011c] eingesetzt, diese Grenzen algorithmisch zu überwinden.

- *Root of a word*: Insbesondere in den semitischen Sprachen, wie dem Arabischen, ist während des *Preprocessing* oftmals der Umgang mit der Wurzel eines Wortes notwendig bzw. hilfreich. Die Wurzel eines Wortes entspricht einer Art "semantischer Morphologie". Für das Verb *schreiben* (in Umschrift *kataba*)[13] können zur selben Wurzel, hier *k-t-b*, auch die semantisch verwandten Wörter *kitāba* (das Schreiben), *kitāb* (Buch), *maktab* (Büro), *maktaba* (Bibliothek), *kitābī* (schriftlich) oder *kātib* (Sekretär, Schriftsteller) abgeleitet werden, die allesamt einen inhaltlichen Bezug zum Wort *schreiben* haben.

- *Historical Variants*: In historischen Texten haben sich nicht selten unterschiedliche Schreibweisen eines Wortes im Laufe der Zeit in verschiedenen geographischen Zonen etabliert. So können allein in diversen englischsprachigen Bibelversionen[14] zum Wort *antothite*[15] vier weitere Schreibweisen *anathothite* vs. *anethothite* vs. *anetothite* vs. *annethothite* beobachtet werden.

- *Spelling errors*: In [Kukich 1992] werden diverse Fehlerklassen für Schreibfehler definiert. In historischen und insbesondere antiken Texten kann davon ausgegangen werden, dass es keine Rechtschreibung im heutigen Sinne gab. Vielmehr gab es verschiedene Dialekte, wie im Altgriechischen, die im Jahre 403 v. Chr. adaptiert worden sind[16]. Dennoch können auch sichere Rechtschreibfehler, wie der *Real Word Error* in *Plutarch, Theseus 26.5*[17], beobachtet werden. In jenem Beispiel wird Ἕρμος (Hermus) versehentlich durch Ἑρμῆς (Hermes) ersetzt. Zu der Verwechslung der *Pythopolitan* ist es durch die Ähnlichkeit der beiden Genitive Ἕρμου und Ἑρμοῦ gekommen, die sich nur in den diakritischen Zeichen unterscheiden.

- *Dialects*: Wie im vorangegangenen Punkt bereits erwähnt, gibt es in historischen Texten zahlreiche *dialektische Varianten* wie συναγαγόντες und ξυναγαγόντες (Beispiel aus [Büchler 2010e] entnommen).

- *Language evolution*: Aufgrund der historischen Zeiträume sind sprachevolutionäre Varianten, wie γίγνηται und γίνηται, auch mit größeren Abweichungen, wie *couldeth* und *could*, beobachtbar (Beispiele sind aus [Büchler 2010e] (Griechisches Beispiel) und [Büchler 2011c] (Englisches Beispiel) entnommen).

[12] `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2004L02`

[13] Dieses Beispiel wurde dankenswerterweise von meiner Kollegin Dr. Ute Pietruschka aus Halle/S. bereitgestellt.

[14] Dieses Beispiel wurde in [Büchler 2011c] veröffentlicht. Untersucht wurden insgesamt sieben Bibelversionen: *King James Version* (KJV), *American Standard Version*, *Webster* (Webster), *World English Bible* (WEB), *Bible in Basic English* (Basic), *Darby* (Darby) sowie der *Young Literal Translation* (YLT). Ziel war es, im Sinne des *Noisy Channel Mining* verschiedene sprachliche Varianten zu extrahieren.

[15] *Antothite* ist die Adjektivierung von *Anathoth*, einer Stadt nordöstlich von Jerusalem nahe dem heutigen *Anata*.

[16] Als erste Rechtschreibreform kann das Adaptieren des ionischen Dialektes durch *Euclides* im Jahre 403 v. Chr. verstanden werden. Zuvor galt *Cadmos*, der etwa 2000 v. Chr. das griechische Alphabet eingeführt hat, lange Zeit als Autorität.

[17] Das folgende Beispiel wurde dankenswerterweise von meiner Kollegin Dr. Monica Berti zur Verfügung gestellt. Der *Real Word Error* in diesem Beispiel hat zur unmittelbaren Folge, dass aus einem Adligen und Edelmann ein Gott wird, wodurch der Kontext sich versehentlich stark verändert.

- *Paradigmatic relation*: In historischen Dokumenten ist oft beobachtbar, dass Wörter und Konzepte im Sinne von de Saussure's paradigmatischer Relation ausgetauscht wurden (vgl. [De Saussure 2001]). Wörter einer paradigmatischen Relation können sowohl Synonyme als auch Kohyponyme sein. Im Altgriechischen gibt es bspw. Evidenzen dafür, dass *Bier* und *Wein* in der Antike als Synonyme benutzt worden sind[18]. Heutzutage werden beide Wörter kohyponym benutzt. In den verschiedenen englischsprachigen Bibelversionen (vgl. [Büchler 2011c]) werden bspw. *sea-beast* und *sea-monster* paradigmatisch, hier Synonyme, verwendet. Andere paradigmatische Relationen, wie u. a. *sea-gull*, *sea-mew* und *sea-hawk* aber auch *apple-tree* und *citrus-tree*, werden kohyponym verwendet. Ferner sind auch komplexere Ersetzungen, hier temporale Konzepte, wie *not defer* (ASV, KJV, Webster), *not delay* (Darby, YLT) und *not wait* (WEB) auf der einen Seite sowie *without loss of time* (Basic) auf der anderen Seite, beobachtbar (Beispiel aus [Büchler 2011c] entnommen).

Die aufgelisteten Äquivalenzklassen und Beispiele zeigen die Komplexität des *Preprocessing*-Level auf, welches durchaus 50 - 70% der Gesamtzeit einnehmen kann. Jedoch zeigen die Beispiele auch, dass bei der dargestellten sprachlichen Vielfalt ein entsprechendes *Mining*-Ergebnis nachhaltig von der Qualität des *Preprocessing* abhängt. In [Heyer 2009] werden aus linguistischer Sicht die Äquivalenzklassen als *Concept* bezeichnet. [Stein 2007] ordnet diesen konzeptuellen Äquivalenzklassen verschiedene automatische Methoden des *Hash-based Text Retrievals*, wie *pLSA* (vgl. [Hofmann 1999]) oder das *Locality Sensitive Hashing* (vgl. [Gionis 1999, Paulevé 2010]), zu[19]. Aber auch *Topic Modelling* (vgl. [Blei 2006, Mimno 2012]) oder *Co-occurrence Similarity* (vgl. [Bordag 2007]) auf semantischer Ebene, der *FastSS*[20] (vgl. [Bocek 2007]) oder der *Levenshtein Distance* (vgl. [Levenshtein 1966]) bzgl. Ähnlichkeiten von *Strings* sowie morphologische Analysen, wie in *Morpheus* (vgl. [Crane 1991]) oder dem *Buckwalter Arabic Morphological Analyzer* (vgl. [Buckwalter 2004a, Buckwalter 2004b]), können als Techniken des *Hash-based Text Retrieval* verstanden werden.

Grundlage hierfür ist die Kombination aus dem *Locality Sensitive Hashing* (kurz *LSH*, vgl. [Charikar 2002]) und der *Min-wise Independent Permutation* (vgl. [Broder 1998]). *LSH* ist eine Hashfunktion, die sich von den in der Informatik bekannten Hashfunktionen, wie *md5* oder *crc32*, insofern unterscheidet, als dass *LSH* bei kleineren Änderungen nicht zum Ziel hat, möglichst genau 50% aller Bits in der Ausgabe zu verändern, um dadurch den größtmöglichen *Avalanche-Effekt* zu erzeugen (vgl. [Büchler 2008a]). Vielmehr ist es das Ziel der *LSH*-Technik, nicht nur für gleiche Eingaben den gleichen *Hash*-Wert bzw. eine Äquivalenzklasse zuzuordnen, sondern auch hinreichend ähnlichen bzw. nur minimal verschiedenen Eingaben einen ähnlichen oder gleichen Hashwert bzw. Äquivalenzklasse zuzuweisen, wodurch diese Klasse von Verfahren ihren Namen *Locality Sensitive Hashing* hat.

In [Charikar 2002] wird hierzu das Paradigma des *Locality Sensitive Hashing* vereinfacht, so dass sie durch die Formel 3.4 ausgedrückt werden kann. Hierbei ist $\mathcal{F}$ eine *Familie* von Hashfunktionen, die Objekte einer Kollektion C so verarbeiten kann, dass für $x, y \in C$ mit einer Ähnlichkeitsfunktion $sim(x, y) \in [0, 1]$ gilt

$$\mathbf{Pr}_{h \in \mathcal{F}}[h(x) = h(y)] = sim(x, y), \tag{3.4}$$

wobei die *Min-wise Independent Permutation* (vgl. [Broder 1998]) mit dem *Jaccard Coefficient* aus Formel 3.5 bestimmt wird.

[18] Dieses Beispiel wurde dankenswerterweise von meiner Kollegin Dr. Monica Berti zur Verfügung gestellt. *Bier* gilt hier als alkoholisches Getränk für Arme, während *Wein* als Getränk für Reiche angesehen wurde.

[19] Verschiedene weitere Verfahren und Techniken des *Hash-based Text Retrievals* sind in [Stein 2007] auf Seite 529 in Tabelle 1 zusammengetragen.

[20] FastSS ist die Abkürzung für *Fast String Similarity*.

$$sim(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.5}$$

Neben den bereits eingeführten *LSH*-Techniken, wie *pLSA* oder *Co-occurrences Similarity*, werden speziell in der Forschung zum *Text Re-use* auch das *T9-like Recoding* (vgl. [Basile 2009]) und das *Word Length Replacement* (vgl. [Barrón-Cedeño 2010b]) eingesetzt. Beim *T9-like Recoding* wird vorgeschlagen, ein Wort ähnlich der Eingabehilfe *T9* für mobilen Endgeräten zu repräsentieren. Beim *Word Length Replacement* wird die Hashfunktion $h(x) = len(x)$ dazu eingesetzt, das Wort durch die Wortlänge darzustellen. Beide Techniken haben den großen Vorteil, dass sie eine enorme Dimensionsreduktion von teilweise mehreren Millionen Wörtern auf wenige Äquivalenzklassen vornehmen. Gerade letztere Methode ermöglicht es, aufgrund ähnlicher Eigenschaften zur menschlichen DNA entsprechende Techniken des *Sequence Alignment* zu adaptieren (vgl. [Kumar 2004]). Im Kontext des *Historical Text Re-use* können beide *LSH*-Techniken eingesetzt werden, um *Meme* vom Typ *Edition* zu erkennen. Angesichts der starken Reduktion der möglichen *Features* wird dadurch die *Performance* (vgl. Abschnitt 2.4) aber auch die Sprachstatistik deutlich verbessert.

Abbildung 3.2 bildet die möglichen *Preprocessing*-Schritte der im Rahmen dieser Arbeit entstandenen *TRACER*-Implementierung ab. Einerseits kann auf Buchstabenebene mit der eingangs genannten *Groß-* und *Kleinschreibung* aber auch mit *diakritische Zeichen* umgegangen werden.

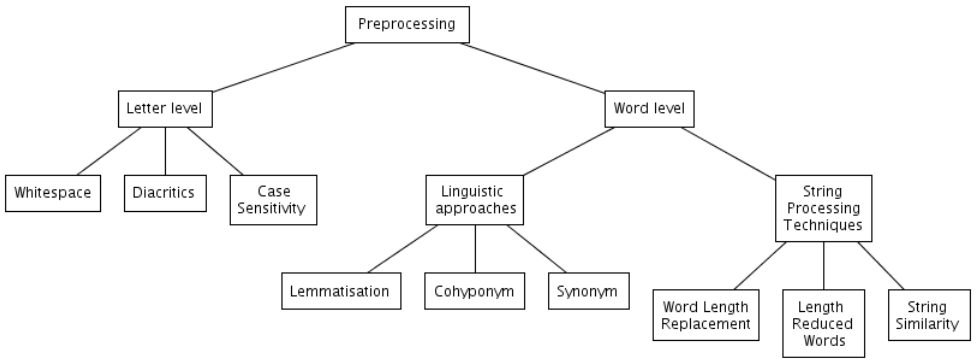


Abbildung 3.2: Taxonomie des Level *Preprocessing* für den *Historical Text Re-use*. Als Entscheidungsbaum werden verschiedene *Preprocessing*-Schritte dargestellt.

Auf der Wortebene kommen verschiedene *LSH*-Techniken zum Einsatz, wie die *Lemmatisation* durch *Morpheus* (vgl. [Crane 1991]) oder der *Buckwalter Arabic Morphological Analyzer* (vgl. [Buckwalter 2004a, Buckwalter 2004b]). *Cohyponyms* und *Synonyms* können sowohl aus automatischen Techniken, wie in [Bordag 2007] durch die *Co-occurrence Similarity* beschrieben[21], aber auch qualitativ gesammelten und lektorierten Daten, wie *WordNet* (vgl. [Miller 1995, Fellbaum 1998]), eingesetzt werden, um die semantische Vielfalt einer *Digital Library* D_S für das *Text Re-use Mining* zu reduzieren.

[21] Diese Methode wurde im *eAQUA*-Projekt eingesetzt. Auch wenn keine Ergebnisse publiziert worden sind, so sei zumindest auf die gefundene semantische Äquivalenz zwischen Τροία (Troja) und Ἰλιάς (Ilias) hingewiesen, welche als Synonym angenommen werden kann. So handelt die homerische *Iliad* von der Eroberung *Trojas* durch das *Trojanische Pferd*.

Die *String Processing Techniques* sind insbesondere bei Texten mit einer Vielzahl von *Historical Variants*, *Spelling Variants*, *Dialects* oder *Language Evolution* hilfreich. *LSH*-Techniken, wie das *Letter Bigram Shingling* und der daraus resultierenden Ähnlichkeit nach Formel 3.5, liefern bereits beachtliche Ergebnisse. In [Büchler 2011c] wurde aufgezeigt, dass beim *Text Re-use* auf sieben verschiedenen Bibelversionen das *Preprocessing* durch *Letter Bigram Shingling* eine nahezu identische *Precision* erreicht werden konnte, wie durch die Lemmatisierung der Texte[22].

Die durch *LSH*-Techniken gebildeten Äquivalenzklassen K werden im *Information Retrieval* eingesetzt, um eine Anfrage Q durch eine *Query Expansion* (vgl. [Efthimiadis 1996] und [Manning 2008] Seiten 173-175) zu erweitern. Bei einer *Text Re-use Analysis* würde eine *Query Expansion* während des *Linking* (vgl. Abschnitt 3.6) umgesetzt werden. Dies hätte jedoch zum Nachteil, dass die *Selection* (vgl. Abschnitt 3.5), welche vor dem *Linking* geschaltet ist, um die Menge der *Features* aus *Performance*-Gründen zu reduzieren, auf nicht oder nur tokenisierten *Features* angewendet werden könnte. *Selection*-Strategien auf Basis des $tf.idf$- (vgl. [Salton 1975]) oder des *Log-Likelihood*-Maßes (vgl. [Dunning 1993]) aber auch der informationstheoretischen *Redundancy* (vgl. [Shannon 1948]) würden dadurch nachhaltig negativ beeinflusst werden (Details zum *Selection* in Abschnitt 3.5).

Der Umgang mit den Elementen einer *LSH*-Klasse während des *Preprocessing* bringt jedoch das Problem hervor, dass ein *Repräsentant* dieser Klasse ausgewählt werden muss. In diesem Sinne können die Elemente einer *Äquivalenzklasse* K als *gerichteter Graph*, wie bei der *Lemmatisierung*, bzw. *ungerichtete Graphen*, wie bei *Synonyms*, *Cohyponyms* oder auch *String Similarity*, verstanden werden. Zu den *gerichteten Graphen* $G = (V, E)$ (vgl. [Aggarwal 2010a, Chakrabarti 2010]) auf Basis einer *LSH*-Funktion, wie in Abb. 3.3, zählen beispielsweise *Lemmatisation*-Graphen[23] oder auch *Disambiguierungsgraphen*.

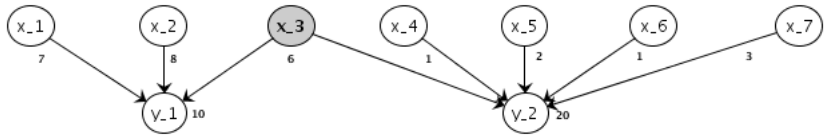


Abbildung 3.3: Gewichtung von *gerichteten Graphen* durch eine *PageRank*-ähnliche Technik (vgl. [Brin 1998]). Für x_3 kann sowohl y_1 als auch y_2 gelten. Auf Basis der Frequenzen $freq(y_1) = 10$ und $freq(y_2) = 20$ müsste sich für $y_{max} = y_2$ entschieden werden. Unter Berücksichtigung der zu y_i eingehenden Kanten werden die Frequenzen $freq(x_i)$ für alle eingehenden x_i zu y_j addiert. Unter dieser zusätzlichen Bedingung gilt nun $y_{max} = y_1$ als der vertrauensvollste Kandidat für x_3.

Für Elemente aus Abb. 3.3, wie x_1, x_2 oder x_4, liefern *gerichtete Graphen* bereits eine eindeutige Abbildungsfunktion mit $x_i \Rightarrow y_j$. Für Elemente, wie x_3, werden jedoch Mengen zurückgeliefert, für welche durch Formel 3.6 der beste Kandidat y_i für ein gegebenes x_i bestimmt werden kann.

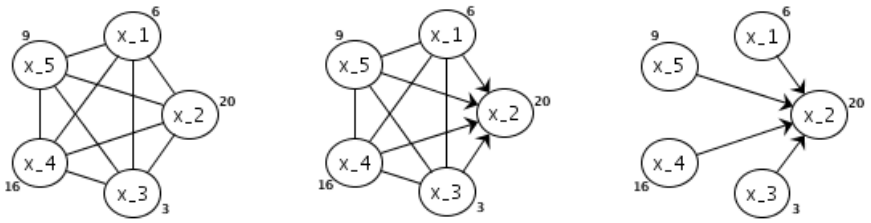
[22] Die guten Ergebnisse konnten in [Büchler 2012c] auf altgriechischen Texten jedoch nicht reproduziert werden. Auch wenn die Gründe dafür nicht tiefer analysiert worden sind, so scheinen zwei Gründe ausschlaggebend dafür gewesen zu sein. Einerseits hat das altgriechische eine deutlich stärkere Morphologie. Andererseits sind die in citeBCMB2012 gemachten Analysen auf die Werke von *Homer* und *Athenaeus* beschränkt, so dass bei der Wahl größerer Textkollektionen die Ergebnisse mitunter positiver ausfallen können.

[23] Es sei an dieser Stelle darauf hingewiesen, dass eine Wortform insbesondere im Altgriechischen mehr als eine Grundform haben kann. Dies kann u. a. durch Sprachevolution aber auch Dialekte begründet sein.

$$y_{max} = \arg\max_{y_i \in Y} \left(freq(y_j) + \sum_{(x_i, y_j) \in E} freq(x_i) \right) \tag{3.6}$$

Die grundlegende Idee hinter Formel 3.6 entspricht der *PageRanking*-Technik[24], welche Graphstrukturen, wie webbasierte Hypertexte (vgl. [Brin 1998]) aber auch einem *Text Re-use Graph* (vgl. [Büchler 2012b]), nutzt, um entsprechende Knoten zu gewichten.

Ungerichtete Graphen, wie die *Synonym-*, *Cohyponym-* oder *String Similarity*-Graphen (vgl. Abb. 3.2), werden im *TRACER*-Tool in drei Schritten zu *gerichteten Graphen* konvertiert (vgl. Abb. 3.4). Im ersten Schritt (vgl. Abb. 3.4(a)) wird auf Basis der Frequenz $freq(x_i)$, der Kandidat mit der höchsten Frequenz x_{max} bestimmt. Im zweiten Schritt (vgl. Abb. 3.4(b)) werden alle Kanten $(x_i, x_{max}) \in E$ zu x_{max} gerichtet. Im dritten Schritt (vgl. Abb. 3.4(c)) werden alle noch verbliebenen, ungerichteten Kanten aus dem Graphen $G = (V, E)$ entfernt.



(a) Ungerichteter Graph mit Frequenzen

(b) Gewichtung aller direkten Kanten zu x_{max}

(c) Löschen aller verbleibenden ungerichteten Kanten

Abbildung 3.4: Konvertierung eines *ungerichteten Graphen* zu einem *gerichteten Graphen*. Während der sprachlichen Normalisierung des *Preprocessing* können *ungerichtete Graphen*, wie *WordNet* (vgl. [Miller 1995, Fellbaum 1998]), dazu eingesetzt werden, um entsprechende sprachliche Vielfalt zu harmonisieren. Hierzu wird auf Basis von Frequenzen $freq(x_i)$ eines Knotens x_i eine Gewichtung vorgenommen.

Nachdem aus einem *ungerichteten Graphen G* ein reduzierter *gerichteter Graph* G' konvertiert worden ist, wird mit dem G' wie in Abb. 3.3 und Formel 3.6 verfahren. Die Notwendigkeit sei an einem einfachen Beispiel gezeigt: Es sei angenommen, dass in Abb. 3.4(a) der Knoten x_5 mit einem Knoten x_6 verbunden ist, welcher ebenfalls eine Frequenz $freq(x_2) = freq(x_6) = 20$ besitzt. In diesem Fall würde x_5 sowohl auf x_2 als auch x_6 gerichtet werden.

Auf Seite 29 im einleitenden Abschnitt 1.2 wurde im Kontext des *Re-use in der Natur* auf die Gleichheit und die Unterschiede von Bäumen verwiesen, die am Rand einer Landstraße stehen. Die im Rahmen des *Preprocessing* eingeführten *LSH*-Methoden fokussieren auf genau diese Wechselwirkung. Es werden gleiche Objekte, hier Bäume, der jeweiligen Baumart, in diesem Abschnitt die *Äquivalenzklassen* genannt, trotz individueller Varianten zugewiesen. Das *Preprocessing* von Texten einer *Digital Library* D_S bildet genau diesen Prozess mit der damit verbundenen Komplexität ab. Vielmehr sichert ein gutes *Preprocessing* bereits einige der *Qualitätskriterien* (vgl. Abschnitt 2.4) wie *Acceptability*, *Permanence* oder auch *Distinctiveness*.

[24] Bei der in Formel 3.6 genannten Vereinfachung des *PageRanking* wird im Wesentlichen auf die Iteration während des Berechnens der Gewichte aus *Performance*-Gründen verzichtet.

3.4 Level 3: Featuring

Nachdem ein Dokument d einer *Digital Library* D_S in einzelne *Re-use Units* gemäß des zu messenden *Text Re-use* (vgl. Abschnitt 3.2) zerlegt worden ist und ebendiese über verschiedene Harmonisierungsschritte (vgl. *Preprocessing* aus Abschnitt 3.3) normalisiert worden sind, müssen die *Re-use Units* partitioniert werden. *Re-use Units* können per se nicht miteinander verglichen werden, sondern sie müssen in einzelne *Atome* zerlegt werden, die die jeweilige *Re-use Unit* repräsentieren.

Definition 12 (Atom[25]). *Sei eine Re-use Unit s_i gegeben. Ein Atom ist die für eine Text Re-use Analysis gewählte Zerlegung einer Re-use Unit s_i in kleinere Einheiten*[26].

Ein *Atom* kann nach Definition 12 als ein *Abtastungstyp* einer *Re-use Unit* verstanden werden. Eine konkrete Instanz, also ein *Token* eines *Atoms*, wird nachfolgend als *Feature* bezeichnet.

Linguistisch relevante bzw. im *Information Retrieval* eingesetzte *Atome* (vgl. in Abschnitt 2.3, *Generationen-Modell*) können in zwei Klassen eingeteilt werden. Einerseits kann zwischen *eingliedrigen* und *mehrgliedrigen Atomen* unterschieden werden. Hierbei wird unterschieden, ob ein *Atom* aus einem Wort oder mehreren Wörtern gebildet wird. Da die automatische *Text Re-use Analysis* von einzelnen Wörtern, wie *Klopstock!*[27], derzeit gar nicht bzw. in manchen Fällen nur schwer möglich ist, haben *mehrgliedrige Atome*, wie *Bigrams* oder *Co-occurrences*, den Vorteil, dass die Frequenz eines *Features* gegenüber *eingliedrigen Atomen* oftmals um mehr als eine Dezimalstelle kleiner ist. So kommt ein *Re-use Atom* vom Typ *Word Bigram* oder *Co-occurrence* in 75% aller Fälle nur einmal, in 95% fünfmal und weniger sowie in etwa 97% seltener als zehnmal vor. Aufgrund der quadratischen Komplexität $O(n^2)$ im *Linking*-Schritt (vgl. Abschnitt 3.6) bedeutet der Einsatz von *mehrgliedrigen Atomen* eine *Performance*-Verbesserung auf in der Regel weniger als 1% der Zeit im Vergleich zu eingliedrigen *Atomen* bei gleichem Ergebnis.

Andererseits können *Atome* und die daraus resultierenden *Features* auch in einen *semantischen* oder *syntaktischen Strukturtyp*[28] eingeteilt werden. Dies sei an der deutschen Übersetzungen des lateinischen *geflügelten Wortes In vino veritas* erklärt[29]. Einerseits kann man sagen: *Die Wahrheit liegt im Wein.* Andererseits ist auch *Im Wein liegt die Wahrheit.* beobachtbar. Ein *syntaktischer Strukturtyp*, wie bspw. das *Bigram Shingling* oder das *Bigram Hashbreaking*, würden das *Bigram Wein liegt* oder *Wahrheit liegt* generieren. Der *semantische Strukturtyp*, wie die *Co-occurrence*, würde bspw. auch das Feature (*Wahrheit, Wein*) ermöglichen. Wissenschaftlich kann der *syntaktische Strukturtyp* in die Regeln der Sprachproduktion eingebunden werden (vgl. u. a. [Miller 1956]). Der *semantische Strukturtyp* ist im Kontext von *de Saussure's* Strukturalismus[30] (vgl. [De Saussure 2001]), genauer der *syntagmatischen Relation*, aber auch der *Hebb Theory* (vgl. [Hebb 1949]) zu betrachten.

Abbildung 3.5 systematisiert die für den *Historical Text Re-use* relevanten *Re-use Atome*. Hierbei werden die *Atome* in drei verschiedene Klassen eingeteilt. Die Klasse der *Non-*

[25] In der Fachliteratur kann für *Atom* auch die Terminologie *Sketch* beobachtet werden. Im Rahmen dieser Arbeit wird hierfür *Atom* festgelegt, da diese Terminologie dem Autor anschaulicher scheint.

[26] Mögliche *Atome* sind in Abb. 3.5 abgebildet.

[27] Dieses Beispiel wurde von meinem Kollegen Prof. Dr. Gerhard Lauer bereitgestellt.

[28] Syntaktisch wird hier abweichend von der Linguistik als eine Sequenz von Tokens definiert, der eine gemeinsamen *Syntax* zugrunde liegt.

[29] Dieses Beispiel wurde im Rahmen des *eTRACES*-Projektes von Annette Geßner zusammengetragen.

[30] Im Kontext der *Humanities* sei darauf verwiesen, dass sich bereits *Aristoteles* mit dieser Frage beschäftigt hat. In seinen drei Gesetzen der Assoziation legte er Grundlagen für die *Assoziationpsychologie*. Das für den *semantischen Strukturtyp* relevante *law of contiguity* ist in seiner Aussage der *Hebb Theory* sehr nah. Vielmehr kann das *law of contiguity* von Aristoteles als eine der ältesten Motivationen und Beschreibungen für die moderne Kookurrenzanalyse verstanden werden (vgl. [Büchler 2006b, Büchler 2008a]).

statistical Approaches zeichnet sich dadurch aus, dass Techniken dieser Klasse sehr gut *streamingfähig* sind. Einerseits zählen hierzu *Pattern based Approaches*, wie das Extrahieren von *Surface Features*, wie *Quotation Marks*, aber auch Techniken, um *Canonical References* (vgl. [Romanello 2009]), wie *Hom. Il. 1 10*[31], zu extrahieren. Sowohl die *Surface Features* als auch die *Canonical References* setzen voraus, dass der Editor eines Textes entsprechende *Marker* im Text gesetzt hat[32].

Marker, die ein antiker Autor bereits gesetzt hat, passen auf das Pattern *<ENTITY>* *<VERBUM DICENDI>* (vgl. Abb. 3.5), wobei *<ENTITY>* eine *Named Entity*, wie eine *Person* oder ein *Ethnikon*, sein kann. *<VERBUM DICENDI>* sind Verben, wie *schreiben* oder *sagen*, die im Rahmen dieses *Patterns* einen *Text Re-use* zwangsweise nach sich ziehen. Diese Form der Patterns wurden in [Moritz 2011] analysiert, um *Textfragemente* einer nicht mehr existierenden *Source* (vgl. Einbettung des *Historical Text Re-use* in Shannon's *Noisy Channel Theorem* in Abb. 1.6 aus Abschnitt 1.7) zu entdecken und schließlich zu extrahieren (vgl. [Berti 2009]). Für einen *Complete Text Re-use*, also wenn sowohl *Source* als auch *Target* eines *Text Re-use* in der *Digital Library* vorliegen, hingegen stellen die *<ENTITY>* *<VERBUM DICENDI>*-Pattern genau wie die *Surface Features* und die *Canonical References* eine gute Möglichkeit der Evaluierung dar.

Signal Processing Techniques verstehen Text als ein fortlaufendes Signal, wie bspw. auch ein *Audio-* oder *Videosignal*, welches in das Spektrum aus k Frequenzkomponenten zerlegt wird. In [Seo 2008] wird hierzu die *Discrete Cosine Transformation* (kurz *DCT*) als eine Vereinfachung der *Fast Fourier Transformation* eingeführt, welche in den dortigen Ergebnissen einem *Bigram Shingling* in der *Precision* bereits sehr nahe kommt[33].

Das *Syntactical Featuring* (vgl. Abb. 3.5) kann als die am stärksten strukturierte und differenzierte *Featuring*-Klasse verstanden werden. Techniken dieser *Featuring*-Klasse werden vornehmlich dazu benutzt, *Dubletten*, *Quasi-Dubletten* oder *Plagiarismus* aufzudecken. Es können Techniken des *Syntactical Featuring* nach zwei Kriterien unterschieden werden. Einerseits kann eine *Re-use Unit* sowohl *Overlapping* als auch *Disjoint* abgetastet werden[34]. Andererseits kann darin unterschieden werden, ob ein *Ngram* eine feste bzw. dynamische Länge hat. Überlappende Techniken werden auch *Shingling* und nicht überlappende Ansätze *Hashbreaking* genannt (vgl Abb. 3.5).

Hashbreaking-Techniken (vgl. [Seo 2008]) nutzen eine Heuristik, um eine Sequenz von Wortformen in nicht überlappende Teilsegmente aufzuteilen. Techniken des *Local Hashbreaking* nutzen bspw. die *Position* eines Wortes im Text, um durch ein $0 = \mod p$, wobei p die Länge des *Ngrams* ist, die *Re-use Unit* aufzusplitten. Bei einer *Re-use Unit* der Länge 10 würden somit genau 5 *Bigrams*, 4 *Trigrams* und 3 *Quattrograms* extrahiert werden. Beim *Global Hashbreaking* wird eine ähnliche Idee verfolgt, nur dass über eine *Digital Library* hinweg globales Wissen, wie bspw. der *Rang* eines *Features*, dazu genutzt wird, um an jeder Position mit $0 = \mod r$ (hier der Rang r eines *Features*) einen *Break* zu setzen. Während beim *Local Hashbreaking* p immer eine konstante *Ngram*-Größe zugrunde liegt, werden beim *Global Hashbreaking* unterschiedlich große *Ngrams* erzeugt.

[31] *Hom. Il. 1 10* ist ein Verweis des Editors auf die *Ilias* von *Homer* und dort genauer auf die zehnte Zeile im ersten Buch.

[32] Eine unmittelbare Anwendung einer automatischen *Text Re-use Analysis* wäre, genau diese Marker für einen gefundenen *Text Re-use* automatisch bzw. semi-automatisch zu setzen. Hierzu sei auf die in Abschnitt 3.2 erwähnten CTS-Identifier hingewiesen, die zu *Canonical References* transformiert werden können.

[33] In eigenen Experimenten konnten diese Ergebnisse jedoch nie reproduziert werden, so dass die *DCT* hier nur vollständigkeitshalber mit erwähnt und keine größere Rolle spielen wird. Jedoch ist die *Streamingfähigkeit* dieser Klasse von Methoden zukunftsweisend, auch wenn noch einiges an Forschung investiert werden muss.

[34] Das überlappende und disjunkte *Featuring* ist nicht mit den *Overlapping* und *Disjoint Re-use Units* zu verwechseln. Letzteres segmentiert eine *Digital Library* zu einzelnen *Re-use Units*. Ersteres hingegen erzeugt eine *Feature*-Menge, den *Digital Fingerprint*, zu jeder *Re-use Unit*.

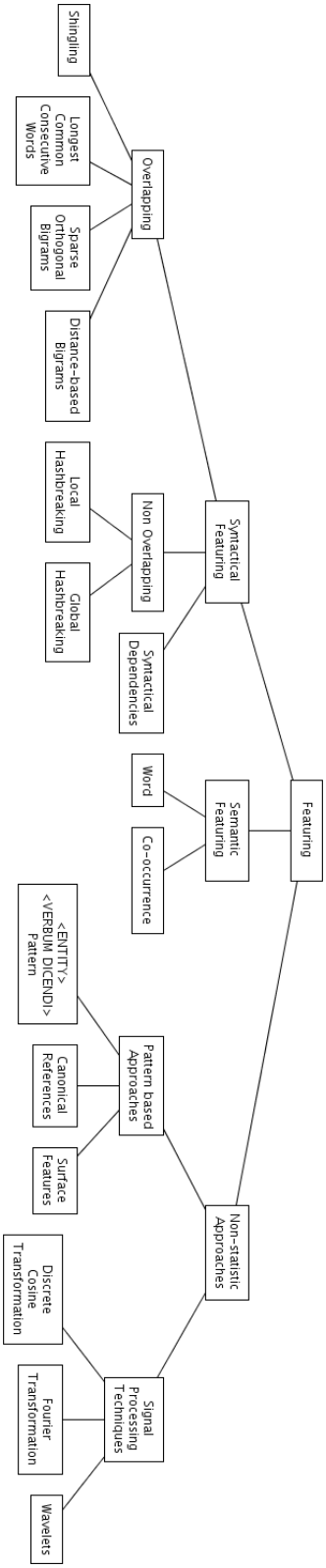


Abbildung 3.5: Taxonomie des Level *Featuring* für den *Historical Text Re-use*. Grundlegend kann zwischen den drei verschiedenen *Featuring*-Klassen *Syntactical Fingerprinting*, *Semantic Fingerprinting* sowie den *Non-statistic Approaches* unterschieden werden.

Zu den *Overlapping*-Techniken zählen das *Ngram Shingling* (vgl. [Seo 2008]), der *Longest Common Consecutive Words* (vgl. [Sediyono 2008]), die *Sparse Orthogonal Bigrams* (vgl. [Siefkes 2004] und die *Distance based Bigrams* (vgl. [Büchler 2008a]). Die beiden letzten Ansätze verfolgen das Ziel, das Paradigma eines klassischen *Bigrams* $w_i w_{i+1}$ dadurch aufzubrechen, dass *Bigrams* unter Hinzunahme der Distanz beider Wörter zu (w_i, w_{i+1}, d) gegenüber Einschüben robuster werden.

Dem *Shingling* liegt eine konstante *Ngram*-Größe n zugrunde. Bei dieser Technik wird vom ersten Wort beginnend das Abtastungsfenster der Länge n sukzessive um ein Wort nach rechts verschoben, so dass aus einer *Re-use Unit* der Länge l genau $l - n + 1$ *Bigrams* abgetastet werden[35] [36].

Der Ansatz des *Longest Common Consecutive Words* (vgl. [Sediyono 2008]) ist eine überlappende Technik mit dynamischer Fenstergröße der jeweiligen *Ngrams*. Hierbei werden alle *Ngram*-Sequenzen mit $freq(w_i w_{i+1} ... w_{i+j}) \geq 2$ als Features extrahiert. *LCCW*-Sequenzen werden auch als *Super Shingles* bezeichnet (vgl. [Broder 1997b]).

Der Vorteil des *Hashbreaking* gegenüber dem *Shingling* ist, dass die Menge der zu extrahierenden *Features* deutlich unter der des *Shinglings* liegt, woraus in den meisten Fällen bereits ein signifikanter *Performance*-Vorteil resultiert. *Hashbreaking*-Techniken sind daher dafür geeignet, offensichtliche *Dubletten* bzw. *Quasi-Dubletten*, wie bspw. mehrere *Editionen* des gleichen Werkes, zu erkennen.

Die Klasse der *Syntactical Dependencies* extrahiert relevante *Features* aus syntaktischen Strukturen wie den *Dependency Trees* (vgl. [Bamman 2008]). Ziel dieser Techniken ist es, Strukturen oberhalb der Wortebene zu identifizieren, um durch Sprachevolution bedingte Veränderungen die *Permanence* und die *Circumvention* zu verbessern (vgl. Abschnitt 2.4). Aufgrund der Schwierigkeiten, qualitativ benutzbare *Dependency Trees* automatisch zu generieren, bleibt diese Klasse zum Zeitpunkt der Erstellung dieser Arbeit noch unberücksichtigt, auch wenn zu erwarten ist, dass diese Technik zukünftig zunehmend stärker eingesetzt wird.

In diesem Abschnitt wurden verschiedene *Featuring*-Techniken vorgestellt. Auch wenn die Herangehensweisen teilweise diametral auseinander liegen, so haben sie alle in ihrer Funktion gemeinsam, dass sie aus einer monolithischen *Re-use Unit* s_i einen Vektor von *Features* $\overrightarrow{s_i}$ erzeugen, wobei alle *Features* vom gleichen *Atom* abgeleitet sind. Der *Feature*-Vektor $\overrightarrow{s_i}$ einer *Re-use Unit* wird in Anlehnung an die Biometrie nachfolgend auch *Digital Fingerprint* genannt.

Definition 13 (Digital Fingerprint)**.** *Sei eine Re-use Unit s_i gegeben. Ein Digital Fingerprint $\overrightarrow{s_i}$ entspricht der durch ein Featuring $\overrightarrow{s_i} = \mu_\Theta(s_i)$ in ihre Atome zerlegten Re-use Unit s_i.*

Das *Featuring* kann in der Matrixschreibweise als eine (n, m)-*Feature*-Matrix F mit n *Re-use Units* und m beobachtbaren *Features* verstanden werden, wobei eine 1 in F (vgl. Formel 3.7) ein Vorkommen eines *Features* innerhalb einer *Re-use Unit* s_i induziert.

[35] Das *Ngram Shingling* einer *Re-use Unit* ist außerhalb des *Text Re-use Detection* der dominanteste *Ngram*-Ansatz.

[36] Das im Rahmen dieser Arbeit entwickelte *TRACER-Tool* (vgl. [Büchler 2013a]) kann sowohl für das *Ngram Shingling* als auch das *Local Hashbreaking* mit einem n von $n \in [2, 10]$ umgehen.

$$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \tag{3.7}$$

Dies sei an einem einfachen Beispiel aus Formel 3.8, bestehend aus fünf *Re-use Units* $V_1 = \{s_1, s_2, s_3, s_4, s_5\}$ sowie 12 *Features* $V_2 = \{A, B, \ldots, J, K\}$, verdeutlicht. Jede der fünf *Re-use Units* sei mit einer Länge von fünf Wörtern angenommen. Weiterhin sei ein *Word*-basiertes *Featuring* μ_Θ (vgl. Abb. 3.5) benutzt, so dass die Menge der Wörter auch zeitgleich der Menge der *Features* entspricht.

$$\begin{array}{lccccc} s_1: & A & B & C & D & E \\ s_2: & A & C & E & F & G \\ s_3: & G & F & A & C & D \\ s_4: & C & F & A & G & E \\ s_5: & D & H & I & J & K \end{array} \tag{3.8}$$

Aus diesen fünf *Re-use Units* s_i wird durch das *Featuring* μ_Θ die $(5, 11)$-*Feature*-Matrix F in Formel 3.9 induziert.

$$\begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} A & C & D & F & G & E & B & H & I & J & K \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array} = F \tag{3.9}$$

Je nach *Preprocessing*- und *Featuring*-Technik kann F eine *Sparse Matrix* sein[37]. Insofern kann F auch als ein *Bipartiter Graph* $G = (V_1, V_2, E)$ (vgl. [Asratian 1998]) mit V_1 der Menge der *Re-use Units*, V_2 der Menge der *Features* sowie $(v_i, v_j) \in E$ der Menge der Kanten mit $v_i \in V_1$ und $v_j \in V_2$ verstanden und repräsentiert werden.

Alle hier vorgestellten *Atome* sind in der Praxis bereits erprobt und haben ihren jeweiligen Anwendungszweck. Kein *Atom* ist in jeder Situation besser geeignet als ein anderes. Die Wahl des *Atoms* ist in erster Linie von den in einer *Digital Library* enthaltenen *Meme* und deren Stabilität im *Re-use Style* abhängig (vgl. Systematisierung aus Abschnitt 2.6).

3.5 Level 4: Selection

Nachdem eine *Re-use Unit* in ihren *Digital Fingerprint*, also ihre Menge von beobachtbaren *Features*, zerlegt worden ist, stellt sich anschließend die Frage, welche dieser *Features* für die *Re-use Unit* relevant bzw. gut beschreibend sind. Dieser Abschnitt stellt einige *Selection*-Strategien $\sigma_\Theta(F)$ vor, die im Forschungsbereich des *Text Re-use* in der Vergangenheit

[37] F ist beispielsweise keine *Sparse Matrix*, wenn als *Preprocessing* das *Word Length Replacement* (vgl. [Barrón-Cedeño 2010b]) mit einem *Featuring* auf Wortebene eingesetzt wird.

berücksichtigt wurden. Die Vielzahl der nachfolgenden Strategien resultiert letztendlich daraus, dass es kein bestes und kein schlechtestes Verfahren gibt. Alle folgenden Strategien haben ihre Vor- und Nachteile. Vielmehr hängt die Wahl der *Selection*-Strategie stark von den in der *Digital Library* enthaltenen und oftmals verborgenen *Meme* ab (vgl. Abschnitt 2.6). Dies sei an vier einfachen Beispielen verdeutlicht:

- *Proverb*: Sein oder nicht sein, das ist hier die Frage. *Source: Shakespeare*
- *Winged Word*: Gleich und gleich gesellt sich gern. *Source: Homer zugeschrieben.*
- *Multi Word Unit*: König Alexander der Große
- *Frequente Sequenz*: im Namen unseres Herren Jesus Christus

Im *Information Retrieval* (vgl. [Manning 2008]) werden Maße und Techniken, wie das $tf.idf$ (vgl. [Salton 1975]) oder die *Differenzanalyse* (vgl. [Witschel 2004]), angewendet, um die relevanten Terme zu extrahieren bzw. wichtigen Termen ein entsprechend hohes Gewicht zu geben. Für den *Text Re-use* hingegen sind außerhalb der *Historiographie* diese Techniken oftmals eher wenig Vorteil bringend, wie die gezeigten Beispiele darstellen. Während bei den letzten beiden Beispielen einige Substantive und Entitäten, wie Vor- oder Personennamen, enthalten sind, bestehen das allgemein bekannte *Proverb* und das *Winged Word* nahezu komplett nur aus Stoppwörter, für die im Kontext der $tf.idf$-Metrik Termgewichte von 0 berechnet werden.

Andere Techniken, wie die der Klasse der Lesbarkeitstests, zu denen der *Dale-Chall Readability Score* (vgl. [RFP Evaluation Centers 2010b]), *Coleman-Liau Grade Level Readability Score* (vgl. [RFP Evaluation Centers 2010a]), *Automated Readability Index* (vgl. [Smith 1967]) gehören, liefern ein ähnliches Bild. So berechnet der *Automated Readability Index* für das *Shakespeare*-Zitat einen *Score*, der dem Text das Niveau einer 2. Klasse zuordnen würde.

Die Frage, die sich aus dieser Betrachtung unmittelbar ergibt, ist: *Warum verwenden wir bestimmte Textpassagen wieder und andere nicht* (vgl. auch [Finnegan 2011])? Während vorgenannte Techniken primär einzelne Worte oder Features zu bewerten versuchen, wird in [Büchler 2010f] sich dieser Frage durch die *Contrastive Semantics* genähert. Hierbei werden semantische Wortgraphen generiert, die letztlich wortpaarweise auf Unähnlichkeit untersucht werden. Ergebnis dieser Analyse sind unerwartete Wortassoziationen[38]. Analysen ergaben, dass in über 90% aller Fälle eine Textpassage, die ein Wortpaar mit einer *Contrastive Semantic* enthielt, am Ende auch mindestens ein zweites Mal wiederverwendet worden ist.

Diese einführenden Beispiele zeigen bereits die wahre Komplexität der *Selection* σ_Θ. Einerseits können *Features* eines *Digital Fingerprint* $\overrightarrow{s_i}$ einzeln betrachtet werden, wie es bspw. beim $tf.idf$-Maß geschieht. Andererseits können *Features* eines *Digital Fingerprint* $\overrightarrow{s_i}$ auf entsprechende Wechselwirkungen bzw. Abhängigkeiten untereinander, wie bspw. bei den *Contrastive Semantics*, analysiert werden. Bei Letzterem liegt immer die Annahme zugrunde, dass *Text Re-use* aus mehr als einem Wort besteht und somit zwangsläufig *Features* eines *Digital Fingerprint* nicht unabhängig voneinander auftreten. Vielmehr bedeutet *Text Re-use*, dass sich ein *Frequent Itemset* bzw. auch ein *Bibliogram* herausbildet.

Alle *Selection*-Strategien σ_Θ haben jedoch gemeinsam, dass sie den *Digital Fingerprint* $\overrightarrow{s_i}$ einer *Re-use Unit* s_i in seiner Länge verkürzen. Das Ergebnis einer *Selection* ist die *Digital Signature* einer *Re-use Unit* s_i.

[38] Eine der prominentesten Wortassoziationen hierzu ist der Zusammenhang zwischen *Bier* und *Windeln* aus eine Analyse von Kassenzetteln einer Supermarktkette (vgl. u. a. `http://www.spiegel.de/spiegel/print/d-83977252.html`).

Definition 14 (Digital Signature)**.** *Sei ein Digital Fingerprint* $\overrightarrow{s_i}$ *einer Re-use Unit* s_i *gegeben. Eine Digital Signature* $\overrightarrow{s_i'}$ *einer Text Re-use Unit* s_i *ist der durch* σ_Θ *verkürzte Digital Fingerprint* $\overrightarrow{s_i}$ *mit der Eigenschaft* $|\overrightarrow{s_i'}| \leq |\overrightarrow{s_i}|$.

Neben der bereits eingeführten Differenzierung nach unabhängigen bzw. bedingten *Features* können *Selection*-Strategien σ_Θ auch nach Art des *Selection Knowledge* sowie dem *Selection Usage* differenziert werden (vgl. Tabelle 3.2). Das *Selection Knowledge* kann sowohl *local* als auch *global* bestimmt werden. Die Klasse der *Local Selection Knowledge* nutzt lediglich die Informationen, die beim Betrachten einer *Re-use Unit* s_i zur Verfügung stehen, wie *Part-Of-Speech*-Tags (kurz *PoS*-Tags) oder auch Wortlängen. Dagegen nutzen *Selection*-Strategien der *Global Knowledge Selection* bspw. Wissen über Verteilungen, Abhängigkeiten zwischen *Features* aber auch nur *Feature*-Häufigkeiten. Während der Vorteil des *Local Selection Knowledge* in seiner perfekten *Streamingfähigkeit* und damit auch einer einfachen *Parallelisierung* im Sinne des *Distributed Text Re-use* aufgrund fehlender Abhängigkeiten liegt, können Methoden des *Global Selection Knowledge*, wie die bereits eingeführte *Contrastive Semantics*, die *Principle Component Analysis* (vgl. [Jolliffe 2002]) oder auch einem *min pruning* oder *max pruning*, durch ihre bzgl. einer *Digital Library* 'globaleren Entscheidungen" motiviert werden.

Genau wie beim *Selection Knowledge* kann auch beim zweiten Kriterium dem *Selection Usage* zwischen *local* und *global* unterschieden werden. Während sich bei einem *Local Selection Usage* die *Selection* auf eine *Re-use Unit* beschränkt, wird beim *Global Selection Usage* auf der Menge alle *Features* selektiert. Der Unterschied zwischen beiden Klassen kann einfach an der *Feature*-Matrix F aus Formel 3.7 verdeutlicht werden. Ausgehend von F kann eine *Signature*-Matrix S durch *Local Selection Usage* erzeugt werden, indem durch σ_Θ als unwichtig angesehene $F_{i,j}$ mit $F_{i,j} = 1$ in S mit $s_{i,j} = 0$ gesetzt werden. Somit wird durch σ_Θ aus einer (n, m)-*Feature*-Matrix $F_{i,j}$ eine (n, m)-*Signature*-Matrix S erzeugt. *Global Selection Usage* reduziert hingegen die Anzahl der Spalten einer (n, m')-*Signature*-Matrix S gegenüber der (n, m)-*Feature*-Matrix F mit $m' < m$. Der daraus resultierende offensichtliche Vorteil des *Global Selection Usage* ist, dass die (n, m')-*Signature*-Matrix S deutlich kleiner (horizontal) und die Datenmenge dadurch reduziert wird. Dies sei am Beispiel der beiden einfachsten *Selection*-Strategien, dem *min pruning* und *max pruning*, verdeutlicht. Durch ein *min pruning* kann die einem meist *Power-Law* zugrunde liegende *Feature*-Menge bereits durch Entfernen aller *Features* mit einer Häufigkeit von 1 um mindestens die Hälfte reduziert werden, welche aufgrund der Häufigkeit von 1 für *Text Re-use* innerhalb einer *Digital Library* nicht relevant sind. Auf der anderen Seite hilft ein *max pruning*, wenige, dafür aber umso häufigere, *Features* aus der *Signature*-Matrix S zu entfernen, um in erster Linie die *Geschwindigkeit* einer *Text Re-use Analysis* sicherzustellen.

Auch wenn aus Datenreduktionssicht bzw. unter Geschwindigkeitsaspekten ein *Global Selection Usage* sinnvoll erscheint, so stellen Verfahren dieser Klasse nicht sicher, dass jede *Re-use Unit* nach der *Selection* σ_Θ auch noch mindestens ein *Feature* in der *Digital Signature* enthält. Vielmehr steigt die Wahrscheinlichkeit mit kleiner werdender *Feature Density* rapide, so dass schnell mehrere *Digital Signature* $\overrightarrow{s_i}$ mit $|\overrightarrow{s_i}| = 0$ erzeugt werden.

Definition 15 (Feature Density)**.** *Sei eine Feature-Matrix* F *und eine Signature-Matrix* S *gegeben. Die Feature Density* $\mathcal{F}$ *ist definiert als das Verhältnis* $\mathcal{F} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m'} s_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}}$ *mit* $\mathcal{F} \in [0, 1]$.

Bezogen auf das in Formel 3.8 eingeführte Beispiel wurde die in Formel 3.9 dargestellte $(5, 11)$-*Feature*-Matrix generiert. Hierbei werden die *Features* $\{B, H, I, J, K\} \subset V_2$ nur einmal beobachtet. Insgesamt gibt es $5 \cdot 5 = 25$ *Feature Tokens*. Durch ein *min pruning* mit

einer *Feature Density* $\mathcal{F} = \frac{20}{25} = 0.8$ wird F aus Formel 3.9 auf eine $(5, 6)$-*Signature*-Matrix S, wie in Formel 3.10, reduziert, welche nun wiederum keine *Features* mehr enthält, die nur einmal aufgetreten sind.

$$\begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{array}{c} \begin{array}{cccccc} A & C & D & F & G & E \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} = S \tag{3.10}$$

Während sich beim *Global Selection Usage* die *Feature Density* auf die Verteilung der *Features* bezieht und dadurch *Digital Signature* $\overrightarrow{s_i}$ mit $|\overrightarrow{s_i}| = 0$ erzeugt werden können, bezieht sich die *Feature Density* beim *Local Selection Usage* auf die *Re-use Unit* selbst, wodurch mit Ausnahme einer *Feature Density* von 0 ausgeschlossen werden kann, dass eine *Digital Signature* $\overrightarrow{s_i}$ mit $|\overrightarrow{s_i}| = 0$ generiert wird.

Tabelle 3.2 spannt eine Matrix zwischen dem eingeführten *Selection Knowledge* und der *Selection Usage* mit den jeweiligen Vor- und Nachteilen auf.

		Selection Knowledge	
		local	global
Selection Usage	local	**Pro**: Streamingfähigkeit, einfache Parallelisierung **Kontra**: ein *Meme* in zwei s_i und s_j muss nicht die gleichen Features haben **Beispiel**: Wortlänge	**Pro**: Berücksichtigung von Abhängigkeiten zwischen *Features* bei *Digital Signature* $\overrightarrow{s_i'}$ mit $\lvert\overrightarrow{s_i'}\rvert \neq 0$ **Beispiel**: *Contrastive Semantics*
	global	**Pro**: Linguistische Features können berücksichtigt werden **Kontra**: Verteilung des *Local Knowledge* muss auf einen Score reduziert werden **Beispiel**: *PoS*-Tags	**Pro**: Berücksichtigung von Abhängigkeiten zwischen *Features* **Kontra**: *Digital Signature* $\overrightarrow{s_i'}$ mit $\lvert\overrightarrow{s_i'}\rvert = 0$ können erzeugt werden **Beispiel**: *max pruning*

Tabelle 3.2: *Selection Knowledge* vs. *Selection Usage*. Die Matrix vergleicht *Pros* und *Kontras* zwischen den jeweiligen Kategorien von *Selection*-Verfahren. Verfahren der vier Kategorien sind in den Abb. 3.6 und 3.7 abgebildet. *Global Selection Knowledge* bei *Local Selection Usage* bietet den besten Kompromiss aus genannten Vor- und Nachteilen.

Der wissenschaftliche Vergleich der verschiedenen *Selection*-Strategien gestaltet sich oftmals als sehr schwierig, da unterschiedliche Techniken andere Parameter benötigen. Ein *min pruning* oder *max pruning* benötigt einen Schwellwert, um bei einer Mindest- oder Höchstfrequenz eines *Features* entsprechend des *Schwellwertes* abzuschneiden. Eine *Minimum Word Length*-Technik (vgl. Abb. 3.6(a)) benötigt andererseits einen Schwellwert, der die Mindestlänge bestimmt. Abhängig von der *Selection*-Strategie sowie den entsprechenden Parametern ist eine Aussage nicht möglich, welche *Selection*-Technik sich für welche Aufgabe besser eignet (vgl. *Data Diversity* in Abschnitt 2.6). Um die Vergleichbarkeit herzustellen, wurde in der *TRACER*-Implementierung (vgl. [Büchler 2013a]) auf diverse Schwellwerte verzichtet und anstelle dessen, die in Definition 15 vorgestellte *Feature Density* $\mathcal{F}$ eingesetzt.

Jede *Selection*-Strategie kann als ein *Scoring* bzw. *Ranking* von *Features* verstanden werden. In der *TRACER*-Implementierung (vgl. [Büchler 2013a]) geschieht das in zwei einfachen Schritten:

- *Feature Scoring*: Beim *Global Selection Usage* wird die für eine *Digital Library* relevante Menge von teilweise mehreren Millionen *Features* gewichtet, während beim *Local Selection Usage* nur die Menge der *Features* einer *Re-use Unit* gewichtet wird.
- *Token Counting*: Für die gewichtete Liste aus dem ersten Schritt werden abschließend so viele *Tokens* ausgewählt, dass die aus Definition 15 gegebene *Feature Density* $\mathcal{F}$ erfüllt wird. Beim *Global Selection Usage* geschieht dies durch Aufsummieren der höchstgewichteten *Features* bis deren Anzahl an Tokens die vorher definierte *Feature Density* $\mathcal{F}$ erreicht. Beim *Local Selection Usage* werden ebenfalls die durch eine *Selection*-Strategie gewichteten *Features*, hier Tokens, aufsummiert bis die *Feature Density* $\mathcal{F}$ erreicht wird[39].

Für das *Global Selection Usage* als auch das *Local Selection Usage* kann auf diese Weise eine reale *Feature Density* $\mathcal{F}'$ erreicht werden, die der vorher definierten *Feature Density*, wie in Formel 3.11 dargestellt, um eine kleine Abweichung von $\varepsilon \leq 10^{-3}$ nicht überschreitet[40].

$$0 \leq \mathcal{F}' - \mathcal{F} \leq \varepsilon \tag{3.11}$$

Der in Tabelle 3.2 aufgezeigten Systematik können mehrere hundert *Selection*-Strategien zugewiesen werden. Im Nachfolgenden liegt der Fokus auf nur einigen Strategien, mit denen in der Vergangenheit bereits gute Ergebnisse erzielt werden konnten.

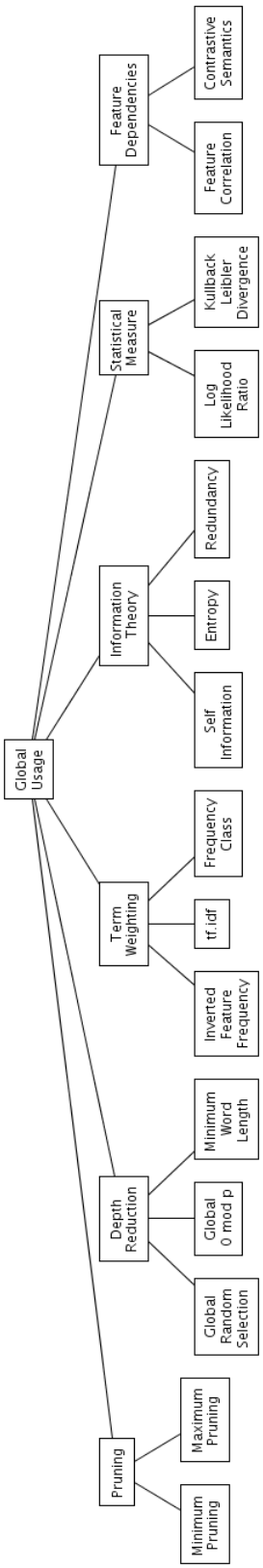
Abb. 3.6 repräsentiert die *Global Selection Knowledge*-Strategien. Sowohl im *Global Selection Usage* (vgl. Abb. 3.6(a)) als auch im *Local Selection Usage* (vgl. Abb. 3.6(b)) können die *Selection*-Strategien σ_Θ in die Klassen *Pruning*, *Depth Reduction*, *Term* bzw. *Feature Weighting*, Ansätze aus der *Information Theory*, *Statistical Measures* bzw. die bereits eingangs genannten *Feature Dependencies* eingeteilt werden.

Die Klasse der *Pruning*-Techniken enthält zwei bereits genannte *Selection*-Strategien des *min pruning* und *max pruning* (vgl. [Guyon 2003]). Der Vorteil dieser Techniken liegt in der Einfachheit und Schnelligkeit. Nachteilig ist jedoch, dass keine linguistische Motivation hinter diesen Techniken steht und deshalb oft relevante Teile des *Digital Fingerprints* nicht in die *Digital Signature* übertragen werden.

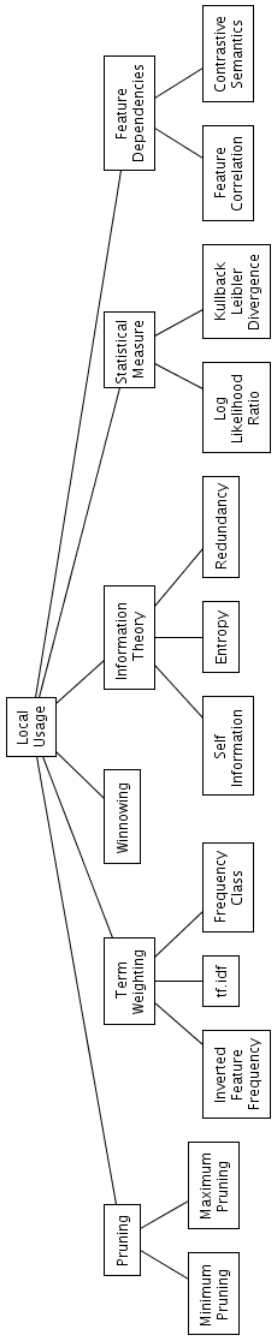
Zu den *Term Weighting*-Strategien zählen z. B. die $tf.idf$-Metrik (vgl. [Salton 1975]), die *Frequency Class* (vgl. [Witschel 2004]) aber auch die *Inverted Feature Frequency*. Der Vorteil der $tf.idf$-Metrik im Allgemeinen ist, dass sowohl häufige als auch seltene *Features* auf 0 gewichtet werden bzw. einen niedrigen *Score* erhalten. Auf historischen Dokumenten, wie dem *Thesaurus Linguae Graecae*, hat sich dieses Maß jedoch als nachteilig herausgestellt, da die $idf = \log \frac{N}{n_i}$ mit N Dokumenten in der *Digital Library* sowie n_i der Anzahl der Dokumente, die ein Feature enthalten, bei kleineren Dokumenten sukzessive vor Probleme gestellt wird. Die Grundannahme ist, dass das *idf* für die häufigen Stoppwörter 0 ist. Sind jedoch kleinere Dokumente Teil der *Digital Library* wird $idf \neq 0$, wodurch das $tf.idf$-Maß für die Stoppwörter die größten *Scores* berechnet (vgl. [Puchalla 2009]). Sowohl die *Frequency Class* als auch die *Inverted Feature Frequency* bevorzugen seltenere *Features* in der *Digital Signature*.

[39] An dieser Stelle sei noch einmal darauf verwiesen, dass bis auf den Fall $\mathcal{F} = 0$ ein *Local Selection Usage* immer mindestens ein *Feature* enthält, wodurch leere *Digital Signatures* $|\overrightarrow{s_i}| = 0$ vermieden werden können.

[40] Das ist durch die *Power Law*-Verteilung der *Features* gegeben, welche sehr viele seltene *Features* induziert. Mit Ausnahme von einem *min pruning* bei einer geringen *Feature Density* von $\mathcal{F} \leq 0.3$, wodurch durch häufiger Wörter größere Sprünge zu beobachten sind, überschreitet $\mathcal{F}'$ nur so gering $\mathcal{F}$, dass die Abweichungen für Vergleiche vernachlässigt werden können.



(a) Taxonomie des *Global Selection Usage*



(b) Taxonomy des *Local Selection Usage*

Abbildung 3.6: Taxonomie des Level *Selection* für den *Historical Text Re-use*. Die Abbildung zeigt die Klasse des *Global Selection Knowledge* sowohl im *Global* als auch *Local Selection Usage*. Grundlegende Techniken des *Global Selection Knowledge* können in den meisten Fällen in beiden *Selection Usage* eingesetzt werden. Alle hier vorgestellten Implementierungen können in der *TRACER*-Implementierung (vgl. [Büchler 2013a]) beliebig miteinander kombiniert werden.

Die Technik der *Redundancy* (vgl. [Shannon 1948]) ist ein Ansatz der *Information Theory* (vgl. Abb. 3.6). Die *Redundancy* kann durch den *Text Re-use* selbst motiviert werden. Durch jeden in einer *Digital Library* eingebrachten *Text Re-use* eines *Meme* erhöht sich dadurch auch die *Redundancy* innerhalb der *Digital Library* und somit auch der einzelnen *Features* des *Meme* (vgl. Abschnitt 2.6)[41]. Sowohl die *Entropy* als auch die *Self Information* sind Bestandteil der *Redundancy*, können aber auch separat von ihr betrachtet werden.

Die *Statistical Measures*, wie das *Log-Likelihood Ratio* (kurz *LLR*) oder die *Kullback-Leibler Divergence* (kurz *KLD*), zielen darauf ab, die statistische Zufälligkeit eines *Features* zu bestimmen. Hierbei werden diejenigen *Features* sowohl global als auch lokal hoch gewichtet, die eine starke innere Plausibilität haben. Im Rahmen der *Selection* können beide Techniken so verstanden werden, dass die Folge der Buchstaben eines *Features* möglichst stark den morphologischen und linguistischen Regeln einer Sprache unterliegen müssen. Etwaige durch ein *Preprocessing* (vgl. Abschnitt 3.3) nicht normalisierte, seltene sprachliche Varianten eines Wortes würden somit durch die *Statistical Measures* ignoriert werden.

Entgegen den bisherigen Klassen des *Global Knowledge Selection*, die die *Features* einzeln analysieren, sind Techniken der Klasse der *Feature Dependencies* (vgl. [Talavera 2000]) darauf ausgerichtet, das Verhältnis zwischen den *Features* für das *Selection* zu berücksichtigen. Im eingangs bereits erwähnten *Shakespeare*-Zitat *Sein oder nicht sein, das ist hier die Frage.* können somit durch ein *Feature Correlation* die Abhängigkeiten der *Bigram Shingles* bestimmt werden, so dass auch die vermeintlich häufigen Wörter dieses Zitates als stabile *Features* extrahiert werden.

Dem *Global Selection Knowledge* kann beim *Global Selection Usage* weiterhin die Klasse des *Depth Reduction* zugeordnet werden. Diese Methoden selektieren rein auf der *Feature*-Menge nach meist einfachen Heuristiken, wie dem *Global Random Selection*, dem *Global 0 mod p* aber auch einer einfachen *Minimum Word* bzw. *Feature Length* (vgl. u. a. [Seo 2008]). Das *Global Random Selection* entfernt zufällig solange *Features* aus der Menge aller *Features* einer *Digital Library*, bis die reduzierte Menge an *Features* der *Feature Density* $\mathcal{F}$ entspricht. Das *Global 0 mod p* nutzt einen *Hashwert* der *Features*, wozu unter anderem auch der Rang eines *Features* in einer *Power Law*-Verteilung zählt, welcher dann durch die *Modulo*-Operation gleichmäßig über der *Re-use Unit* selektiert.

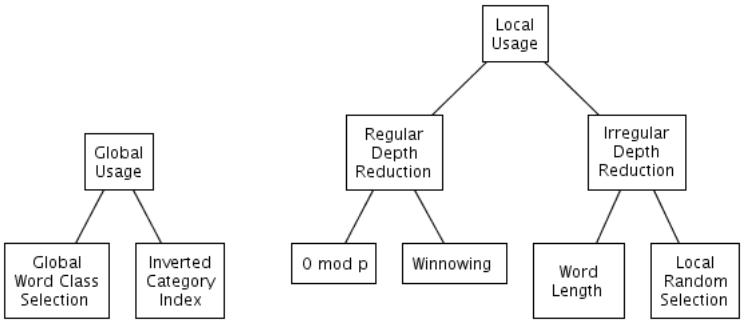
Im Vergleich zum *min* und *max pruning* bleiben sowohl häufige als auch seltenere *Features* in der *Feature*-Menge erhalten. Die Klasse der *Depth Reduction*-Strategien eignet sich sehr gut, um schnell und effektiv *Features* für das *Edition*-Detection auszuwählen.

Eine weiterführende Strategie des *Local Selection Usage* ist das *Winnowing* von *Features* (vgl. [Schleimer 2003]). Bei den *Winnowing*-Strategien gleitet ein *Moving Window* über die gemäß der *Re-use Unit* sortierten Menge der *Features*, welches kleiner als die *Re-use Unit* selbst ist. Jedes *Feature* besitzt einen *Hashwert*. Aus dem *Moving Window* werden durch Verschieben immer nur die *Features* selektiert, die innerhalb des Fensters den größten Hashwert besitzen. Der Vorteil dieser *Selection*-Strategie ist, dass Cluster von *Features* in der *Digital Signature* vermieden werden können. Vielmehr verteilen sich die *Features* der *Digital Signature* über die gesamte *Re-use Unit*.

Abbildung 3.7 fasst einige Methoden des *Local Selection Knowledge* zusammen, welche sich dadurch auszeichnen, dass die Entscheidung über die *Selection* mit dem beschränkten Wissen über die *Re-use Unit* selbst getroffen werden muss.

Techniken des *Local Selection Usage* können in die beiden Klassen der *Regular Depth Reduction* sowie *Irregular Depth Reduction* unterteilt werden (vgl. Abb. 3.7(b)). *Regular* und *irregular* bezieht sich hierbei auf die Verteilung der *Feature*-Selektion innerhalb der *Re-use*

[41] Die *Redundancy* ist somit nicht nur eine *Selection*-Strategie, sondern motiviert auch die *Text Re-use Compression* in Kapitel 3.10, welches als ein Maß für den *Text Re-use* innerhalb einer *Digital Library* verstanden werden kann.



(a) Taxonomie des *Global Selection Usage*

(b) Taxonomie des *Local Selection Usage*

Abbildung 3.7: Taxonomie des Level *Selection* für den *Historical Text Re-use*. Die Abb. zeigt die Klasse des *Local Selection Knowledge* sowohl im *Global* als auch *Local Selection Usage*.

Unit. Beim *Regular Depth Reduction* verteilen sich die ausgewählten *Features* über die gesamte *Re-use Unit*, während sich bei den *Irregular Depth Reduction*-Techniken auch *Feature Clusters* bzw. Textstellen einer *Re-use Unit* bilden können, die nicht durch entsprechende *Features* repräsentiert werden.

Techniken der lokalen *Regular Depth Reduction* sind bspw. die *0 mod p*-Technik und das *Local Winnowing*. Beim lokalen *0 mod p* (vgl. u. a. [Seo 2008]) wird die Position eines *Features* innerhalb der *Re-use Unit* benutzt, um diese numerische Position p durch eine *Modulo*-Operation zu selektieren. Sei $p = 3$, so wird jedes dritte *Feature* innerhalb der *Re-use Unit* ausgewählt. Das *Local Winnowing* (vgl. [Schleimer 2003]) unterscheidet sich vom bereits vorher eingeführten *Global Winnowing* (vgl. Abb. 3.6(b)) nur darin, dass der Algorithmus nicht auf der globalen *Feature*-Menge ausgeführt wird, sondern auf den lokalen *Features* innerhalb einer *Re-use Unit*. Als zugrunde liegende *Hashfunktion* können bspw. gewichtete *PoS*-Tags oder auch die Wortlänge herangezogen werden. Es sei ein *Winnowing*-Fenster $w = 5$ angenommen, so ist garantiert, dass mindestens jedes fünfte innerhalb einer *Re-use Unit* auftauchende *Feature* vom *Digital Fingerprint* in die *Digital Signature* übertragen wird.

Techniken des *Global Selection Usage* mit *Local Selection Knowledge*, wie die *Global Word Class Selection* oder der *Inverted Category Index* (vgl. Abb. 3.7(a)), charakterisieren, dass ihr lokales Wissen, wie ein *PoS*-Tag für ein Wort bzw. eine *PoS*-Tag-Sequenz für *Ngrams*, zu Verteilungen im globalen Kontext führen. So kann ein Wort lokal ein Verb oder Substantiv sein. Im globalen Kontext ergibt sich für ein *Feature* daher eine Verteilung von *PoS*-Tag-Klassen, die zu einem *Score* normalisiert werden muss. Vielmehr können die Wahrscheinlichkeiten der Verteilung zu einem *Feature* bereits als Normalisierungsgewichte eingesetzt werden.

Der *Inverted Category Index* basiert auf Metainformationen zu einem Dokument einer *Digital Library*. Hierbei wird ein lokales *Feature* daraufhin analysiert, wie sehr es für ein Metadatum, wie eine literarische Klassifikation (bspw. *Philosophie*, *Mathematik* oder *Historiographie*), repräsentativ ist. Der *Inverted Category Index* (kurz *ICI*) misst, wie in Formel 3.12 dargestellt, die Zugehörigkeit eines *Features* zu verschiedenen Klassen c_i eines Metadatums.

$$\sigma_{\Theta}^{ICI} = \frac{1}{c_i} \tag{3.12}$$

Aus Formel 3.12 wird ersichtlich, dass die Zugehörigkeit zu mehreren Klassen sukzessive bestraft wird. So kann davon ausgegangen werden, dass Stoppwörter den kleinstmöglichen Score mit $\frac{1}{C}$ mit C der Anzahl der Kategorien eines Metadatums erhalten.

Alle in diesem Abschnitt vorgestellten Techniken sind bereits in diversen Publikationen veröffentlicht und erprobt worden. Das Ziel dieser Techniken ist, aus einem *Digital Fingerprint* eine für eine *Re-use Unit* relevante *Digital Signature* zu erstellen. Diese adressiert die beiden Aspekte der *Geschwindigkeit* sowie der *Genauigkeit* des *Qualitätskriteriums Performance* (vgl. Qualitätskriterien für die *Text Re-use Analysis* in Abschnitt 2.4).

Eine der wesentlichen Fragen des *Selection* ist daher, die Überlappung aus berechneter *Signifikanz* und wissenschaftlicher bzw. fachwissenschaftlicher *Relevanz* zu messen. Für die Beurteilung der Qualität einer *Digital Signature* muss daher entschieden werden, wie gut die Auswahl der selektierten *Features* getroffen worden ist. Aus der Biometrie kann in diesem Zusammenhang die Terminologie der *Minutiae* adaptiert werden. Ein *Minutiae* eines menschlichen Fingerabdruckes entspricht eines von sieben Hauptkomponenten, wie einem *Branch* (vgl. [Maltoni 2009]). Die Terminologie des *Minutiae* kann auch für das *Historical Text Re-use*, wie in Definition 16, adaptiert werden.

Definition 16 (Minutiae)**.** *Sei eine Re-use Unit s_i gegeben. Als ein Minutiae wird ein Feature bezeichnet, welches für die Repräsentation einer Re-use Unit s_i relevant ist.*

Der Unterschied zwischen einem *Minutiae* aus Definition 16 und einem selektierten *Feature* aus der *Digital Signature* besteht in der bereits erwähnten Unterscheidung zwischen informationstechnischer *Signifikanz* und fachwissenschaftlicher *Relevanz*. Aus diesem Verständnis heraus kann neben den quasi als Optimum anzusehenden *Features* einer *Re-use Unit*, den *Minutiae*, auch eine optimale *Signature*, dem *Re-use Nucleus*, wie in Definition 17 manifestiert werden.

Definition 17 (Re-use Nucleus)**.** *Sei eine Re-use Unit s_i gegeben. Ein Re-use Nucleus setzt sich aus den relevanten Minutiae einer Re-use Unit s_i zusammen.*

Aus dieser eingeführten Terminologie des *Re-use Nucleus* ergeben sich unmittelbar zwei relevante Fragestellungen. Einerseits muss gefragt werden, wie stark ein *Re-use Nucleus* im Sinne der *Feature Density* $\mathcal{F}$ einen *Digital Fingerprint* komprimiert. Andererseits ergibt sich auch die Frage, wie gut sich die *Digital Signature* einer *Selection*-Strategie am relevanten Optimum des *Re-use Nucleus* annähert. Aus dieser Betrachtung kann eine solche Analyse als eine *Digital Signature Analysis* verstanden werden, auf die in Abschnitt 5.1 im Detail eingegangen wird.

Zusammenfassend kann das Level der *Selection* als eines der komplexesten vorgestellt werden. Die Wahl der *Selection*-Strategie ist von vielen Parametern abhängig. Neben den bereits vielfach erwähnten *Meme* innerhalb einer *Digital Library* (vgl. Abschnitt 2.6) sind auch Aspekte, wie die Streamingfähigkeit oder die Parallelisierbarkeit, ggf. von Interesse. Vielmehr wird auch deutlich, dass oftmals nicht eine *Selection*-Strategie zum Ziel führt, sondern eine Untermenge der hier genannten Techniken.

Im Kontext des *Re-use Nucleus* und dementsprechend den *Minutiae* stellt sich noch die Frage nach der Effizienz von *Features*. Ein gutes *Feature* zeichnet sich im Idealfall dadurch aus, dass es, wenn es ausgewählt worden ist, möglichst immer Teil des *Re-use Nucleus* ist, so dass ein entsprechender Versuch, andere *Re-use Unit* mit diesem *Feature* zu finden (vgl. *Linking*-Level in Abschnitt 3.6), bei einem Treffer größtmöglichen Erfolg verspricht. Besonders bei einem *Linking* sind in einer verteilten Umgebung die Kosten für ineffiziente *Features* besonders hoch.

3.6 Level 5: Linking

In Anlehnung an das *Information Retrieval* (vgl. [Manning 2008]) kann eine *Digital Signature* einer *Re-use Unit* s_i als eine Anfrage q beschrieben werden, die diejenigen Dokumente $D \backslash \{s_i\}$ gewichtet zurück geliefert bekommt, die am besten auf die Anfrage q passen. Im Kontext der *7-Level-Architektur* für den *Historical Text Re-use* entspricht genau dieser Prozess den beiden folgenden Level des *Linking* und des *Scoring* (vgl. Abschnitt 3.7).

Das *Linking* λ_Θ kann unter drei verschiedenen Aspekten betrachtet werden, wodurch sich entsprechende Techniken klassifizieren lassen:

- *On The Fly Linking* vs. *Precomputation*
- *Intra Digital Library Linking* vs. *Inter Digital Library Linking*
- *Local Linking* vs. *Distributed Linking*

On The Fly Linking wird in interaktiven Umgebungen, wie dem *Crowd Sourcing*, eingesetzt. Dabei wird beim Erstellen einer Online-Edition eine zu bearbeitende Textpassage zu einem Server mit einer indexierten *Digital Library* übertragen, um dort entsprechenden *Text Re-use* festzustellen (vgl. [Geßner 2012]). Auch in anderen *Crowd Sourcing*-Umgebungen, wie *Homer MultiText* (vgl. [Smith 2010]), *SoSOL* (vgl. [Beaulieu 2012]) und *DM2E* (vgl. [Gradmann 2012]), kann *On The Fly Linking* eingesetzt werden, um das Auffinden von *Text Re-use* zu erleichtern bzw. zu beschleunigen. Dem steht eine vollständige *Precomputation* von *Text Re-use* gegenüber. Hierbei liegt der Vorteil im *Serendipity Effect* begründet, der mit Massendatenanalyse einhergeht (vgl. [Büchler 2013b]). Der Nutzen des *On The Fly Linking* hingegen liegt darin, dass sowohl der *Re-use Style* vom Fachwissenschaftler bestimmt aber auch entsprechende Knoten und Kanten eines *Re-use Graph* $G = (V, E)$ getypt werden können (vgl. [Moritz 2013]).

Der Unterschied zwischen einem *Intra Digital Library Linking* und einem *Inter Digital Library Linking* besteht darin, ob *Text Re-use* innerhalb einer *Digital Library* oder zwischen mindestens zwei verschiedenen *Digital Libraries* berechnet werden soll. Hinter beiden *Linking*-Ansätzen stecken jedoch unterschiedliche wissenschaftliche Paradigmen, die durch Aspekte gegenüber gestellt werden können:

- *Größe der Korpora*: Während bei einem *Intra Digital Library Linking* meist sehr große Textmengen verarbeitet werden, sind die Textmengen beim *Inter Digital Library Linking* meist nur zwei oder einige wenige Dokumente groß. Bedingt dadurch ist das gemeinsame Auftreten eines selteneren *Features* bereits ein guter Indikator für einen *Text Re-use* ([Lee 2007]). Gleiche Parameter bei einem *Intra Digital Library Linking* würden jedoch bedeuten, dass Kosten für eine *Text Analyse* progressiv steigen.

- *Abhängigkeitsannahme*: Bei einem *Intra Digital Library Linking* gibt es aufgrund der Datenmenge meist keine Annahmen über die Abhängigkeit zwischen einzelnen *Re-use Units* bzw. Werken ([Clough 2002, Seo 2008, Büchler 2010d]). Dem *Inter Digital Library Linking* liegt gegenteilig die Annahme zugrunde, dass es zwischen den zu untersuchenden Werken, meist einen paarweisen Vergleich zweier Dokumente, einen berechtigten und meist auch offensichtlichen *Text Re-use*-Kontext gibt. Zwei sehr prominente Arbeiten hierzu sind die Bibelvergleiche von John Lee ([Lee 2007]) und Ron Hose ([Hose 2004]). John Lee verglich in seiner Arbeit das Lukas- mit dem Markus-Evangelium. Ron Hose hingegen untersuchte die hebräisch geschriebenen *Dead Sea Scrolls* mit einer hebräischen Übersetzung der Bibel. Hierbei geht es nicht mehr darum, sich des *Serendipity Effects* zu bedienen, sondern vielmehr zu dem bekannten *Text Re-use* alle möglichen Parallelen zwischen den Werken aufzudecken.

Aus informationstechnischer Sicht wird der Unterschied beim Formalisieren deutlich. Es sei eine (n, m')-*Signature*-Matrix S gegeben. Ein *Intra Digital Library Linking* λ_Θ entspricht der Matrixmultiplikation von S mit der transponierten Matrix $S^\intercal$ (vgl. Formel 3.13).

$$L = S \cdot S^\intercal \tag{3.13}$$

Bezogen auf das in Formel 3.8 eingeführte Beispiel mit der *Signature*-Matrix aus Formel 3.10 wird die *Link*-Matrix L in Formel 3.14 generiert. Ein Wert $L_{i,j}$ gibt die Anzahl der überlappenden *Features* zweier *Re-use Units* s_i und s_j bzgl. ihrer *Digital Signature* an. Weiterhin ist L bzgl. ihrer Hauptdiagonalen immer symmetrisch.

$$\begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{array}{c} \begin{array}{ccccc} s_1 & s_2 & s_3 & s_4 & s_5 \end{array} \\ \begin{pmatrix} 0 & 3 & 3 & 3 & 1 \\ 3 & 0 & 4 & 5 & 0 \\ 3 & 4 & 0 & 4 & 1 \\ 3 & 5 & 4 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array} = L \tag{3.14}$$

Beim *Inter Digital Library Linking* wird aus der (n, m')-*Signature*-Matrix S sowie deren transponierte Matrix keine (n, n)-*Linking*-Matrix generiert, sondern eine (n, n')-Matrix, die aus der Matrixmultiplikation von S für eine *Digital Library* sowie einer zweiten transponierten (n', m')-*Signature*-Matrix $T^\intercal$ einer zweiten *Digital Library* resultiert.

$$L = S \cdot T^\intercal \tag{3.15}$$

Das *Inter Digital Library Linking* begleitet hierbei zwei Nebenbetrachtungen. Einerseits können nicht wie beim *Intra Digital Library Linking* alle *Features* durch ein *min pruning* aus der *Feature*-Menge mit der Quantität m' entfernt werden, die nur einmal auftreten, da der *Match* nicht innerhalb S stattfindet, sondern mit den *Digital Signature* einer zweiten *Digital Library*. Dies bedeutet, dass der Index für die *Digital Signature* ungleich größer ist als beim *Intra Digital Library Linking*. Andererseits ist durch die beiden (n, m')- und (n', m')-*Signature*-Matrizen S und T auch ein gemeinsamer *Feature*-Vektor der Länge m' nötig (vgl. die Problemklassen des *Preprocessing* in Abschnitt 3.3).

Sowohl das *Intra Digital Library Linking* als auch das *Inter Digital Library Linking* λ_Θ haben eine quadratische Laufzeitkomplexität $O(n^2)$. Für jedes *Feature* einer *Digital Signature* wird ein Link zu allen anderen *Re-use Units* gesetzt, die das gleiche *Feature* beinhalten. Insbesondere bei großen *Digital Libraries* können sowohl das Mapping von einer *Re-use Unit* auf ihre *Features* als auch von den *Features* auf die *Re-use Units* nicht im Hauptspeicher vorgehalten werden, wodurch ein *Caching* nötig ist. Die Wahl der *Caching*-Strategie ist jedoch stark von der Verteilung der Daten abhängig (vgl. [Olaru 2004]).

Ausgehend von der (n, m')-*Signature*-Matrix S können die Längen der einzelnen *Re-use Units* aus S durch eine rechtsseitige Matrixmultiplikation eines Einheitsvektors E der Länge m' bestimmt werden (vgl. Formel 3.16).

$$S^s = \begin{pmatrix} len(s_1) \\ len(s_2) \\ \vdots \\ len(s_n) \end{pmatrix} = S \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{3.16}$$

Auf die gleiche Weise kann auch die Verteilung der *Features* durch eine linksseitige Matrixmultiplikation mit einem transponierten Einheitsvektor der Länge n bestimmt werden (vgl. Formel 3.17).

$$S^{w\intercal} = \begin{pmatrix} freq(w_1) \\ freq(w_2) \\ \cdots \\ freq(w_{m'}) \end{pmatrix}^{\intercal} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}^{\intercal} \cdot S \tag{3.17}$$

Im Beispiel der *Signature*-Matrix S aus Formel 3.10 ergeben sich somit die Vektoren $S^s = \begin{pmatrix} 5 & 5 & 5 & 5 & 5 \end{pmatrix}$ und $S^w = \begin{pmatrix} 4 & 4 & 3 & 3 & 3 & 3 \end{pmatrix}$. S^s ist leicht überprüfbar, da alle *Re-use Units* s_i aus dem Beispiel genau 5 Wörter enthalten. S^w kann durch spaltenweises Aufsummieren ebenfalls einfach überprüft werden.

Die Verteilung aus S^s ist vom ersten Level, der *Segmentation*, abhängig. Bei einer *Moving Window*-Segmentierung (vgl. Abb. 3.1) haben wie im Beispiel aus Formel 3.8 alle *Re-use Units* die gleiche Länge. Bei einer disjunkten *Sentence*-weisen Segmentierung hingegen folgt die Verteilung der Satzlängen der *negativen Binomialverteilung* (vgl. [Kelih 2005]).

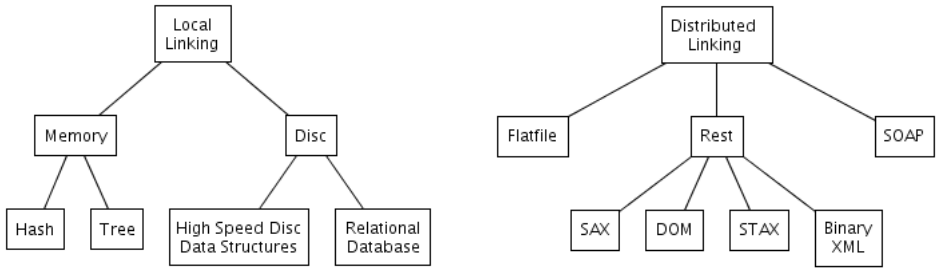
Die Verteilung der Features aus S^w hingegen kann in den meisten Fällen als eine *Power Law*-Verteilung angenommen werden, was ein effizientes *Caching* oftmals sehr einfach gestaltet. In wenigen Ausnahmen kann auch eine *negative Binomialverteilung* erkannt werden. Dies geschieht bspw. dann, wenn im *Preprocessing* eine *Word Length Reduction* eingesetzt wird (vgl. Abb. 3.2). Die Wahl eines guten und verteilungsabhängigen *Caching* (vgl. u. a. [Olaru 2004]) hat oftmals enorme Auswirkungen auf die *Performance* (vgl. *Qualitätskriterien* in Abschnitt 2.4). Nicht selten kann bei einem falschen *Caching* ein Sprung von einer Einheit, wie *Prozessortage* zu *Prozessorwochen* oder gar *Prozessormonate*, beobachtet werden.

Neben den beiden bereits eingeführten Klassifizierungen von *Linking*-Techniken kann auch zwischen einem *Local Linking* und einem *Distributed Linking* unterschieden werden. Das *Local Linking* (vgl. Abb. 3.8(a)) wird beispielsweise bei einer kompletten *Precomputation* des *Text Re-use* innerhalb einer *Digital Library* eingesetzt (vgl. u. a. [Büchler 2012c]). Hierbei können Zugriffstechniken als *Memory*- und *Disc*-basiert klassifiziert werden. Zu den *In-Memory*-Zugriffstechniken[42] zählen u. a. verschiedene *Hash*- und *Tree*-basierte Datenstrukturen, die Zugriffe entweder konstant in $O(1)$ oder logarithmisch in $O(\log(n))$ ermöglichen (vgl. [Ottmann 1996, Knuth 1997a]). Diese Techniken werden auch für die Erstellung von *Indexstrukturen* in Datenbanken eingesetzt. Die Qualität der *Disc*-basierten Ansätze (vgl. Abb. 3.8(a)) sind jedoch oftmals von der Implementierung und dem *Overhead* des darunterliegenden Protokolls abhängig. So ist eine Softwarebibliothek wie *cooccaccess* (vgl. [Bordag 2008]) etwa 10- bis 20mal schneller als eine relationale Datenbank[43].

Dem *Distributed Linking* (vgl. Abb. 3.8(b)) liegt meist ein XML-basiertes Protokoll, wie *SOAP*, zugrunde, welches zwei Arten von Daten enthält. Einerseits umfasst dies die inhaltsbehafteten Daten mit dem Mapping einer *Re-use Unit* s_i auf die *Features* $\overrightarrow{s_i'}$ der *Digital Signature* und andererseits XML-basierte Strukturdaten, wie Tags. Das Verhältnis aus diesen beiden Daten fällt bei einfachen *Flatfiles* um ein Vielfaches besser aus, als bspw. bei *SOAP*-Nachrichten, auf die kein Einfluss genommen werden kann. Bei *Rest*-basierten Protokollen jedoch kann dieses Verhältnis entsprechend den Anforderungen optimiert werden. In [Sander 2010] wurde dieses Optimierungsproblem auf verschiedene Nachrichtengrößen untersucht. Unter anderem wurde herausgefunden, dass ab einer gewissen Größe, die

[42] bspw. das *Collection*-Package in Java Development Kit oder die *Primitive Collection for Java* (vgl. `http://pcj.sourceforge.net/`

[43] *cooccaccess* wird sowohl seit sechs Jahren in *Medusa* als auch im *TRACER*-Tool erfolgreich eingesetzt. Auch wenn natürlich ein Unterschied zwischen *RAM*- und *DISC*-basiertem Zugriff mit *cooccaccess* messbar ist, so scheint diese Bibliothek für eine derartige *Linking*-Aufgabe die schnellste Standalone-Implementierung zu sein, ohne dass erst ein Index einer Datenbank optimiert werden muss.



(a) Taxonomy des *Local Linking*

(b) Taxonomy des *Distributed Linking*

Abbildung 3.8: Taxonomie des Level *Linking* für den *Historical Text Re-use*. Die Abbildung zeigt die Klassen des *Local Linking* und *Distributed Linking*.

relativen Kosten pro Datensatz bei SOAP-Nachrichten wieder steigen, wodurch *SOAP* bedingt durch diesen Effekt für größere Anwendungen nachteilig wird. Auf der anderen Seite hängt bei *Rest*-basierten und verteilten Services (vgl. [Büchler 2005]) die Geschwindigkeit auch davon ab, ob ein *DOM*, *Stax* oder *Binary XML* generiert bzw. verarbeitet wird. Der Unterschied liegt hierbei darin, ob ein *Memory*-basiertes Modell, wie *DOM*, eine *Streaming*-basierte Technologie, wie *Stax*, welche auch sehr große XML-Outputs effizient verarbeitet, oder eine binäre Repräsentation, wie *Binary XML*[44], eingesetzt werden. Mit *Binary XML* konnte in unveröffentlichten Analysen im Rahmen der Entwicklung von *TRACER* (vgl. [Büchler 2013a]) eine Verbesserung der *Performance* von bis zu 40% beim *Linking* gegenüber einer *DOM*-basierten *Rest*-Anwendung festgestellt werden.

3.7 Level 6: Scoring

Wie eingangs zum vorangestellten Abschnitt 3.6 bereits erwähnt, bilden die *Linking*- und *Scoring*-Level den Prozess des *Information Retrieval* ab. *Scoring* kann im *Information Retrieval* mit dem *Vector Space Model* (vgl. [Manning 2008]) verglichen werden. Hierbei werden zwei Termvektoren einer Anfrage und eines Dokumentes verglichen. Die Terme selbst sind durch ein *Term Weighting* beschrieben. Hierbei bekommen durch eine $tf.idf$-Gewichtung (vgl. [Salton 1975]) sowohl häufige als auch seltene Terme einen niedrigen *Score*.

Genau dieser letztere Prozess des Gewichtens eines paarweisen *Re-use Overlap* zweier *Re-use Unit* s_i und s_j aus Formel 3.14 ist Gegenstand des *Scoring* θ_Θ. Die Schwierigkeit hierbei ist, dass Metriken, wie das $tf.idf$-*Measure*, zwar in *Information Retrieval*-Systemen auf Dokumentenebene eingesetzt werden und dort auch funktionieren. Jedoch wird es bei kleineren *Re-use Units*, wie einem *Sentence* oder gar einem noch kleineren *Moving Window*, schwierig, dies durch *Term Weighting*-Techniken auszudrücken. Auch wenn Metriken, wie $tf.idf$, für die Gewichtung beim *Text Re-use* eingesetzt werden (vgl. [Hose 2004, Metzler 2005, Barrón-Cedeño 2010a]), so sei sich dem Shakespeare-Zitat *Sein oder nicht sein, das ist hier die Frage.* bedient, um aufzuzeigen, dass gerade beim *Text Re-use* die enthaltene Semantik nicht oder nur schwer an einzelne Wörter gebunden ist.

Vielmehr wird bereits während der *Selection* (vgl. Abschnitt 3.5) durch verschiedene Strategien entschieden, welche *Features* als wichtig erachtet werden und welche nicht. Bedingt dadurch haben sich in der *Scientific Community* eher einfache mengenorientier-

[44] vgl. bspw. die Implementierung *bnux*, `http://acs.lbl.gov/software/nux/`

te und ungewichtete Metriken, wie die *Resemblance* oder das *Containment* (vgl. beide [Broder 1997a]), etabliert. Während die *Resemblance* $\theta^R_\Theta(s_i, s_j) = \frac{|\overrightarrow{s_i} \cap \overrightarrow{s_j}|}{|\overrightarrow{s_i} \cup \overrightarrow{s_j}|}$ ein symmetrisches Maß mit $\theta(s_i, s_j) = \theta(s_j, s_i)$ ist, kann das *Containment* $\theta^C_\Theta(s_i, s_j) = \frac{|\overrightarrow{s_i} \cap \overrightarrow{s_j}|}{|\overrightarrow{s_i}|}$ zu den asymmetrischen *Scoring*-Maßen mit $\theta(s_i, s_j) \neq \theta(s_j, s_i)$ gezählt werden.

Scoring-Maße können sehr vielfältig ausgewählt werden. So kann bspw. auch der *Levenshtein*-Abstand als ein Maß verstanden werden, welches den Abstand zweier *Re-use Units* s_i und s_j bestimmt. Prinzipiell können *Scoring*-Techniken wie folgt eingeteilt werden:

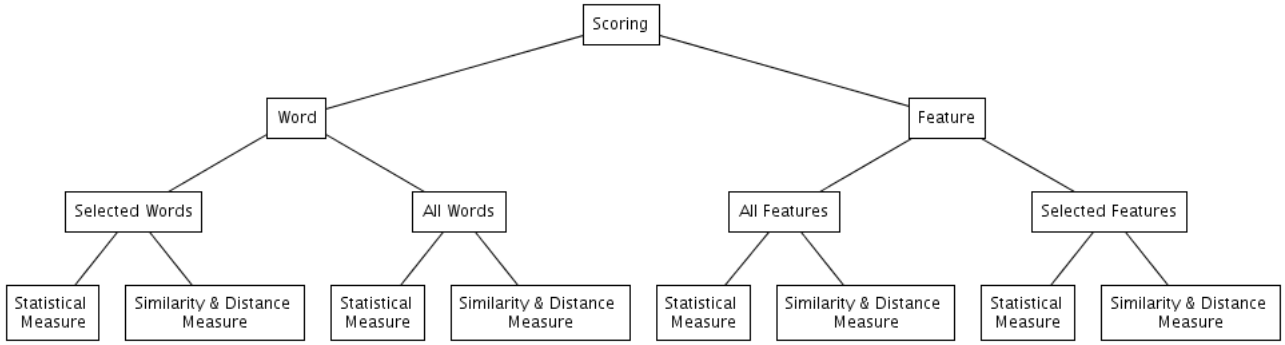
- *Scoring*-basierend auf den *Features* vs. auf Basis der Wörter einer *Re-use Unit*,
- *Scoring* auf Basis der selektierten Features einer *Digital Signature* bzw. den daraus resultierenden Wörtern vs. allen *Features* oder Wörtern sowie
- *Scoring* auf Basis von *Statistical Measure* vs. *Similarity Measure*.

Die Unterscheidung zwischen *Scoring* auf *Word*- bzw. *Feature*-Ebene (vgl. Abb. 3.9) scheint im Fokus der *Digital Signature* (vgl. Abschnitt 3.5) nicht sinnvoll. Es hat sich jedoch in der fachwissenschaftlichen Anwendung, wie in *eAQUA*[45], gezeigt, dass zwischen einem technischen *Scoring* auf der *Feature*-Ebene, um eine qualifizierte Entscheidung zu treffen, ob ein *Text Re-use Candidate* akzeptiert werden kann, und einem anzuzeigenden *Score* in einer fachwissenschaftlichen Anwendung unterschieden werden muss. Beim *Preprocessing* ist es das Ziel, zwei daraus resultierende *Digital Signatures* durch das Entfernen von Varianten so zu normalisieren, dass beide *Re-use Units* auch während des *Scoring* bzgl. eines Schwellwertes als hinreichend ähnlich erkannt werden. In Abb. 1.1 auf Seite 35 wurde bereits die *Microview* vorgestellt, welche die Varianten eines Zitates aufzeigt. In den meisten Fällen werden diese Varianten während des *Preprocessing* entweder normalisiert oder während der *Selection* aufgrund ihrer Seltenheit entfernt. So ist oft zu beobachten, dass sich zwei *Re-use Units* s_i und s_j bzgl. ihrer *Digital Signature* wesentlich ähnlicher oder gar identisch sind, wenn gleich ein Fachwissenschaftler beim Benutzen dieser Daten eine deutlich größere Unähnlichkeit auf Basis aller Wörter feststellt. Aus diesem Grund empfiehlt sich daher oftmals, das technische *Scoring* auf den *Features* der *Digital Signature* zu bestimmen und für den daraus akzeptierten *Text Re-use* einen *Score* auf *Word*-Ebene basierend zu berechnen, der wiederum auch von Sprachwissenschaftlern, wie Gräzisten und Latinisten, akzeptiert und nachvollzogen werden kann.

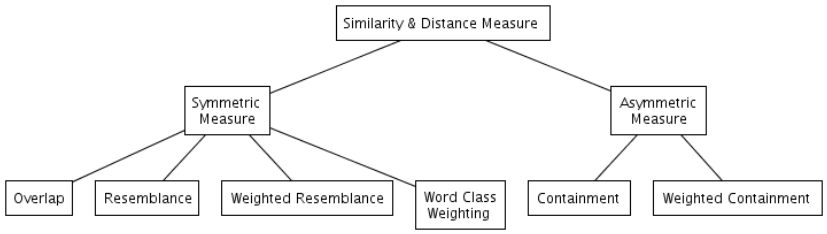
Mit der vorigen Diskussion einhergehend kann beim *Scoring* unterschieden werden, ob er auf allen *Words* bzw. *Features* oder nur auf der Datenbasis der *Digital Signature* berechnet wird (vgl. Abb. 3.9). Für den technischen *Score* empfiehlt sich, auf den selektierten *Features* den *Score* zu berechnen, während sich für den *Score*, welcher einem Nutzer angezeigt wird, der Vergleich aller *Words* oder *Features* zweier *Re-use Units* empfiehlt.

Der *Re-use Overlap* $\theta^O_\Theta(s_i, s_j) = |\overrightarrow{s_i} \cap \overrightarrow{s_j}|$ aus Abb. 3.9(b) kann als das einfachste aller *Similarity Measures* verstanden werden. Jedoch ist dieses Maß bereits nicht mehr geeignet, sobald die *Re-use Units* unterschiedlich lang sind. Ein *Re-use Overlap* zweier kurzer *Re-use Units* ist wesentlich höher zu bewerten als ein gleich großer *Overlap* zwischen zwei längeren *Re-use Units*. In [Broder 1997a] wird das Problem der Vergleichbarkeit durch die bereits eingeführten *Resemblance* $\theta^R_\Theta(s_i, s_j)$ und *Containment* $\theta^C_\Theta(s_i, s_j)$ gelöst, so dass der *Re-use Overlap* durch $|\overrightarrow{s_i} \cup \overrightarrow{s_j}|$ bzw. $|\overrightarrow{s_i}|$ mit $\theta^R_\Theta(s_i, s_j) \in [0, 1]$ und $\theta^C_\Theta(s_i, s_j) \in [0, 1]$ normalisiert wird. Hierbei kann wiederum ein *Score* für zwei *Re-use Units* s_i und s_j den gleichen Wert

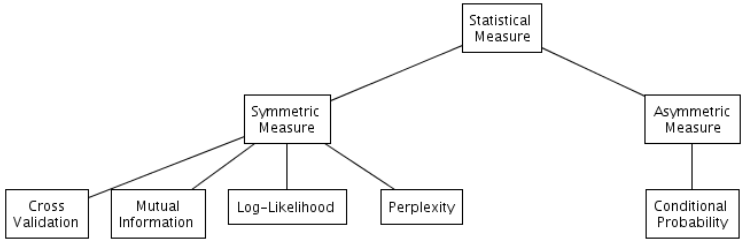
[45] In *eAQUA* wurde dem Fachwissenschaftler der *Score* auf *Feature*-Ebene angezeigt. Dies hat jedoch den Nachteil, dass immer wieder die Frage aufkam, was dieser *Score* bedeutet. Da eine *Selection* vorangestellt ist, kann in einer Anwendung nachträglich nicht mehr transparent festgestellt werden, welche *Features* Teil des *Scoring* gewesen sind.



(a) Taxonomie des Level *Scoring* für den *Historical Text Re-use*.



(b) Überblick über *Similarity*- und *Distance* Measure



(c) Überblick über *Statistical Measures*

Abbildung 3.9: Taxonomie des Level *Scoring* für den *Historical Text Re-use*. Diese Taxonomie schlüsselt verschiedene Techniken auf. Die Unterscheidung zwischen einem *Scoring* auf der *Word*- und *Feature*-Ebene hängt vom Einsatz ab. Während die *Feature*-Ebene aus technischer Sicht zu bevorzugen ist, sind für Anzeigen in Benutzerschnittstellen oftmals die *Word*-Ebene besser geeignet, da die *Scoring*-Werte einfacher nachzuvollziehen sind.

sowohl für zwei als auch 20 *Features* oder *Words* im *Re-use Overlap* haben, wobei Letzterer selbst bei gleicher Ähnlichkeit wesentlich signifikanter ist. Dieses Problem wurde im Rahmen von *TRACER* so gelöst, dass eine gewichtete *Resemblance* durch

$$\theta_{\Theta}^{R'}(s_i, s_j) = \theta_{\Theta}^{O}(s_i, s_j) \cdot \theta_{\Theta}^{R}(s_i, s_j) = \frac{|\overrightarrow{s_i} \cap \overrightarrow{s_j}|^2}{|\overrightarrow{s_i} \cup \overrightarrow{s_j}|} \tag{3.18}$$

und ein gewichtetes *Containment* durch

$$\theta_{\Theta}^{C'}(s_i, s_j) = \theta_{\Theta}^{O}(s_i, s_j) \cdot \theta_{\Theta}^{C}(s_i, s_j) = \frac{|\overrightarrow{s_i} \cap \overrightarrow{s_j}|^2}{|\overrightarrow{s_i}|} \tag{3.19}$$

mit $\theta_{\Theta}^{R'}(s_i, s_j) \in \mathbb{R}^+$ und $\theta_{\Theta}^{C'}(s_i, s_j) \in \mathbb{R}^+$ eingeführt wurde (vgl. [Büchler 2011c]).

Das *Word Class Weighting* hingegen nutzt Wortartklassen, um den *Overlap* zu gewichten. Hierbei bekommen Substantive und Verben einen hohen *Score*, während Wortartklassen für Stoppwörter mit Ausnahme der *Personalpronomen* einen niedrigen *Score* erhalten. Personalpronomen als eine Klasse von Stoppwörtern haben bei einer *Text Re-use Analysis* in Verbindung mit einem *Verbum Dicendi*, wie *sagen* oder *schreiben*, die Funktion, dass sie einen *Text Re-use* oftmals einleiten oder abschließen.

Neben den *Similarity Measures* können auch *Statistical Measures* für die Gewichtung des *Re-use Overlaps* eingesetzt werden (vgl. Abb. Abb. 3.9(c)). Hierzu zählen u. a. die *Mutual Information* (vgl. [Church 1990]), das *Log-Likelihood-Ratio* (vgl. [Dunning 1993]), die *Perplexity* (vgl. [Kenne 1996]) aber auch die *Conditional Probability*. Während letzteres Maß sich asymmetrisch $\theta(s_i, s_j) \neq \theta(s_j, s_i)$ verhält, sind die zuvor genannten statistischen Maße allesamt symmetrisch $\theta(s_i, s_j) = \theta(s_j, s_i)$.

Die *Statistical Measures* werden derzeit weitestgehend nicht im Bereich einer *Text Re-use Analysis* eingesetzt. Dies kommt mit dem Fakt daher, dass *Features* einer *Power Law*-Verteilung mit vielen seltenen *Features* folgen, wodurch ein *Overlap* aus mehreren *Features* nahezu immer statistisch signifikant ist. In Abschnitt 4.2 wird gezeigt, dass etwa zwei Drittel aller *Bigrams* selbst bei einmaligem Auftreten bereits statistisch signifikant sind. In [Church 1990] wird daher auch vorgeschlagen, nur diejenigen *Bigrams* oder *Co-occurrences* in Betracht zu ziehen, die mindestens fünfmal beobachtet werden. Auf der anderen Seite wird in [Büchler 2010d] gezeigt, dass ab fünf *Words* im *Overlap* nahezu alle Überlappungen selbst von Stoppwörtern immer als statistisch signifikant bzw. auffällig gelten.

Aufgrund der Gewichtung der *Features* während der *Selection* liegt der vornehmliche Einsatz von *Scoring*-Techniken im Bereich der *Similarity Measures*. Bezogen auf das in Formel 3.8 eingeführte Beispiel mit der *Signature*-Matrix aus Formel 3.10 sowie der *Link*-Matrix L aus Formel 3.14 entspricht eine *Resemblance*-basierte *Scoring*-Matrix T^R mit $T^R = \theta_{\Theta}^{R}(s_i, s_j)$ der Matrix in Formel 3.20.

$$\begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{array}{c} \begin{array}{ccccc} s_1 & s_2 & s_3 & s_4 & s_5 \end{array} \\ \begin{pmatrix} 0 & 0.43 & 0.43 & 0.43 & 0.11 \\ 0.43 & 0 & 0.66 & 1.00 & 0 \\ 0.43 & 0.66 & 0 & 0.66 & 0.11 \\ 0.43 & 1.00 & 0.66 & 0 & 0 \\ 0.11 & 0 & 0.11 & 0 & 0 \end{pmatrix} \end{array} = T^R \tag{3.20}$$

Für eine *Text Re-use Analysis* reicht es in den meisten Fällen aus, wenn eine *Scoring*-Matrix T, wie in Formel 3.20, bzw. der daraus resultierende *Re-use Graph* berechnet wird.

3.8 Level 7: Postprocessing

Auch wenn für die meisten geisteswissenschaftlichen Anwendungen das Auffinden von *Text Re-use* bereits genügt, so kann ein gutes *Postprocessing* sehr dabei helfen, die Daten zu verstehen, den *Re-use Graph* zu säubern aber auch genau diejenigen Formen von *Text Re-use* zu identifizieren, die nicht zu erwarten sind, um die Fachwissenschaften explizit darauf hinzuweisen und somit nicht nur neue Methoden zu etablieren, sondern vor allem auch schnelleren Zugriff auf Neues bzw. Unbekanntes zu geben. Das *Postprocessing* ist daher immer im Kontext des *Information Overload vs. Information Poverty* aus Abschnitt 1.4 zu sehen. Eine *Text Re-use Analysis* produziert auf einer großen *Digital Library* ebenfalls sehr viele *Text Re-use Data*, wodurch sich der Fachwissenschaftler schnell in der Datenmenge verlieren kann. Das *Postprocessing* kann daher als ein Nachbearbeitungsschritt verstanden werden, der für eine konkrete fachwissenschaftliche Fragestellung einen *Re-use Graph* entsprechend nachbereitet und ggf. nicht relevante Daten dafür entfernt bzw. ignoriert.

Das *Postprocessing* eines *Re-use Graph* G bzw. einer *Scoring*-Matrix T wird sehr vielschichtig gestaltet. Abb. 3.10 zeigt daher nur einen Umriss möglicher *Postprocessing*-Schritte auf. Jede *Graph Mining*-Technik (vgl. [Aggarwal 2010b]) kann faktisch auch auf einem *Text Re-use Graph* angewendet werden. Mögliche *Postprocessing*-Schritte können in die Klassen *Bibliometry*, *Phonetic Postprocessing*, *Text Re-use Tasks*, *Cluster Analysis* sowie *Further Graph Mining Approaches* eingeteilt werden[46] (vgl. Abb. 3.10).

Von den fünf *Postprocessing*-Clustern wird *Bibliometry* nicht nur als eine Klasse von Techniken sondern vielmehr als ein eigener Wissenschaftsbereich (vgl. [Archambault 2004]) aufgefasst. Bibliometrische Analysen umfassen mehr oder weniger alle Analysen, um Dokumente bzw. Abhängigkeiten zwischen ihnen aufzudecken, wie es auch durch *Text Re-use* gegeben ist (vgl. *Einführung in die Bibliometrie* in [Havemann 2009]). Neben Gesetzmäßigkeiten, wie *Lotka's Law* (vgl. [Lotka 1926]), beschreiben Kennzahlen, wie der *Impact Factor* (vgl. u. a. [Havemann 2009] oder der *h-Index* (vgl. [Hirsch 2005]), wie wichtig ein Dokument bzw. Journal ist, was anhand der von einem Autor gegebenen Abhängigkeit durch Zitieren anderer Werke gemessen wird. Einen ähnlichen Grundgedanken verfolgen Brin & Page, die einen zugrunde liegenden *Hypertext* im Web dazu nutzen, um Dokumente mittels *PageRanking* innerhalb eines *Information Retrieval*-Systems zu gewichten[47] (vgl. [Brin 1998]). Die *PageRank*-Technik ist jedoch nicht nur auf Webtexte eingeschränkt, sondern kann auf jede Form eines *Hypertextes* angewendet werden, wie auch dem *Text Re-use Graph* (vgl. [Büchler 2012b]). Neueste Ergebnisse hingegen zeigen, dass innerhalb eines Zitationsgraphen wenige Dokumente den Graphen dominieren (vgl. [Barabási 2012]), so dass vermutet werden kann, dass die *PageRank*-Technik durch den wesentlich einfacher zu bestimmenden *h-Index* approximiert werden kann.

Die Klasse des *Phonetic Postprocessing* ist in der Informatik weitestgehend nicht verbreitet und sehr aus dem Bereich der Geisteswissenschaften forciert (vgl. [Coffee 2012b]). Das *Phonetic Postprocessing* umfasst alle Techniken, die sich mit der Aussprache und Stimmlage beschäftigen. Neben der metrischen Analyse (vgl. Abb. 3.10), in welcher das Versmaß einer Wortsequenz (vgl. [Bobenhausen 2009, Küper 2011]) bestimmt wird, sind auch die *Alliteration*- und *Rhyming*-Techniken für das *Text Re-use Postprocessing* von großem Interesse.

[46] Insbesondere beim *Postprocessing* ist die Klassifizierung nicht immer eindeutig. So kann faktisch nahezu jede bibliometrische Analyse auf einem *Text Re-use Graph* auch unter *Graph Mining* klassifiziert werden.

[47] Die Google zugrunde liegende *PageRank*-Technik wurde absichtlich nicht in *Graph Mining* in Abb. 3.10 klassifiziert, sondern in die Menge der bibliometrischen Analysen, da die Nähe des Grundgedankens zur Bibliometrie offensichtlich ist. Der Unterschied besteht letztlich eher in den Daten. Während bei *bibliometrischen Analysen* der *Hypertext* durch Zitationsabhängigkeiten gegeben ist, sind es beim *PageRanking* zumindest in [Brin 1998] Links in Webdokumenten. Beide haben jedoch gemeinsam, dass sowohl die explizite Angabe des Verweises als auch eine gerichtete Kante gegeben sind.

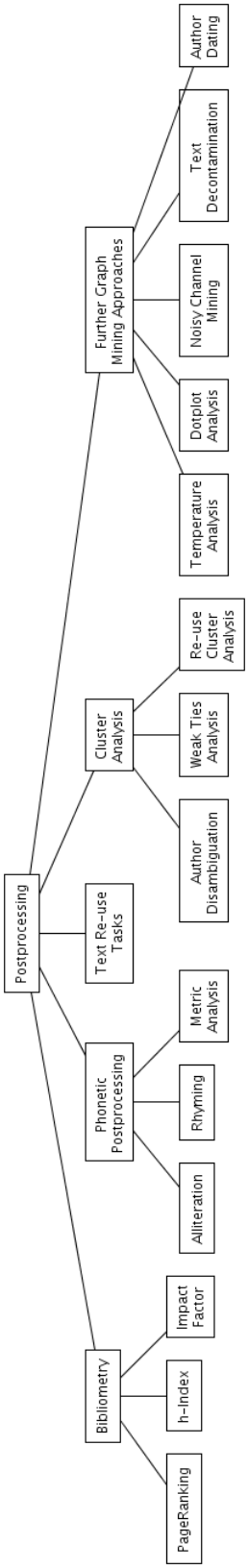


Abbildung 3.10: Taxonomie des Level *Postprocessing* für den *Historical Text Re-use*. Diese Taxonomie schlüsselt verschiedene Techniken aus der Bibliometrie, der phonetischen Analyse, den *Text Re-use Tasks* aus Abschnitt 2.7, den *Clusteranalysen*, sowie weiteren Techniken des *Graph Mining* auf. Gemeinsames Ziel aller *Postprocessing*-Techniken ist das Vereinfachen der Daten im Sinne des *Information Overload* vs. der *Information Poverty* aus Abschnitt 1.4.

Sowohl eine *Alliteration*, wie die v-Alliterationen in *veni, vidi, vici* oder in *in vino veritas*, als auch ein *Rhyming*, wie in *Viele Hände, schnelles Ende.* oder *Nicht suchen, buchen.*, können bei einem kurzen *Text Re-use* von weniger als fünf Wörtern helfen, um zwischen einem *Text Re-use* und dem *Language Re-use*, wie bspw. *Co- occurrences*, *Bigrams* oder *Trigrams*, zu unterscheiden.

Die Klasse der *Cluster Analysis* von *Text Re-use Graphs* beschreibt (vgl. [An 2004]) oder analysiert (vgl. [Bolelli 2006]) *Text Re-use Data*. Solche *Cluster*-Analysen können sehr vielschichtig sein. In einer *Re-use Cluster Analysis* wird ein *Re-use Graph* so partitioniert (vgl. [Aggarwal 2010b]), dass möglichst natürlich nachvollziehbare *Re-use Cluster* gebildet werden, die nicht mit anderen *Clustern* verbunden sind (vgl. Adaption für *Text Re-use Graphs* aus [Bolelli 2006]). Bei der *Re-use Cluster Analysis* werden hierbei sukzessive schwache bzw. schwach signifikante *Relationen* zwischen zwei *Re-use Units* entfernt. Dieser Form der Analyse steht die *Weak Ties Analysis* (vgl. [Granovetter 1983]) gegenüber. Granovetter zeigte wiederum, dass insbesondere die *Weak Ties* in *Graphs* von netzwerktheoretischem Interesse sind, da sie nicht nur *Cluster* verbinden, sondern auch seither fundamentaler Bestandteil von Graphmodellierungen sind (vgl. [Barabási 2003]). In diesem Sinne kann eine *Weak Ties Analysis* eingesetzt werden, um speziell dem *Serendipity Effect* aus *Text Re-use Graphs* (vgl. [Büchler 2013b]) so auszunutzen, dass ein nicht zu erwartender *Text Re-use* besonders hervorgehoben wird, um in einer fachwissenschaftlichen Anwendung entsprechend den Nutzer darauf hinzuweisen. Eine Technik, um *Weak Ties* zu messen und zu bestimmen, ist der Ansatz der *Contrastive Semantic* (vgl. [Büchler 2010f]). Weiterhin kann ein *Re-use Graph* auch dazu eingesetzt werden, um ein *Author Disambiguation* durchzuführen (vgl. [McRae-Spencer 2006]). Hierbei werden Verhältnisse zwischen Dokumenten bzw. Textpassagen aber auch *Self Text Re-use* gemessen, um unabhängig von Metadaten einen Autor zu bestimmen. So gibt es im Kanon der griechischen Literatur bspw. zwei verschiedene Personen mit dem Namen *Platon*. Einerseits kann der berühmte Philosoph aus dem 4. Jh. v. Chr. gemeint sein. Andererseits ist von Platon, dem Komiker, die Rede. Durch die Zitationsmuster ist es möglich, auf diese Weise ein *Pattern*, wie *Platon sagte*, zu disambiguieren.

Zu der letzten Klasse der hier vorgestellten *Postprocessing*-Techniken (vgl. Abb. 3.10) zählen u. a. die *Temperature Analysis* (vgl. [Büchler 2013c] sowie Abb. 1.3 auf Seite 36), die *Dotplot Analysis* (vgl. [Lee 2007] sowie Abb. 1.2 auf Seite 35) und das *Noisy Channel Mining* (vgl. Abschnitte 1.7 und 5.5 sowie [Büchler 2011c]). Ergänzend dazu können auch die *Text Decontamination* aber auch das *Author Dating* dieser Klasse zugewiesen werden. Beim *Text Decontamination* wird ein *Re-use Graph G* dazu eingesetzt, um im Rahmen eines Vorverarbeitungsschrittes den *Text Re-use* zu entfernen, um wiederum die Sprachstatistik für andere Lernverfahren zu verbessern. In Abschnitt 3.3 wurde eingangs ausgeführt, dass bedingt durch die *Power Law*-Verteilung von *Co-occurrences* oder *Bigrams* 95% dieser Daten fünfmal und seltener auftreten. Dementsprechend wird die Sprachstatistik durch nicht *Text Re-use*-dekontaminierte Daten verfälscht. Eine *Text Re-use Analysis* tritt hierbei als eine Hilfswissenschaft für andere Bereiche der *Text Analysis* auf.

Weiterhin kann ein *Text Re-use Graph* als ein komplexes Konstrukt von Abhängigkeiten verstanden werden. Insbesondere bei antiken Texten sind oftmals die Datierungen von Autoren oder gar Werken schwierig. Ein gerichteter *Text Re-use Graph* kann vielmehr dazu eingesetzt werden, Datierungen für eine Person oder ein Werk auf Basis des *Text Re-use* sukzessive einzuschränken.

3.9 Wechselwirkungen zwischen den einzelnen Level

Die in diesem Kapitel vorgestellte *7-Level-Architektur* repräsentiert eine komplexe Softwarearchitektur, mit welcher durch alle Implementierungen auf den jeweiligen Level insgesamt über eine Million Kombinationen möglich sind, um mit den verschiedenen *Meme* und *Re-use Styles* (vgl. geisteswissenschaftliche *Systematisierung* des *Historical Text Re-use* in Abschnitt 2.6) umgehen zu können. Jedoch sind diese Permutationen nicht frei von Wechselwirkungen. Ziel dieses Abschnittes ist es, kurz auf die wichtigsten Wechselwirkungen einzugehen. Auch wenn es verschiedene fachwissenschaftliche Aspekte hierzu gibt, so liegt der technische Fokus in diesem Abschnitt auf der Geschwindigkeit einer *Text Re-use Analysis*.

Bei der Segmentierung kann eine *Sentence*-basierte *Segmentation* mit einer *Moving Window*-Segmentierung verglichen werden (vgl. [Büchler 2012c]). Während die erste Segmentierung zu der *Disjoint Segmentation* zählt, ist Letztere der *Overlapping Segmentation* zuzuordnen. Der Unterschied zwischen beiden kann schnell verdeutlicht werden. Es sei eine *Sentence*-segmentierte Zerlegung gegeben, dann kann aus jeder disjunkten *Re-use Unit* s_i eine Menge von überlappenden *Moving Windows* generiert werden. Es sei weiterhin zur Vereinfachung eine fixe Satzlänge l gegeben. Bei einer fixen Länge des *Moving Window* von w werden somit $n \cdot (l - w + 1)$ überlappende *Re-use Units* erzeugt. Bei einer durchschnittlichen Satzlänge von knapp 20 Wörtern im Deutschen entstehen somit bei einem *Moving Window* mit $w = 5$ 16-mal mehr *Re-use Units*. In [Büchler 2012c] werden konkret aus 452.138 *Sentences* der *Perseus Digital Library* 8.015.511 *Re-use Units* aus einem *Moving Window* der Länge $w = 5$ generiert, was knapp der 18-fachen Menge an *Re-use Units* nur durch eine überlappende gegenüber einer disjunkten *Segmentation* entspricht. Während des *Linking* mit einer quadratischen Komplexität $O(n^2)$ bedeutet das bezogen auf die *Perseus Digital Library* etwa die 280-fache Laufzeit, die nur mit einer anderen *Segmentation* begründet werden kann.

Ähnlich der *Segmentation* hat auch das *Preprocessing* einen großen Einfluss. Jede Form eines *Preprocessing* verändert je nach den Eigenschaften einer Sprache, wie Flexion oder anderen sprachlichen Variationen z. B. Dialekte, sprachevolutionäre Veränderungen oder editorspezifische Modifikationen, das *Type-Token-Ratio* $R_{TTR} = \frac{t}{T}$ mit t *Word Types* und T *Word Tokens*. Die in Abschnitt 3.3 beschriebenen *Preprocessing*-Techniken verändern die Anzahl der *Tokens* T in R_{TTR} nicht. Durch ein *Preprocessing* wird im Kontext des *Type-Token-Ratio* R_{TTR} durch die in Abschnitt 3.3 genannten *Äquivalenzklassen* die Anzahl der *Types* t sukzessive reduziert. Im Umkehrschluss bedeutet das jedoch auch, dass sich die durchschnittliche Worthäufigkeit mit $\frac{1}{R_{TTR}}$ durch das *Preprocessing* erhöht. In [Büchler 2010d] wird auf ein längeres Platon-Zitat eingegangen, welches μῆτραί (dt. Gebärmutter, 3. Fall, weiblich) enthält. Im gesamten *Thesaurus Linguae Graecae* kann diese Wortform exakt siebenmal beobachtet werden, wobei jedes einzelne Vorkommen dieses Wortes Teil eines *Text Re-use* ist. Durch eine Lemmatisierung kann jedoch fast 1000-mal häufiger die Grundform μήτρα beobachtet werden, wovon nur ein sehr geringer Teil mit *Text Re-use* assoziiert werden kann. Bei einer *Text Re-use Analysis* mit einem *Word*-basierten *Featuring* (vgl. Abschnitt 3.4) vergrößert sich u. a. durch Lemmatisieren die Laufzeit dementsprechend deutlich und signifikant.

Generell muss bei einer *Text Re-use Analysis* deshalb immer ein Kompromiss aus *Vollständigkeit* und *akzeptabler Laufzeit* gefunden werden. Durch vorangestellte und nicht vollständige, sondern nur ausgewählte Probleme, kann sich die Laufzeit bei kleineren Änderungen auf den Level *Segmentation* und *Preprocessing* in der Zeiteinheit schnell von Tagen oder wenigen Wochen auf mehrere Monate oder gar Jahre erhöhen. Aus diesem Grund wird Vollständigkeit im Rahmen dieser Arbeit wie in Definition 18 festgelegt.

Definition 18 (Technical Completeness)**.** *Sei eine Text Re-use Analyse ϕ_Θ gegeben. Das Ergebnis einer Text Re-use Analyse ϕ_Θ wird zu einem Parameterraum Θ als technisch vollständig bezeichnet.*

Diese Definition bzw. Einschränkung ist insofern wichtig, als dass es niemals eine *Text Re-use Analysis* bzw. keine *Hybrid Text Re-use Analysis* geben wird, die jeglichen *Text Re-use* bestimmen wird. Dies sei an dem bereits erwähnten Beispiel für einen *Cognitive Re-use* zwischen *like will to like* und *Birds of same feather flock together* verdeutlicht. Es ist nahezu unmöglich, die Assoziation zwischen beiden *Re-use Units* automatisch aufzudecken.

Einhergehend mit den Komplexitätsbetrachtungen sei eine abschließende Sichtweise auf die *Text Re-use Analysis* gegeben. Eine *Brute-Force*-Technik des *Text Re-use* wäre, jede *Re-use Unit* s_i mit jeder anderen *Re-use Unit* s_j zu vergleichen. Bei n *Re-use Units* ergeben sich daraus $n \cdot (n-1)$ asymmetrische bzw. $\frac{n \cdot (n-1)}{2}$ symmetrische Vergleiche. Bei der in diesem Abschnitt vorgestellten Architektur wird während des *Linking* durch $S \cdot S^\intercal$ jeder *Re-use Unit* mit jeder anderen *Re-use Unit* verlinkt, die mindestens ein *Feature* gemeinsam haben. Bedingt dadurch, dass *Text Re-use* nicht so definiert ist, dass zwischen zwei *Re-use Units* s_i und s_j ein *Re-use Overlap* (vgl. Formel 3.14) von nur einem *Feature* besteht, sondern dass ein möglichst signifikant großer *Re-use Overlap* bestimmt werden kann, werden für jedes *Feature* gemäß dem inversen *Typ-Token-Ratio* im Durchschnitt pro *Feature* $\frac{1}{R_{TTR}} - 1$ *Links* generiert. Je nach Länge der *Re-use Unit* können hierdurch teilweise mehrere tausend oder gar zehntausende Links pro *Re-use Unit* generiert werden. Sowohl bei sehr langen *Re-use Units* als auch sprachlich stark normalisierten bzw. wenig variantenreichen Texten einer *Digital Library* mit einem hohen Wert für $\frac{1}{R_{TTR}}$ sind somit *Linking*-Kosten nahe der *Brute Force*-Methode oder gar noch laufzeitorientiert schlechtere *Text Re-use Analysis* möglich. Insbesondere das *Preprocessing* ist in diesen Laufzeitbetrachtungen sehr gegensätzlich. Um dennoch aus einem erforderlichen *Preprocessing* Vorteil für die *Text Re-use Analysis* ziehen zu können, ist eine gute *Selection* nötig. Um *Text Re-use* zu bestimmen, ist es nicht erforderlich, dass alle möglichen überlappenden *Features* bestimmt werden. Wenn eine *Re-use Unit* auf 3 bis 7 *Features* reduziert werden kann und diese zeitgleich beschreibend für den *Text Re-use* sind (vgl. *Minutiae* aus Def. 16 auf Seite 112), so ist das für die meisten *Text Re-use Analysis* oftmals bereits ausreichend und stellt einen guten Kompromiss aus Qualität der Ergebnisse sowie einem akzeptablen Laufzeitverhalten dar.

Ferner kann der im vorigen Absatz erwähnte *Re-use Overlap* als ein Maß für die *Redundanz* verstanden werden. Je mehr Redundanzen in einer *Digital Library* enthalten sind bzw. durch eine geeignete *Text Re-use Analysis* aufgezeigt werden können, desto stärker kann eine *Digital Library* bspw. durch den LZ77-Algorithmus (vgl. [Ziv 1977]) komprimiert werden.

3.10 Text Re-use Compression

Texte zu komprimieren hat den Ursprung in der Notwendigkeit, die exponentiell wachsende Datenmenge möglichst effizient und platzsparend zu speichern (vgl. [Salomon 2002]). In der Automatischen Sprachverarbeitung haben sich hierzu verschiedene Techniken, wie der *Huffman Code* (vgl. [Huffman 1952]) oder auch die *Lempel-Ziv*-Komprimierung (vgl. [Ziv 1977]), etabliert.

Während die Grundidee der Komprimierung von Texten im effizienten und platzsparenden Speichern von Dokumenten liegt, kann sie im Kontext einer *Text Re-use Analysis* als ein Evaluierungsmaß verstanden werden, um die Effektivität eines *Preprocessing*-Schrittes oder einer *Featuring*-Methode zu bestimmen. Hierzu wird aus einer *Text Re-use Analysis*

ϕ_Θ eine Methode durch eine andere ausgetauscht, um letztlich den Unterschied der *Text Re-use Compression* bzgl. beider Methoden zu vergleichen. Um diese Vergleichbarkeit auch für Techniken des *Selection* herstellen zu können, wurde in Abschnitt 3.5 die *Feature Density* $\mathcal{F}$ eingeführt.

Das *Lempel-Ziv*-Komprimierungsverfahren zeichnet sich dadurch aus, dass es einfach zu implementieren und zeitgleich schnell zu berechnen ist. Diese Komprimierungstechnik zählt zu den *Dictionary*-basierten Komprimierungsverfahren. Hierbei wird ein leeres Wörterbuch initialisiert. Ist das Zeichen oder das Wort nicht im Wörterbuch enthalten, so wird es unter einer ID im Wörterbuch aufgenommen. Ist ein Zeichen oder Wort im Wörterbuch enthalten, so wird sukzessive ein weiterer Buchstabe bzw. ein weiteres Wort dem beobachteten Token solange hinzugefügt, bis das Token nicht mehr im Wörterbuch enthalten ist. Dieses wird nun schließlich ebenfalls in das Wörterbuch eingetragen. Dieser simple Wörterbuchansatz kann mit dem *Longest Common Consecutive Words*-Ansatz (vgl. [Sediyono 2008]) aus Abschnitt 3.4 verglichen werden. Trotz seiner Einfachheit hat er jedoch zwei signifikante Nachteile, um die durch *Text Re-use* verursachte Kompression zu messen. Einerseits beschränkt sich der *Lempel-Ziv*-Ansatz auf den *Syntactic Re-use* (vgl. Abschnitt 2.6), bei welchem die Reihenfolge der Wörter zweier *Re-use Units* identisch bzw. weitestgehend identisch sein muss. Sobald sich selbst bei gleichen Wörtern, wie in *Die Wahrheit liegt im Wein* und *Im Wein liegt die Wahrheit*, die Reihenfolge ändert, komprimiert der *Lempel-Ziv*-Ansatz deutlich schlechter. Andererseits ist bei dieser Technik der Einfluss von allen Wörtern im Text, insbesondere der Stoppwörter, gegeben, so dass ggf. reine Stoppwortsequenzen, die durch *Text Re-use* induzierte *Kompression* größer erscheinen lassen, als sie realistisch ist.

Aus diesem Grund wird im Rahmen dieser Arbeit bei der *Text Re-use Compression* als *Evaluierungsgrundlage* auf den *Re-use Overlap* aus Formel 3.13 zurückgegriffen. Sowohl die symmetrische *Resemblance* $\theta^R_\Theta(s_i, s_j)$ als auch das asymmetrische *Containment* $\theta^C_\Theta(s_i, s_j)$ erzeugen *Scores* mit $\theta^R_\Theta(s_i, s_j) \in [0, 1]$ und $\theta^C_\Theta(s_i, s_j) \in [0, 1]$. Ist der *Score* $\theta_\Theta(s_i, s_j) = 0$, dann liegt kein *Text Re-use* vor. Bei einem *Score* von $\theta_\Theta(s_i, s_j) = 1$ kann s_j vollständig durch s_i repräsentiert bzw. komprimiert werden. Jeder *Score* zwischen 0 und 1 impliziert eine anteilige Komprimierung. Die *Text Re-use Compression* zu einem *Text Re-use* ϕ_Θ mit den Parameterraum Θ ergibt sich somit wie in Formel 3.21 abgebildet.

$$\mathcal{C}_\Theta = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} \theta_\Theta(s_i, s_j)}{n \cdot m} \tag{3.21}$$

Wenn für jedes Element $\theta_\Theta(s_i, s_j) = 0$ gilt, also kein *Text Re-use* zu einem Parameterraum Θ erkannt werden konnte, dann ist dementsprechend auch die *Text Re-use Compression* $\mathcal{C}_\Theta = 0$. Es sei nun bei einem *Intra Linking Text Re-use* mit $n = m$ angenommen, dass jede *Re-use Unit* s_i eine exakte Kopie der ersten *Re-use Unit* s_1 ist. Dann gilt $\theta_\Theta(s_i, s_j) = 1$ für T^R mit $i \neq j$. Für die entsprechende obere Abschätzung des *Text Re-use Compression* ergibt sich somit die Schranke wie in Formel 3.22.

$$\mathcal{C}_\Theta = \frac{n \cdot (n-1)}{n^2} = 1 - \frac{1}{n} \tag{3.22}$$

Die Formel 3.22 zeigt klar, dass bei einer Duplizierung einer *Re-use Unit* die *Text Re-use Compression* mit steigendem Grad der Duplizierung gegen 1 konvergiert, jedoch 1 niemals annehmen kann. Im Beispiel zu diesem Kapitel aus Formel 3.8 kann aus der *Resemblance*-basierten *Scoring*-Matrix eine *Text Re-use Compression* von 0.3064 und bei einer durch 5 exakte Duplikate maximal möglichen *Text Re-use Compression* von 0.8 bestimmt werden.

Da in der praktischen Anwendung der *Text Re-use Compression* die *Scoring*-Matrix T meistens eine *Sparse Matrix* ist, werden oft sehr kleine $\mathcal{C}_\Theta$ berechnet. Da das Maß der *Text Re-use Compression* $\mathcal{C}_\Theta$ nicht zum Ziel hat, den benötigten Speicherplatz zu reduzieren,

sondern als Evaluierungsmaß bzw. Vergleich zweier bspw. *Selection*-Techniken dienen soll, ist es bereits ausreichend, zwei *Text Re-use Compressions* miteinander zu vergleichen. In diesem Sinne wird das zwei Techniken vergleichende *Compression Ratio* $\mathcal{R}$ definiert als

$$\mathcal{R} = \frac{\mathcal{C}_{\Theta'}}{\mathcal{C}_{\Theta}}. \tag{3.23}$$

Gilt hierbei $\mathcal{R} > 1$ oder gar $\mathcal{R} \gg 1$, dann stellt die modifizierte Technik hinter $\mathcal{C}_{\Theta'}$ eine Verbesserung zum Ausgangsszenario $\mathcal{C}_{\Theta}$ dar.

Eine dünnbesetzte *Scoring*-Matrix T kann auch durch eine Aggregation verschiedener *Re-use Units* bspw. zu Werken bzw. Autoren aber auch literarischen Klassifikationen oder Datierungen zusammengefaltet werden. Hierzu wird eine (l, n)-*Meta Data*-Matrix M mit der (n, m)-*Scoring*-Matrix T von links bzw. mit einer entsprechend transponierten (m, l)-*Meta Data*-Matrix $M^{\intercal}$ von rechts, wie in Formel 3.24 mit $l \ll n$ und $l \ll m$, multipliziert.

$$T' = M \cdot T \cdot M^{\intercal} \tag{3.24}$$

In Verbindung mit dem zuvor eingeführten *Compression Ratio* $\mathcal{R}$ kann das auf Metadaten basierende Zusammenfalten der *Scoring*-Matrix gut eingesetzt werden, um Textsorten-abhängige Optimierungen am *Text Re-use Model* vorzunehmen.

Das *Text Re-use Compression* kann von seiner Anlage her als ein spezieller *Postprocessing*-Schritt verstanden werden, der das Ergebnis einer *Text Re-use Analysis* in einem Wert zusammenfasst. Für die *Text Re-use Analysis*, insbesondere im Hinblick auf die sprachliche Vielfalt, des *Historical Text Re-use* mit den in Abschnitt 2.6 genannten *Meme* und *Re-use Styles* kann die *Text Re-use Compression* bzw. das *Compression Ratio* als ein quantitatives Evaluierungsmaß verstanden werden.

Kapitel 4

Zufall und Struktur

Contents

4.1 Einführung 128
4.2 Probleme von Sprachmodellen 129
4.3 Evaluierung von Sprachmodellen 137
4.4 Noisy Channel Evaluation 141
4.5 Arten einer *Randomised Digital Library* 149
4.6 Eigenschaften der *Noisy Channel Evaluation* 151
4.6.1 *Mining Ability* in Abhängigkeit von der Größe einer *Digital Library* bei konstantem Parameterraum Θ 154
4.6.2 *Mining Ability* in Abhängigkeit vom Parameterraum Θ bei konstanter Größe einer *Digital Library* 158
4.6.3 Minimale und maximale *Mining Ability* $\mathcal{L}_{min}(\Theta)$ und $\mathcal{L}_{max}(\Theta)$ bei einem dynamischen Parameterraum Θ sowie unterschiedlich großen *Digital Libraries* 161
4.7 Einbettung dieses Kapitels in die gesamte Arbeit 163

Language is never ever random.

Adam Killgariff, (1960-)

Strukturen, wie *Morphologie*, *Syntax*, oder *Semantik*, geben einem Text seine Natürlichsprachlichkeit und unterscheiden ihn von einer zufälligen Anordnung der Wörter bzw. der Buchstaben. Im Kontext des *Text Re-use* würde einer *Digital Library* etwas fehlen, wenn nicht wenigstens entsprechende Phrasen und Redewendungen Teil der Texte wären. Daher kann *Text Mining* als die Summe aller Ansätze verstanden werden, welche entsprechend der gestellten Aufgabe natürlichsprachliche Strukturen aus einer *Digital Library* extrahieren. Für die *Text Re-use Analysis* bedeutet dies, dass einerseits relevanter *Text Re-use* extrahiert werden und auf der anderen Seite das zufälliges Rauschen möglichst klein bleiben soll.

In diesem Kapitel wird das rein quantitative Maß der *Mining Ability* eingeführt, welche die Fähigkeit einer *Mining*-Methode, wie es auch der *Text Re-use* ist, misst, um zwischen Struktur und Zufall unterscheiden zu können.

Einige grundlegende Eigenschaften der quantitativen Evaluierung mit der *Mining Ability* werden in diesem Kapitel erklärt. Da für den *Historical Text Re-use* mit der *7-Level-Architektur* ein komplexes Modell zugrunde liegt, wird der *Text Re-use* in diesem Abschnitt auf die beiden *Atome Bigram* und *Co-occurrence* reduziert, welche den kleinstmöglichen mehrgliedrigen *Text Re-use* entsprechen. Ferner dient diese Vereinfachung der Reduktion der Komplexität, so dass in diesem Kapitel letztlich nur die beiden Abhängigen der Größe der *Digital Library* sowie dem Schwellwert eines Signifikanzmaßes untersucht werden.

4.1 Einführung

Wörter werden nicht zufällig gebildet bzw. stellen nicht nur eine Aneinanderreihung von Buchstaben dar. Wörter wiederum werden nicht regellos in eine Sequenz von Wörtern, wie einem Satz, gebracht, sondern genügen einer Sprachsyntax. Die Wörter innerhalb eines Satzes haben untereinander semantische Bezüge und werden daher nicht kontextfrei benutzt. Sätze dagegen folgen einem argumentativen Diskurs.

All diese verschiedenen Aspekte unterscheiden Sprache deutlich von einer reinen Zufallssequenz von Buchstaben. Diese von Zufall abweichenden Strukturen aufzudecken, ist Gegenstand der Automatischen Sprachverarbeitung bzw. im Falle des *Text Re-use* des *Text Re-use Mining*. Entsprechende Verfahren können gemäß ihrer Methodik bzw. des Vorwissens klassifiziert werden. Ein Beispiel für ein regelbasiertes Verfahren ist *Morpheus* (vgl. [Crane 1991]), welches eine oder mehrere Grundformen zu einer gegebenen Wortform bestimmt. Eine andere Klasse wird durch die probabilistischen Sprachmodelle (vgl. [Manning 1999, Heyer 2006]) repräsentiert, welche sich der Sprachstatistik oder auch den Hypothesentests aus der Schätz- und Testtheorie bedienen.

Die grundlegende Fragestellung eines jeden *Text Mining*-Verfahren ist, wie das entsprechende Sprachmodell und dessen Ergebnisse evaluiert werden können. Hierzu wird in den meisten Fällen ein *Gold Standard* definiert, gegen die das Ergebnis eines *Mining*-Verfahrens evaluiert wird. Kenngrößen, wie *Precision*, *Recall* und *F-Measure* (vgl. Abschnitt 4.2), haben sich in den letzten Jahrzehnten für die Evaluierung etabliert.

Im Kontext des *Historical Text Re-use Detection* und der in Abschnitt 2.6 dargelegten *Re-use Diversity* würde das jedoch bedeuten, dass für jedes *Meme* kombiniert mit jedem *Re-use Style* ein adäquater *Evaluationsstandard* generiert werden müsste. Das ist aber aus Zeit- und Arbeitsaufwandsgründen nicht umsetzbar. Insofern resultiert aus der *Diversity* das grundlegende Problem der fehlenden Evaluierungsmöglichkeit für den *Historical Text Re-use*.

Soll nun die Qualität eines Verfahrens zum Aufdecken einer solchen im Text enthaltenen Struktur aufgedeckt werden, gib es zwei Möglichkeiten, sich diesem Problem zu nähern. Einerseits kann jedes einzelne kleine Teilergebnis auf seine Richtigkeit überprüft werden. Dies ist bspw. bei einem statistischen Hypothesentest der Fall, bei welchem abschließend die Gesamtfehlerwahrscheinlichkeit für die *Digital Library* bestimmt wird. Andererseits kann eine Evaluierung gegen einen *Gold Standard* durchgeführt werden, um ein Gesamtergebnis betrachten zu können. Beide Techniken sind wohl etabliert, haben aber zeitgleich auch signifikante Nachteile. In den folgenden Abschnitten 4.2 und 4.3 wird auf beide Ansätze eingegangen.

In diesem Kapitel wird neben der in Abschnitt 3.10 vorgestellten *Text Re-use Compression*, welche sich lediglich für die Evaluierung von *Text Re-use* empfiehlt, eine weitere Evaluierungsmethode eingeführt. Die Methode *Noisy Channel Evaluation* kann auch in der Breite für andere Bereiche des *Text Mining* adaptiert werden. Ziel dieser rein quantitativen Evaluierung ist es, die Strukturen in einem Text im Vergleich zu einer willkürlichen Sequenz von Buchstaben oder Wörtern zu testen. Hierbei wird eine ausgewählte Eigenschaft einer *Digital Library* (vgl. Abschnitt 4.5) systematisch in einer *Randomised Digital Library* ersetzt. Forschungsgegenstand ist, die Ergebnisse zu einem gegebenen Parameterraum nicht nur zu vergleichen, wenn es faktisch keine enthaltenen Strukturen gibt, sondern einen quantitativen *Score*, die *Mining Ability*, zu definieren (vgl. Abschnitt 4.4). *Text* als auch *Text Re-use Mining* kann daher als die Fähigkeit verstanden werden, nicht zufällige Strukturen aus einer *Digital Library* aufzudecken, die durch die *Mining Ability* quantifiziert wird. Diese Form der quantitativen Evaluierung ist, wie in Abschnitt 1.7 bereits aufgezeigt, in Shannon's *Noisy Channel Theorem* (vgl. [Shannon 1948]) eingebettet, so dass

die Methode nachfolgend dementsprechend *Noisy Channel Evaluation* genannt wird. Diese Evaluierungsmethode zeichnet sich dadurch aus, dass immer die Qualität durch messbare Evaluierungskriterien der gesamten *Digital Library* und nicht nur die Qualität bzgl. eines *Gold Standards* gemessen wird, die bei einer zweiten Evaluierungsgrundlage wiederum völlig anders ausfallen kann.

4.2 Probleme von Sprachmodellen

Text Mining benötigt ein Sprachmodell. Da es oftmals schwer ist, das nötige und nicht selten auch sehr umfangreiche Wissen einem Computer anzulernen, so dass er anschließend entsprechend nicht zufällige und absichtliche Strukturen aus dem Text aufdecken kann, werden in vielen Bereichen *probabilistische Sprachmodelle* eingesetzt. Der Vorteil dieser Klasse von Sprachmodellen liegt darin, dass statistisch auffällige Zusammenhänge aus Texten erlernt und diese als Wissensbasis wiederverwendet werden können. Im Kontext des *Text Re-use* wäre es bspw. mathematisch möglich, die statistische Fehlerwahrscheinlichkeit aus einem *Re-use Overlap* aus den Wortwahrscheinlichkeiten zu bestimmen. Die Fehlerwahrscheinlichkeit kann durch einen χ^2-Test oder einem sich asymptotisch zum χ^2-Test verhaltenden Maß berechnet werden. Wie in den Tabellen A.1 bis A.10 aufgezeigt, kann aus einem solchen *Score* unter Berücksichtigung der Freiheitsgrade der statistische Fehler bestimmt werden. Es kann dementsprechend als Grundidee verstanden werden, die Ergebnisse einer *Mining*-Analyse einzeln auf ihre Fehlerwahrscheinlichkeit zu untersuchen und daraus den statistischen Gesamtfehler innerhalb einer gesamten *Digital Library* zu berechnen.

Nach einer kurzen Einführung gliedert sich dieser Abschnitt in zwei Teile. Zunächst wird *Text Re-use* auf die beiden kleinstmöglichen, mehrgliedrigen *Atome*, dem *Bigram* sowie der *Co-occurrence*, reduziert. Anschließend werden auch größere *Re-use Overlaps* aus statistischer Sicht betrachtet.

Grundlage für solche Hypothesentests ist in den meisten Fällen die stochastische Unabhängigkeit aus Formel 4.1 mit w_i als Wörtern und $p(w_i)$ der entsprechenden Wahrscheinlichkeit des Wortes w_i, welche sowohl eine Alternativ- als auch eine Nullhypothese benötigen. Während die Alternativhypothese der realen Beobachtung entspricht, wird das zufällige gemeinsame Auftreten oftmals als Nullhypothese eingesetzt. Dabei ist es meist eher schwierig, eine qualifiziertere und gesichert bestimmbare Nullhypothese zu formulieren.

$$p(w_1 w_2 \cdots w_n) = p(w_1)p(w_2)\cdots p(w_n) \tag{4.1}$$

Während die linke Seite der Formel 4.1 zur stochastischen Unabhängigkeit die beobachtete Wahrscheinlichkeit eines *Re-use Overlaps* reflektiert, entspricht die rechte Seite der statistischen und zufälligen Erwartung. Durch Umstellen und Ziehen des Logarithmus folgt Formel 4.2.

$$0 = \log\left(\frac{p(w_1 w_2 \cdots w_n)}{p(w_1)p(w_2)\cdots p(w_n)}\right) \tag{4.2}$$

Formel 4.2 macht deutlich, dass, wenn sowohl die Beobachtung als auch die Erwartung gleich sind, dementsprechend die Gleichung gilt. Interessanter jedoch sind die Fälle, die sich möglichst stark von 0 unterscheiden, um bspw. eine statistische Über- bzw. Unterbenutzung zu messen (vgl. [Hirschmann 2012]). Church & Hanks formulierten 1989 hierzu das einfachste aller statistischen Maße, die *pointwise Mutual Information*, aus Formel 4.3 (vgl. [Church 1989]).

$$sig_{MI} = \log\left(\frac{p(w_1 w_2 \cdots w_n)}{p(w_1)p(w_2)\cdots p(w_n)}\right) \tag{4.3}$$

Die *Mutual Information* wird vielseitig eingesetzt. In der Bioinformatik wird sie benutzt, um eine statistische Aussage über die Qualität des *Sequence Alignments*[1] auszudrücken (vgl. [Penner 2011]). Weiterhin gibt es auch Anwendungen in der Bioinformatik, in welcher die *Mutual Information* zum Erstellen von Graphen eingesetzt wird (vgl. [Hsu 2012]). Auch in der Automatischen Sprachverarbeitung hat sie seit Church & Hanks ihre Anwendung in der automatischen Generierung von Graphen gefunden (vgl. u. a. [Baroni 2004]).

Das Hauptproblem bei der Anwendung der stochastischen Unabhängigkeit in Form der *Mutual Information* ist jedoch durch die *Power Law*-Verteilung der Wörter[2] gegeben, wodurch sehr viele Wörter nur selten in einer *Digital Library* beobachtet werden können.

Hierzu sei zunächst die stochastische Unabhängigkeit aus Formel 4.1 zu einem *Re-use Overlap* aus nur noch zwei Wörtern, bspw. eines syntaktischen *Bigram* oder einer semantischen *Co-occurrence*, mit $p(w_i) = \frac{|w_i|}{N}$ und $p(w_i w_j) = \frac{|w_i w_j|}{M}$ vereinfacht, wobei N der Anzahl der *Tokens* bzw. der fortlaufenden Wörter und M der Summe aller *Bigram*- bzw. *Co-occurrence*-Häufigkeiten entspricht.

$$\frac{|w_i w_j|}{M} = \frac{|w_i||w_j|}{N^2} \tag{4.4}$$

Um die durch die *Power Law*-Verteilungen verursachten statistischen Probleme der stochastischen Unabhängigkeit zu verdeutlichen, sei aus der vorigen Formel die *erwartete Frequenz* in Formel 4.5 aufgestellt, wobei $\frac{M}{N^2}$ für eine *Digital Library* als eine fixe Normalisierungskonstante angesehen werden kann.

$$f_{epx} = |w_i||w_j|\frac{M}{N^2} \tag{4.5}$$

Die statistisch erwartbare Frequenz f_{exp} aus Formel 4.5 kann als diejenige *Bigram*- oder *Co-occurrence*-Häufigkeit verstanden werden, von welcher sich eine beobachtete Wortassoziation durch $f_{obs} = |w_i w_j|$ möglichst deutlich unterscheiden sollte. Es ist nun ausgehend von dem Ergebnis einer *Bigram*- oder *Co-occurrence*-Analyse einfach, danach zu fragen, für wie viele der beobachteten *Bigrams*- oder *Co-occurrences* gilt, dass nicht nur $f_{obs} \geq 1$ ist, sondern auch $f_{epx} \geq 1$. Die Grundidee dieses Experimentes besteht darin, dass, wenn $0 < f_{epx} \ll 1$ gilt, genau dann eine nicht-natürliche, sondern aus einem statistischen Problem heraus resultierende Abweichung zwischen der beobachteten Wahrscheinlichkeit $p(w_i w_j)$ und der erwarteten Wahrscheinlichkeit $p(w_i)p(w_j)$ resultiert. Es sei bspw. f_{epx} mit $f_{epx} = 10^{-5}$ angenommen. Es wird weiterhin ein *Bigram* bzw. eine *Co-occurrence* nur einmal beobachtet, dann würde die Beobachtung 10^5-mal wahrscheinlicher sein als die statistische Erwartung, was wiederum zwangsläufig einer statistisch positiven Auffälligkeit entspricht, auch wenn es sich in den meisten Fällen nur um Rauschen handelt. Für das eben genannte Beispiel würde sich ein χ^2-Wert von $\chi^2 = 99998$ berechnen, was einem α-Signifikanzniveau bzw. einer vermeintlichen statistischen Fehlerwahrscheinlichkeit $P_{error} \ll 10^{-4}$ entspricht. Angesichts dessen, dass selbst bei großen *Digital Libraries* für die Wortwahrscheinlichkeit $p(w_i)$ in den meisten Fällen $p(w_i) \leq 10^{-7}$ gilt, tritt genau dieser Fall häufig auf.

Dieser statistische Nebeneffekt soll nachfolgend in einer Reihe von Experimenten analysiert werden, welche auf unterschiedlich großen Textmengen, den sogenannten Normkorpora (vgl. [Biemann 2007a]), aus der *Leipzig Corpora Collection* (vgl. [Goldhahn 2012]) durchgeführt werden. Es wird ein 250 Millionen Sätze umfassendes, aus dem Web gesammeltes, deutschsprachiges Korpus zusätzlich auf die Normgrößen[3] *1k*, *3k*, *10k*, *30k*, *100k*, *300k*,

[1] *Sequence Alignment* kann als das Pendant der Bioinformatik zum *Historical Text Re-use* der *eHumanities* verstanden werden.

[2] Im Falle der Wortverteilung wird das *Power Law* auch *Zipfsches Gesetz* genannt (vgl. [Zipf 1949]).

[3] k ist die Abkürzung für 1000. M entspricht einer Millionen.

1M, *3M*, *10M*, *30M* sowie *100M* Sätze mit einer durchschnittlichen Satzlänge von etwa 18 normalisiert[4]. Die unterschiedlichen Normgrößen unterstützen hierbei, das Verhalten bei wachsender Datenmenge zu verstehen.

In einem ersten Experiment werden all diejenigen *Bigrams* der unterschiedlichen Normkorpora ausgewählt, für welche die Bedingung $f_{epx} \geq 1$ gilt. In Abb. 4.1 wird der Verlauf jener ausgewählter *Bigrams* bzgl. unterschiedlich großer Normkorpora dargestellt. Sowohl die x- als auch die y-Achse sind bedingt durch die *Power Law*-Verteilung der Wörter logarithmisch skaliert. Beide Achsen repräsentieren nach dem Zipfschen Gesetz Rang-sortierte Wortlisten[5]. Die x- und y-Achsen aller zwölf Plots sind auf 36 Millionen Wörter normiert, die im Maximum beim Ausgangs- und somit größtmöglichen Korpus beobachtet werden konnten. Ein untersuchtes *Bigram* wird genau dann schwarz eingezeichnet, wenn es die minimalste Testbedingung $f_{epx} \geq 1$ erfüllt. Die weiße Fläche entspricht denjenigen *Bigrams*, die entweder nicht die Testbedingung erfüllt haben oder nicht im Normkorpus aufgetreten sind. Bereich *Data Sparseness* herrscht, während im schwarzen Bereich eine vergleichbar hohe Datendichte vorliegt.

Die verschiedenen Plots in Abb. 4.1 zeigen deutlich auf, dass die Menge der *Bigrams*, die die Testbedingung $f_{epx} \geq 1$ erfüllen, sukzessive (schwarzer Fläche) bei zunehmender Größe des Normkorpus steigt. Daraus kann direkt abgeleitet werden, dass sich dementsprechend die Sprachstatistik verbessert.

	1	3	5	10	20	50	75	100
1M	0.2629	0.1044	0.0657	0.0344	0.0178	0.0073	0.0049	0.0037
30M	0.3343	0.1556	0.1056	0.0611	0.0345	0.0157	0.0110	0.0085
100M	0.3462	0.1636	0.1127	0.0668	0.0390	0.0185	0.0132	0.0103

Tabelle 4.1: Diese Tabelle stellt die Ergebnisse von Experimenten mit einer Normgröße von *1M*, *30M* und *100M* Sätzen dar. Untersucht werden die Anzahl derjenigen *Bigrams*, die die Testbedingung $f_{epx} \geq i$ mit $i \in \{1, 3, 5, 10, 20, 50, 75, 100\}$ erfüllen.

Es muss jedoch auch festgehalten werden, dass $f_{epx} \geq 1$ lediglich eine Minimalbedingung ist. Auf einem Normkorpus mit 100 Millionen Sätzen[6] erfüllen die minimalste Testbedingung lediglich 0.3462 aller möglichen *Bigrams* (vgl. Spalte *1* aus Tabelle 4.1). Für fast zwei Drittel aller beobachteten Daten muss festgestellt werden, dass $f_{epx} < 1$ gilt. Zum Verdeutlichen dieses Ergebnisses sei es mit den in der Wahrscheinlichkeitstheorie üblichen Würfelexperimenten verglichen. Für die gleiche Analyse mit zwei regulären Würfeln, bei der insgesamt 36 Kombinationen möglich sind, reichen bereits 200 Versuche, um einen solchen Plot vollständig bzw. nahezu vollständig schwarz einzufärben. Selbst eine Analyse mit einem hypothetischen regulären Würfel mit 10^5 Seiten gleicher Wahrscheinlichkeit $\frac{1}{n}$ bei vergleichbarer Anzahl von Versuchen, wie es *Bigrams* im *100M*-Normkorpus gibt, würde dieser Plot nahezu komplett schwarz eingefärbt sein. Es kann also kurzum gesagt werden, dass selbst bei großen Textmenge für zwei Drittel aller *Bigrams* und *Co-occurrences* das einmalige Auftreten dieser Wortkombination bereits über der statistischen Erwartung liegt.

[4]Die Wahl von Normgrößen wie *3k* oder *30k* kann durch oftmals logarithmische Skalierung der Achsen von Ergebnissen, wie in Abb. 4.1, begründet werden. Während Normgrößen, wie *100k* oder *1M*, in einer logarithmischen Darstellung durch den dekadischen Logarithmus immer genau einer Darstellungseinheit entsprechen, wie es durch $log_{10}100k = 6$ oder $log_{10}1M = 7$ gegeben ist, genügt der dreifache Werte mit $log_{10}3 = 0.477$ in etwa der Hälfte zwischen zwei Darstellungseinheiten.

[5]Im Detail bedeutet dies, dass ein Wort umso häufiger beobachtet werden kann, je näher es sich am Koordinatenursprung befindet.

[6]Das entspricht in etwa 1.8 Milliarden fortlaufenden Wörtern bzw. *Tokens*.

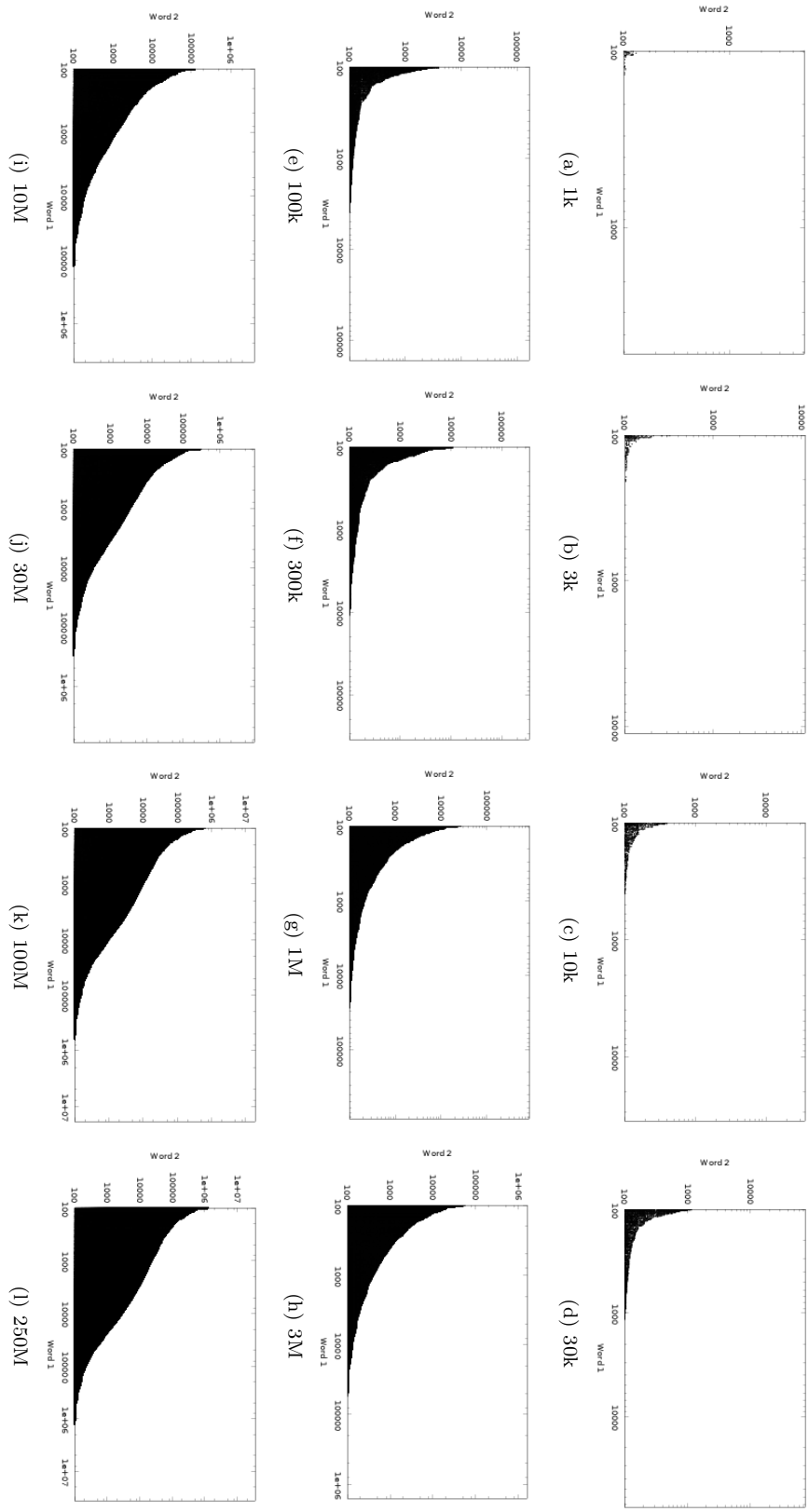
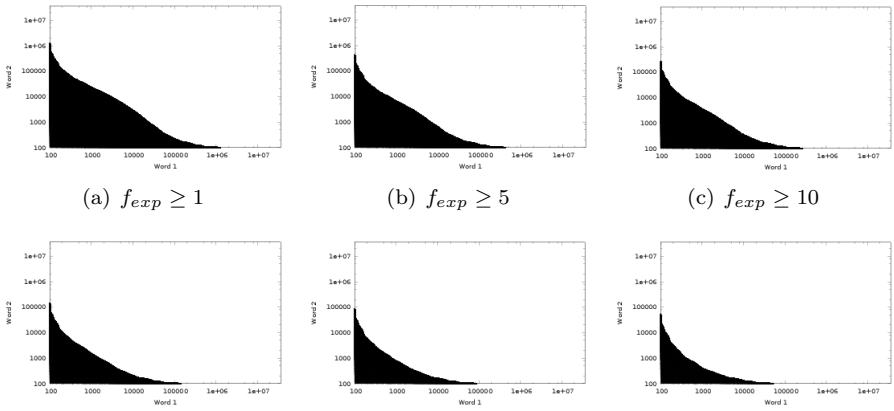


Abbildung 4.1: Über 90% aller in einer *Digital Library* enthaltenen Wörter werden zehnmal oder seltener beobachtet. Bedingt durch die zugrunde liegende *Power Law*-Verteilung werden bei einer Wortassoziationsanalysen statistische Probleme induziert. Den Abbildungen 4.1(a) bis 4.1(l) liegt eine *Bigram*-Analyse zugrunde, wobei die Wörter auf den beiden Achsen mit logarithmischer Skalierung abgetragen werden. Jedes beobachtete *Bigram* wird nach Formel 4.5 mit $f_{eps} \geq 1$ analysiert. Die schwarze Fläche entspricht den *Bigrams*, welche den Test bestanden haben. Die weiß markierten Flächen repräsentieren die *Bigrams*, welche den Test nicht bestanden haben oder nicht beobachtet worden sind.

Das stellt wiederum den mathematischen Test und insbesondere deren Anwendung im Bereich der Automatischen Sprachverarbeitung in Frage. In einer *Text Re-use*-Anwendung würde das im Detail bedeuten, dass viele binäre *Re-use Overlaps* statistisch vielleicht auffällig oder gar zu einem α-Fehlerniveau signifikant sind, jedoch hat ein solcher teilweise sehr hoher *Score* keinerlei statistische Relevanz. Sie können daher auch als "*Phantomsignifikanzen*" bezeichnet werden, die aufgrund der Bedingung $f_{epx} \ll 1$ meistens sogar die höchsten *Scores* in einer *Digital Library* erzielen. Gemäß dem *Gesetz der großen Zahlen*, welches einem statistischen Test zugrunde liegen sollte, muss die Minimalbedingung deutlich erhöht werden, um eine haltbare statistische Aussage machen zu können. Abb. 4.2 zeigt den Verlauf für das *250M* Korpus unter der Berücksichtigung einiger ausgewählter f_{exp} mit einer Erwartung von bis zu $f_{exp} \geq 100$. Zusätzlich kann aus Tabelle 4.1 (vgl. Spalte *100*) abgelesen werden, dass in diesem Fall einen solcher Test nur noch 0.0103 aller *Bigrams* bestehen. Auch wenn es zweifelsohne keine feste Definition der Größe bzw. des Umfanges des *Gesetzes der großen Zahlen* gibt, da es sich lediglich um ein statistisches Verhalten handelt, so scheint ein $f_{exp} \geq 100$ nicht zu restriktiv bzw. zu hoch angesetzt zu sein. Selbst bei einem $f_{exp} \geq 20$ oder $f_{exp} \geq 50$ erfüllen diese Testbedingung nur noch 0.0668 bzw. 0.039 aller *Bigrams*, so dass gesichert festgehalten werden kann, dass deutlich mehr als $\frac{9}{10}$ *Bigrams* einer "*Phantomsignifikanz* "unterliegen[7].



(a) $f_{exp} \geq 1$ (b) $f_{exp} \geq 5$ (c) $f_{exp} \geq 10$

(d) $f_{exp} \geq 25$ (e) $f_{exp} \geq 50$ (f) $f_{exp} \geq 100$

Abbildung 4.2: Ausgehend von einer *Digital Library* mit *250M* Sätzen wird der Verlauf der Testbedingungen von $f_{epx} \geq 1$ bis $f_{epx} \geq 100$ abgebildet.

Zurückkommend zu dem Fall, dass der *Re-use Overlap* aus mehr als zwei Wörtern besteht, sei die Frage aufgrund der vorigen Ergebnisse, dass für etwa zwei Drittel aller *Bigrams* $f_{epx} < 1$ gilt, nicht mehr danach gestellt, wie viele *Ngrams* die minimalste Testbedingung $f_{epx} \geq 1$ erfüllen, sondern es sei verschärft danach gefragt, ab welcher Größe ein *Re-use Overlap* immer statistisch signifikant ist.

Ausgehend von der *stochastischen Unabhängigkeit* $p(w_1 w_2 \cdots w_n) = p(w_1)p(w_2)\cdots p(w_n)$ aus Formel 4.1, sei diese Formel für die gestellte Abschätzung vereinfacht. Dies ist insofern der Sache dienlich, als dass ohne Vereinfachung insgesamt n Parameter berücksichtigt werden müssten, was eine Abschätzung deutlich verkomplizieren würde. Aus diesem Grund sei $p(w_j)$ das harmonische Mittel aller Einzelwortwahrscheinlichkeiten des *Ngrams*, so dass die

[7]Diese Ergebnisse konnten sowohl für Daten einer *Co-occurrence*-Analyse als auch auf einem vergleichbar großen Englischkorpus bestätigt werden. Da die Ergebnisse dieser Analysen keinen weiteren Mehrwert enthalten, wurde an dieser Stelle darauf verzichtet, doppelte Resultate einzufügen.

Approximation aus Formel 4.6 aufgestellt werden kann.

$$p(w_1 w_2 \cdots w_n) \approx p(w_j)^n \tag{4.6}$$

Weiterhin sei als Testkriterium für $p(w_1 w_2 \cdots w_n) = \frac{\lfloor w_1 w_2 \cdots w_n \rfloor}{M}$ mit M als der Anzahl aller *Bigrams* die Bedingung des einmaligen Auftretens des *Bigrams* mit einer Wahrscheinlichkeit von $\frac{1}{M}$ definiert. Ausgehend von dieser Bedingung kann die Formel 4.6 nach n umgestellt werden. Formel 4.7 reflektiert diese Umstellung, wobei zusätzlich aufgerundet wird. Dies ist insofern sinnvoll, als dass die Frage nach der Mindestgröße des *Re-use Overlaps* gestellt worden ist, ab welchem er immer signifikant ist.

$$n = \left\lceil -\frac{\log_{10}(M)}{\log_{10}(p(w_i))} \right\rceil \tag{4.7}$$

Tabelle 4.2 reflektiert die Analysen für die Normgrößen *1k*, *10k*, *100k*, *1M*, *10M*, *100M* sowie *250M* zu vier verschiedenen harmonischen Mitteln für das häufigste Wort und den fünf, 100- sowie 1000-häufigsten Wörtern, welche gemäß dem *Zipfschen Gesetz* stark kleiner werdende Wahrscheinlichkeiten $p(w_j)$ besitzen.

Als gröbste Abschätzung sei angenommen, dass jedes Wort eines *Ngrams* mit der Wahrscheinlichkeit des häufigsten Wortes (vgl. $j = 1$ in Tabelle 4.2) ersetzt wird, dann kann bei einer Größe der *Digital Library* von 250 Millionen Sätzen und einer durchschnittlichen Satzlänge von etwa 18 für n ein Wert von $n = 7$ ausgerechnet werden (vgl. letzte Spalte in der ersten Zeile mit einem Wert von $\lceil 6.881 \rceil$). In der Interpretation bedeutet dies, dass bei der genannten Textmenge das einmalige Auftreten eines *Septogram*[8] immer statistisch auffällig oder gar signifikant ist. Wird eine weniger großzügige Abschätzung zugrunde gelegt, in dem bspw. die Wahrscheinlichkeit des harmonischen Mittels der Wortwahrscheinlichkeiten des *Ngrams* durch das 1000 häufigste Wort repräsentiert ist, so gilt die Testbedingung bereits für jedes *Trigram* als erfüllt (vgl. Tabelle 4.2).

j	$p(w_j)$	$1k$	$10k$	$100k$	$1M$	$10M$	$100M$	$250M$
1	$2.57 \cdot 10^{-2}$	$\lceil 4.259 \rceil$	$\lceil 3.895 \rceil$	$\lceil 4.552 \rceil$	$\lceil 5.428 \rceil$	$\lceil 5.925 \rceil$	$\lceil 6.667 \rceil$	$\lceil 6.881 \rceil$
5	$2.01 \cdot 10^{-3}$	$\lceil 1.994 \rceil$	$\lceil 2.338 \rceil$	$\lceil 2.731 \rceil$	$\lceil 3.075 \rceil$	$\lceil 3.587 \rceil$	$\lceil 3.872 \rceil$	$\lceil 4.055 \rceil$
100	$7.81 \cdot 10^{-4}$	$\lceil 1.789 \rceil$	$\lceil 2.123 \rceil$	$\lceil 2.438 \rceil$	$\lceil 2.706 \rceil$	$\lceil 3.069 \rceil$	$\lceil 3.361 \rceil$	$\lceil 3.519 \rceil$
1000	$7.86 \cdot 10^{-5}$	$\lceil 1.293 \rceil$	$\lceil 1.635 \rceil$	$\lceil 1.876 \rceil$	$\lceil 2.104 \rceil$	$\lceil 2.313 \rceil$	$\lceil 2.555 \rceil$	$\lceil 2.665 \rceil$

Tabelle 4.2: Diese Tabelle stellt die Ergebnisse aus Formel 4.7 dar. Hierzu wurden Experimente auf verschiedene Normgrößen, den Spalten, sowie zu vier verschiedenen Approximationen der Wahrscheinlichkeit $p(w_j)$, den Zeilen, durchgeführt.

Da in Tabelle 4.2 noch immer relativ häufige und dementsprechend wenig restriktive Abschätzungen für $p(w_j)$ gemacht worden sind, können die genannten Wahrscheinlichkeiten $log_{10}(p(w_j))$ in Formel 4.7 für große Korpora als mehr oder weniger konstant angenommen werden. Da die Tabelle 4.2 aufzeigt, dass bereits für ein *Tri-* bzw. *Quatrogram* angenommen werden kann, dass alle entsprechenden Instanzen selbst bei einmaligem Auftreten statistisch auffällig sind, sei abschließend die Frage danach gestellt, um wie viel Text eine *Digital Library* vergrößert werden muss, damit ein *Re-use Overlap* um genau ein weiteres Wort erweitert werden kann, um eine statistische Analyse zu zulassen, für die nicht $f_{epx} < 1$ gilt.

[8] *Ngram* der Länge 7.

Das sei an der *Ngram*-Approximation durch das 1000-häufigste Wort mit einer Wahrscheinlichkeit von $p(w_{1000}) = 7.86 \cdot 10^{-5}$ aus Tabelle 4.2 verdeutlicht. Gemäß der Formel 4.7 resultiert daraus, dass $log_{10}p(w_j) = -4.104$ ist. Um nun n um 1 zu erhöhen, müsste sich $log_{10}(M)$ um exakt 4.104 vergrößern, was wiederum einer Vergrößerung der Textbasis von $10^{4.104}$ entspräche. Die mehr als zehntausendfache Menge an Text kann aus Tabelle 4.2 abgelesen werden. Bei einer Normgröße von *10k* entspricht die zehntausendfache Menge der Normgröße *100M* Sätzen. Die entsprechenden numerischen Werte von $\lceil 1.635 \rceil$ und $\lceil 2.555 \rceil$ reflektieren die Größenabschätzung der vorangestellten mathematischen Überlegung.

Für ein vergleichbar großes englischsprachiges Normkorpus können diese Ergebnisse reproduziert werden. Jedoch wird für die *j=1*-Approximation ein Wert von $\lceil 8.107 \rceil$ anstelle von $\lceil 6.881 \rceil$ berechnet. Bei $j = 1000$ berechnet sich ein Wert von $\lceil 2.75 \rceil$, der nur unerheblich vom deutschsprachigen Korpus abweicht (vgl. Tabelle 4.2).

Fälschlicherweise wird aufgrund dieser sprachstatistischen Gegebenheiten die *Mutual Information* als schlechtes Maß bezeichnet, auch wenn sie, wie eingangs gezeigt, direkt aus der stochastischen Unabhängigkeit abgeleitet werden kann. Vielmehr sind es die seltenen Ereignisse, die es nicht ermöglichen, eine ernsthafte Sprachstatik zu betreiben. Wird die Menge der seltenen Ereignisse im Kontext des *Gesetzes der großen Zahlen* betrachtet, so scheint es eher nicht sinnvoll, für die meisten *Bigrams* oder *Co-occurrences* derartige *Scores* zu berechnen.

Dunning nähert sich diesem Problem aus der Testtheorie (vgl. [Dunning 1993]). Um zu messen, wie sehr einem *Ngram* oder einer *Co-occurrence* vertraut werden kann, wird hierzu eine Punktschätzung eingesetzt. In der statistischen Schätztheorie muss eine Verteilung angenommen werden. Aufgrund ihrer binären Entscheidung, ob etwas entweder mit einer Wahrscheinlichkeit p bestimmt oder mit der Gegenwahrscheinlichkeit $1-p$ nicht beobachtet werden kann, wird hierzu oftmals die Binomialverteilung $B(k, p, n)$, wie in Formel 4.8, mit $k \in [0, n]$ zugrunde gelegt. In der statistischen Schätztheorie können prinzipiell auch andere Verteilungen, wie die Normalverteilung, angenommen werden. Jedoch wird sich nachfolgend noch zeigen, dass die Binomialverteilung $B(k, p, n)$ einige Vorteile besitzt.

$$B(k, p, n) = \binom{n}{k} p^k (1-p)^{n-k} \tag{4.8}$$

Hierbei entspricht n der Anzahl aller Versuche und k der Anzahl der Versuche, welche ein positives Ergebnis geliefert haben. In der statistischen Schätztheorie wird die Binomialverteilung $B(k, p, n)$ als statistisches Modell eingesetzt, so dass die Wahrscheinlichkeit p als unbekannt angenommen wird und stellt somit den zu schätzenden Parameter θ dar (vgl. Formel 4.9).

$$B(k, \theta, n) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{4.9}$$

Formel 4.9 ist somit zu einer Funktion geworden, die zu gegebenen n und k nur noch von θ abhängig ist. Diese Funktion wird auch *Likelihood*-Funktion mit $L(\theta) = B(k, \theta, n)$ genannt. Von besonderem Interesse ist hierbei der Wert des Parameters θ, für welchen $L(\theta)$ maximal ist. Die zugrunde liegende Binomialverteilung bietet den Vorteil, dass es immer genau einen solchen Punkt gibt, für den $L(\theta)$ maximal wird. Dieser Punkt wird bei der hier angenommenen Binomialverteilung immer genau dann erreicht, wenn $\theta = \frac{k}{n}$ ist. Diese Eigenschaft ist insofern von Vorteil, dass $\frac{k}{n}$ exakt der Wahrscheinlichkeit entspricht, mit welcher bspw. ein *Ngram* oder eine *Co-occurrence* innerhalb einer *Digital Library* beobachtet werden kann.

Die Grundidee des *Likelihood-Ratio* ist, zwei Hypothesen unter einer gegebenen Beobachtung, die durch n und k gegeben ist, zu vergleichen (vgl. Formel 4.10).

$$\Lambda = \frac{L(\theta_0)}{L(\theta_1)} \tag{4.10}$$

Hierbei repräsentiert θ_1 die Alternativ- und θ_0 die Nullhypothese. Die Alternativhypothese θ_1 ist durch $\theta_1 = \max(\theta) = \frac{k}{n}$ gegeben. Während die Alternativhypothese dem Maximum der *Likelihood*-Funktion entspricht, wird genau wie bei der stochastischen Unabhängigkeit oftmals die statistisch erwartbare Wahrscheinlichkeit von n Wörtern eines *Re-use Overlaps* $p(w_1)p(w_2)\cdots p(w_n)$ angenommen. Beim *Likelihood Ratio* ist es jedoch auch möglich, für θ_0 die beobachtbare Wahrscheinlichkeit einer zweiten *Digital Library* zu wählen. Das *Likelihood Ratio* kann dementsprechend so verstanden werden, dass zu einer gegebenen Beobachtung die Nullhypothese im Kontext dieser Beobachtung interpretiert wird. Umso größer das *Likelihood Ratio* ist, um so weniger passt die Nullhypothese auf die beobachteten Daten und kann dementsprechend abgelehnt werden. Hierzu wird das *Likelihood Ratio* weiterführend logarithmisch zum *Log Likelihood Ratio* λ, wie in Formel 4.11, skaliert. Die logarithmische Skalierung erleichtert nicht nur die Lesbarkeit der Ergebnisse, sondern vielmehr verhält sich -2λ asymptotisch zum χ^2-Test, wodurch aus dem *Score* eines *Log Likelihood Ratio* direkt auf die Fehlerwahrscheinlichkeit, wie in den Tabellen A.1 bis A.10, geschlossen werden kann. So entspricht ein $-2\lambda = 10.828$ einem α-Signifikanzniveau bzw. einem α-Fehler von 0.001.

$$\lambda = \log\left(\frac{L(\theta_0)}{L(\theta_1)}\right) \tag{4.11}$$

Neben dem asymptotischen Verhalten des *Log Likelihood Ratio* -2λ zum χ^2-Test gibt es einige weitere Parallelen, auf die nachfolgend kurz eingegangen werden soll. Hierbei wird als Nullhypothese die statistische Erwartung zugrunde gelegt. Ziel ist es, dass *Log Likelihood Ratio* durch Umstellen sukzessive zu vereinfachen.

Da dem *Log Likelihood Ratio* durch die Binomialverteilung ein statistisches Modell mit genau einem n und einem k zugrunde liegt, ist der Binomialkoeffizient $\binom{n}{k}$ in $L(\theta_0)$ sowie $L(\theta_1)$ identisch und kann dementsprechend, wie in Formel 4.12, gekürzt werden.

$$-\lambda = \log\left(\frac{\theta_1^k(1-\theta_1)^{n-k}}{\theta_0^k(1-\theta_0)^{n-k}}\right) \tag{4.12}$$

In Formel 4.12 wurden zusätzlich Zähler und Nenner vertauscht, wodurch die nachfolgende Formel nun $-\lambda$ entspricht. Diese Formel kann wiederum durch einfaches Umstellen zu der Formel 4.13 umformuliert werden.

$$-\lambda = \log\left(\left(\frac{\theta_1}{\theta_0}\right)^k\left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-k}\right) \tag{4.13}$$

Unter Anwendung der Logarithmengesetze kann schließlich die Formel 4.14 abgeleitet werden.

$$-\lambda = k \cdot \log\left(\frac{\theta_1}{\theta_0}\right) + (n-k) \cdot \log\left(\frac{1-\theta_1}{1-\theta_0}\right) \tag{4.14}$$

Aus Formel 4.14 ist ersichtlich, dass für den Fall $\theta_1 = \theta_0$ immer $-\lambda = 0$ gilt. Somit kann die Nullhypothese nicht zurückgewiesen werden. Während die Alternativhypothese der beobachteten Wahrscheinlichkeit eines *Bigrams* oder einer *Co-occurrence* entspricht, ist die Nullhypothese als die statistische Erwartung angenommen. Deshalb entspricht $\theta_1 = p(w_1 w_2 \cdots w_n)$ und $\theta_0 = p(w_1)p(w_2)\cdots p(w_n)$.

Da in den einführenden Experimenten zu diesem Abschnitt bereits gezeigt werden konnte, dass über 90% aller Daten zehnmal und seltener beobachtet werden können, kann das zweite Glied für ebendiese und damit dem größeren Teil vernachlässigt werden. Da sowohl $1 - \theta_1 \approx 1$ als auch $1 - \theta_0 \approx 1$ gelten und sich somit für niedrige Frequenzen mehr oder weniger neutralisieren, wird für den Logarithmus ein Wert von nahe 0 berechnet.

Daher soll nur das erste Glied aus Formel 4.14 im Detail betrachtet werden. Es ergibt sich die Formel 4.14, welche der *Local* oder *Lexicographer's Mutual Information* entspricht (vgl. [Evert 2004]), die aus der bereits erwähnten *Mutual Information* besteht, welche wiederum mit der Häufigkeit des Auftretens multipliziert wird.

$$LMI = k \cdot \log \left(\frac{p(w_1 w_2 \cdots w_n)}{p(w_1)p(w_2) \cdots p(w_n)} \right) \tag{4.15}$$

Des Weiteren entspricht die *Local Mutual Information*[9] genau dann der *Kullback-Leibler-Divergenz*, wenn k mit $k = |w_1 w_2 \cdots w_n|$ nicht absolut eingesetzt wird, sondern die relative Wahrscheinlichkeit $p(w_1 w_2 \cdots w_n)$ benutzt wird (vgl. [Kullback 1951]). Sowohl das *Log Likelihood Ratio* als auch die *Local Mutual Information* und die *Kullback-Leibler-Divergenz* haben gemeinsam, dass sie aus zwei grundlegenden Komponenten bestehen. Einerseits misst die *Mutual Information* die Abweichung einer Beobachtung von der statistischen Erwartung. Auf der anderen Seite wird durch die Multiplikation mit k sichergestellt, dass selbst bei gleicher *Mutual Information* durch ein häufigeres Auftreten ein höherer *Score* zugewiesen wird. Aus dieser Wechselwirkung wird der mathematische Nachteil der *Mutual Information* bedingt verbessert. Da mit dem statistischen Problem einhergeht, dass k mit $1 \leq k \leq 10$ relativ klein ist, können häufigere *Bigrams* oder *Co-occurrences* mit einer kleinen und nicht durch Rauschen verzerrten *Mutual Information* dennoch einen höheren *Score* erreichen.

Der Fokus dieses Abschnittes ist auf die probabilistischen Modelle beschränkt, die für diese Arbeit relevant sind. Es kann sich jedoch einfach überlegt werden, dass die hier genannten und durch die *Power-Law*-Verteilung von Wörtern induzierten Probleme einfach auf Ansätze adaptiert werden können, die beispielsweise mit *bedingten Wahrscheinlichkeiten* arbeiten. Weiterhin gilt auch für das *Log Likelihood Ratio*, dass es keine statistische Aussage treffen kann, wenn *Ngrams* oder *Co-occurrences* zu selten auftreten.

Dementsprechend ist es zwar mathematisch möglich, jedoch in der konkreten Anwendung aus wissenschaftlicher Sicht nicht akzeptierbar, ein *Mining*-Ergebnis oder im Fall des *Text Re-use* einen *Re-use Overlap* darauf zu untersuchen, wie groß der statistische Einzelfehler ist, um daraus einen Gesamtfehler ableiten zu können.

4.3 Evaluierung von Sprachmodellen

Entgegen der statistischen Einzelsicht aus dem vorigen Kapitel können Sprach- und *Information Retrieval*-Modelle auch systemorientiert in ihren Ergebnissen untersucht werden. Grundlage für diese Form der Evaluierung ist ein *Gold Standard* bzw. *Ground Truth* (vgl. [Manning 2008] Seite 139 ff.). Eine solche Evaluierungsmethode setzt immer voraus, dass eine möglichst große Menge an qualitativen Daten vorliegt.

Anhand des *Gold Standards* kann demnach entschieden werden, ob ein Datensatz für ein nutzerspezifisches Bedürfnis *relevant* bzw. *nicht relevant* ist. Andererseits kann ein Datensatz durch eine *Retrieval*- oder *Mining*-Analyse entweder *erkannt* bzw. *nicht erkannt*

[9] Die gewichteten *Scores* aus Abschnitt 3.7 können nun anhand der *Local Mutual Information* motiviert werden. Das Multiplizieren der Häufigkeit einer Beobachtung mit der *Mutual Information* bzw. einem *Similarity-Score*, wie der *Resemblance*, bietet den Vorteil, dass größere *Re-use Overlaps* gegenüber kleineren bei einer gleichen *Mutual Information* oder *Resemblance* besser bewertet werden.

werden. Aus diesen zwei Dimensionen ergibt sich eine Matrix, wie in Tabelle 4.3 abgebildet, welche insgesamt vier mögliche Ausgänge der Analyse gegen einen *Gold Standard* aufzeigt.

	relevant	**nicht relevant**
erkannt	*true positives* (tp)	*false positives* (fp)
nicht erkannt	*false negatives* (fn)	*true negatives* (tn)

Tabelle 4.3: Vier mögliche Ausgänge einer Evaluierung gegen einen *Gold Standard*.

Aus den vier möglichen Ausgängen einer Analyse gegen einen *Gold Standard* (vgl. Tabelle 4.3) können verschiedene *Evaluierungsmaße* abgeleitet werden (vgl. [Manning 2008]). Die *Precision* P ist, wie in Formel 4.16, definiert.

$$P = \frac{tp}{tp + fp} \tag{4.16}$$

Die *Precision* P kann als der Anteil der *erkannten* Datensätze verstanden werden, welche gemäß des *Gold Standards* auch als *relevant* angesehen werden können. Dementgegen entspricht der *Recall* R dem Anteil derjenigen *relevanten* Datensätze, die auch erkannt (vgl. Tabelle 4.3) und dementsprechend, wie in Formel 4.17, definiert worden sind.

$$R = \frac{tp}{tp + fn} \tag{4.17}$$

Ein grundlegendes Problem einer Evaluierung mit beiden Maßen ist, dass sich *Precision* und *Recall* umgekehrt proportional zueinander verhalten. In der konkreten Anwendung stellt sich demnach immer die Frage, ob nach *Precision* oder *Recall* optimiert werden soll (vgl. u. a. [Büchler 2012c]). Das *F-Measure* kombiniert sowohl die *Precision* P als auch den *Recall* R zu einem *Score*. Hierbei wird die *Precision* P mit einem Wert $\alpha \in [0, 1]$ und der *Recall* R mit $1 - \alpha$, wie in Formel 4.18, gewichtet.

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} \tag{4.18}$$

Würde $\alpha = 0.5$ gewählt werden, was dem gleichen Gewicht für *Precision* und *Recall* entspricht, kann das *F-Measure* zur folgenden Formel vereinfacht werden.

$$F_{\alpha=0.5} = \frac{2PR}{P + R} \tag{4.19}$$

Um einen hinreichend großen *Gold Standard* für Evaluierungszwecke aufzubauen, gibt es verschiedene Methoden. Einerseits können Daten bspw. innerhalb einer *Crowd Sourcing*-Umgebung gesammelt werden. Andererseits besteht auch die Möglichkeit, die *Pooling*-Technik einzusetzen (vgl. [Harman 1995]), welche die Ergebnisse verschiedener Ansätze kombiniert. Hierbei sind diejenigen Datensätze hochgerankt und anschließend reviewt, die von möglichst vielen Systemen erkannt werden.

Der Vorteil einer solchen Evaluierung durch *Precision*, *Recall* und *F-Measure* besteht darin, dass das Ergebnis ganzheitlich bzw. eine Methode als System evaluiert und bewertet wird. Dennoch hat diese Form der Evaluierung auch signifikante Nachteile.

Während die *Precision* meist einfach bestimmbar ist, sind die *false negatives* innerhalb einer Evaluierung meist nicht bzw. nur unvollständig definiert (vgl. [Büttcher 2007]), wodurch der *Recall* eher als optimistische untere Schranke verstanden werden kann. Aus

diesem Grundproblem der Evaluierung durch einen *Gold Standard* resultieren nachhaltige Einschränkungen in der Bewertung von Evaluierungsergebnissen.

Erstens, ist die Stichprobe des *Gold Standards* repräsentativ genug für die Grundgesamtheit? Es sei hierzu ein Gedankenexperiment gemacht. In jeder *Digital Library* gibt es ein theoretisches Maximum an Informationen, welche mit jedweder Form von *Mining* aufgedeckt werden könnte. Im Vergleich zu diesem endlichen Maximum ist die Größe der *Evaluierungsdaten* meist um mehrere Dezimalstellen kleiner, so dass genau dieser *Gold Standard* einem unklaren *Sampling Bias* unterliegt. Dies bedeutet, dass oftmals unklar ist, inwiefern die Stichprobe auch unterschiedliche Phänomene der Grundgesamtheit in adäquater Weise repräsentiert. Das resultiert aus dem Paradoxon, bei dem versucht wird, sich dem Maximum zu nähern, ohne es jedoch bzw. bestenfalls fragmentarisch zu kennen.

Zweitens, im Weiterdenken des vorigen Gedankenexperimentes muss gefragt werden, wie sich ein *Gold Standard* bzgl. *Precision*, *Recall* oder *F-Measure* gegenüber des hypothetischen Maximums verhalten würde. Aufgrund dessen, dass ein *Gold Standard* meistens deutlich kleiner als das hypothetische Maximum ist, wäre selbst bei einer hohen *Precision*, der *Recall* sehr niedrig und damit zwangsläufig auch das *F-Measure*.

Drittens, eine Evaluierung durch einen *Gold Standard* kann auf eine maximale *Precision*, einem optimalen *Recall* oder das beste *F-Measure* optimiert werden. Eine Evaluierung reflektiert jedoch oftmals genau dieses Optimierungsverhalten nicht. Bedingt dadurch können Ergebnisse ähnlich gut jedoch durch andere Parameter auf unterschiedliche Kriterien optimiert sein. Letztlich ist es das Ziel, die Fähigkeit eines Algorithmus auf seine *Mining*-Fähigkeit zu überprüfen. Optimierungen auf bestimmte Kriterien sind jedoch bereits anwendungsspezifische Sichten und keine Evaluierung mehr.

Viertens, fokussierend auf die zu evaluierende *Mining*-Fähigkeit eines Verfahrens bzw. eines Ergebnisses muss im Kontext der *eHumanities* sowohl *Precision* als auch *Recall* speziell auf die fachwissenschaftliche Anwendung betrachtet werden. Während eine manuelle Annotation von Strukturen, bspw. *Text Re-use*, einer anzunehmenden hohen *Precision* und einem sehr geringen *Recall* unterliegt, erreicht eine automatische *Mining*-Methode einen deutlich höheren *Recall* bei einer geringeren *Precision*. Um dennoch eine ganzheitliche Vergleichbarkeit zwischen der automatischen als auch der manuellen Gewinnung von Ergebnissen herstellen zu können, wird oftmals das *F-Measure* eingesetzt. Selbst wenn bei einer manuellen Gewinnung von Ergebnissen die *Precision* mit $P = 1$ angenommen sei, gilt für den *Recall* R oftmals $R \ll 10^{-4}$, wodurch ein *F-Measure* mit $F \leq 2 \cdot 10^{-4}$ bestimmt werden würde. Auch für einen sehr schlechten automatischen Algorithmus, für den $R = P = 0.1$ angenommen sei, würde bereits ein *F-Measure* von $F = 0.1$ berechnet werden, welcher dem Vielfachen der fachwissenschaftlichen Leistungsfähigkeit entspricht. Unter diesem Aspekt muss die Frage neu gestellt werden, was wirklich gemessen wird. Dieses einfache Gedankenexperiment zeigt jedoch bereits einen signifikanten Widerspruch bzgl. der suggerierten *Mining*-Fähigkeit einer *Methode* durch einen Evaluierung gegen einen *Gold Standard* auf. De facto würde selbst der schlechteste automatische Algorithmus immer noch bzgl. des *F-Measure* als besser in einer Evaluierung gegen einen *Gold Standard* angesehen werden als das manuelle und mit einer hohen kognitiven Fähigkeit ausgestattete *Crowd Sourcing*.

Fünftens, es muss immer gefragt werden, ob die in allen Parametern identische Analyse auf der gleichen *Digital Library* auch dann das gleiche Evaluierungsergebnis reproduzieren würde, wenn ein anderer *Gold Standard* zugrunde läge. Beim wissenschaftlichen *Review* der Forschungsliteratur können nicht nur teilweise sehr diametral auseinander liegende Ergebnisse beobachtet werden, sondern diese Ergebnisse in einem Nachbau des Experimentes auf gleichen oder ähnlichen Daten teilweise bzw. nicht einmal ansatzweise reproduziert werden.

Sechstens, Ergebnisse sind oftmals sehr domänenspezifisch. Das bedeutet, dass sowohl die *Digital Library* als auch der *Gold Standard* aus der entsprechend gleichen Domäne, wie

der *Philosophie* oder der *Historiographie*, entstammen müssen. Ferner sind durch existierende Ergebnisse keine Rückschlüsse auf andere Domänen möglich. Dadurch ist die Aussagekraft einer Evaluierung deutlich eingeschränkt.

Siebtens, muss hinterfragt werden, ob ein gutes Ergebnis wirklich auch dieses beinhaltet. Letztlich kann nur die Kombination aus leicht zu findenden Strukturen im *Gold Standard* kombiniert mit einer einfach zu analysierenden *Digital Library* ein fälschlicherweise gutes Ergebnis generieren. Einfach zu findende Strukturen können sehr häufig auftretende Sprachphänomene sein, die offensichtlich sind. Ein einfacher Text kann u. a. durch eine geringe sprachliche Vielfalt beschrieben sein. So entscheidet ein Autor während des Schreibens von Texten, ob ein Objekt, wie in der Mathematik üblich, durch feste Bezeichner immer und immer wieder ausgedrückt wird oder ob im Sinne eines guten Schreibstils Wortcluster auf verschiedenste Art und Weise vermieden werden. Ersteres ermöglicht ein deutlich besseres Mining. Im zweiten Fall ist ein deutlich stärkeres Preprocessing nötig bzw. die Ergebnisse werden deutlich schlechter ausfallen, wenn die Texte gleich vorverarbeitet werden. Je nach Kombination aus *Gold Standard* und *Digital Library* können dadurch auch *Pseudo*-Effekte während der Evaluierung auftreten, die das wahre Ergebnis verschleiern.

Achtens, es kann unter Betrachtung der *Qualitätskriterien* aus Abschnitt 2.4 eine solche Evaluierung im Sinne des *Circumvention* sehr leicht vom Forscher beeinflusst werden. Immer wieder können Tricks beobachtet werden, wobei die vier Zellen der Tabelle 4.3 so modifiziert werden, dass die Evaluierung zugunsten eines Verfahrens und damit auch zum Vorteil einer Publikation eines Forschers ausfällt. Das kann oftmals sehr einfach dadurch erreicht werden, dass zugunsten des Interesses abgeschnitten wird. Dies sei an einem einfachen Beispiel kurz erklärt. Es sei wieder die Evaluierung von *Co-occurrences* bzw. deren Signifikanzmaße aus Abschnitt 4.2 zugrunde gelegt. Es kann gegen einen *Gold Standard* wie *WordNet*, *GermanNet* oder dem *Leipzig Annotation Project* (LAP) evaluiert werden, welche allesamt englisch bzw. deutsch getypte Wortpaarassoziationen enthalten. Da diese *Evaluierungsbasen* nicht für eine bestimmte *Digital Library* entwickelt worden sind, gibt es immer *Datensätze*, die nicht beobachtet werden können. Aus diesem Grund werden richtigerweise diese Datensätze aus der Evaluierung ausgeschlossen. Dies sei einmal als geschehen angenommen. Bei der Evaluierung von *Co-occurrences* liegt nun immer ein Signifikanzmaß zugrunde, welches einen *Score* berechnet und Daten unterhalb eines Schwellwert t abschneidet. Deshalb werden durch die *Power Law*-Verteilung selbst bei geringen Schwellwerten bereits große Datenmenge abgeschnitten, welche auch eine große Menge von Daten aus dem *Gold Standard* enthalten. Dementsprechend klein ist auch der *Recall* gegenüber dem *Gold Standard*. Wird nun der Aufbau eines Experimentes durch eine weitere Einschränkung begleitet wie *Es werden nur diejenigen Daten aus dem Gold Standard verwendet, welche mindestens dreimal beobachtet werden konnten*, dann bedeutet dies, dass die relevanten Daten des *Gold Standards* um etwa 70% reduziert werden[10]. Bedingt dadurch verringert sich bspw. der Nenner aus Formel 4.16, wodurch die *Precision* besser erscheint. Solche Formen des "intelligenten" Abschneidens haben auch einen großen Einfluss auf die *Mean Average Position* (vgl. [Buckley 2000]), bei welcher selbst bei einem top_x-Ansatz, die Ergebnisse zugunsten des Algorithmus verändert werden können.

Neuntens, es fehlt bei einem Test gegen einen *Gold Standard* die Sicht, wie das Evaluierungsergebnis ausfallen würde, wenn die Strukturen nicht manuell oder automatisch aufgedeckt werden, sondern eine zufällige Ergebnisliste generiert werden würde. Aufgrund dessen, dass sich das Ergebnis der Evaluierung auf die vier Zellen der Tabelle 4.3 nur anders

[10] Diese Zahl wurde im Rahmen eines Experimentes ermittelt. *Co-occurrences* unterliegen einem *Power Law*. Auch die Selektion von bestimmten Wortpaaren aus einem *Gold Standard*, wie *GermanNet*, unterliegt dieser Verteilung. Bedingt dadurch werden sehr viele der qualitativen Assoziationen nur ein- oder zweimal beobachtet.

verteilen würde, ist es offensichtlich wenig wahrscheinlich, dass sowohl *Precision* als auch *Recall* null sind. Deshalb ist ein *Mining*-Ergebnis immer aus der Sicht zu betrachten, wie gut es sich von einer Zufallsanalyse unterscheidet.

Zusammenfassend muss festgestellt werden, dass ein Test gegen einen *Gold Standard* zwar oft als einziges Instrument angesehen werden kann, jedoch aufgrund der neun genannten Aspekte ein Zweifel an der Aussagekraft der *Evaluation* bleibt. Im Rahmen dieser Arbeit würde eine solche Form der Evaluierung bedeuten, dass für jedes *Meme* und jeden *Re-use Style* eine adäquate Menge an Evaluierungsdaten zur Verfügung stehen müsste. Ferner sollte ein für die Evaluierung einzusetzender hybrider *Gold Standard* einer realistischen Verteilung der *Meme* und *Re-use Styles* entsprechen, die auch nach akkurater Bestimmung für eine andere *Digital Library* nicht mehr oder nur teilweise passend wäre.

Ziel dieser Form eines Tests ist das Bestimmen der *Mining*-Fähigkeit einer Methode. Realistisch jedoch wird die Übereinstimmung einer *Digital Library* mit einem *Gold Standard* mit vielen absichtlichen und unabsichtlichen ergebnisverfälschenden Einflüssen gemessen, wodurch oftmals Ergebnisse nicht adäquat reproduziert werden können, so dass die wissenschaftliche Aussagekraft sowohl des Tests selbst als auch der Evaluierungsmethode in Frage gestellt ist. Letztlich darf es neben den unabsichtlichen Einflüssen nicht möglich sein, dass Evaluierungen im Sinne einer forensischen Arbeitsweise durch das Zutun des Menschen aktiv beeinflusst werden können.

4.4 Noisy Channel Evaluation

In den Abschnitten 4.2 und 4.3 wurden sowohl ein statistischer Einzel- als auch ein systemorientierter Evaluierungsansatz vorgestellt. Beide haben zum Teil signifikante Nachteile. Der mathematische Ansatz zieht essenzielle statistische Probleme auf natürlichsprachlichen Texten nach sich. Der systemorientierte Ansatz hingegen ist zu vielen absichtlichen und unabsichtlichen Einflüssen während der Evaluierung ausgesetzt. Unabhängig davon kann keine der genannten Methoden in adäquater Weise mit der *Data Diversity* des *Historical Text Re-use* umgehen (vgl. Abschnitt 2.6). Aus diesem Grund wird in diesem Abschnitt eine neue und rein *quantitative Evaluierung* vorgestellt, welche sowohl die genannten Nachteile nicht auf sich vereint als auch resistent gegen die *Data Diversity* ist.

Hierbei sind zwei Fragestellungen wichtig. Erstens, wie kann die *Mining*-Fähigkeit einer Methode quantifiziert werden? Zweitens, wie muss ein *Evaluation Score* beschaffen sein, so dass er sowohl manuelle als auch automatische Methoden in gleicher Weise behandelt? Wie bereits in Abschnitt 4.3 dargestellt worden ist, würde faktisch jedes manuelle *Mining* aufgrund der Defizite in der Masse selbst einem schlechten Algorithmus in einer Evaluierung gegen einen *Gold Standard* unterliegen. Auch wenn anzunehmen ist, dass aufgrund der kognitiven Fähigkeit eines Menschen bzw. im Sinne des *Crowd Sourcing* einer Gruppe von Menschen die *Mining*-Fähigkeit der manuellen Methode deutlich größer sein sollte.

Speziell eine Evaluierung nach *Precision* und *Recall* zeigt in den beiden Maxima für $P = 1$ und $R = 1$ signifikante Nachteile auf. So kann ein Maximum in der *Precision* einfach dadurch erreicht werden, dass nur ein Datensatz manuell extrahiert wird, der auch im *Gold Standard* enthalten ist. Daraus würde bereits eine *Precision* von $P = 1$ resultieren. Eine *Precision* von $P = 1$ bei 10, 100, 1000 oder gar einer Millionen extrahierter Daten wäre deutlich höher zu bewerten. Der *Recall* hingegen kann einfach dadurch auf $R = 1$ gesetzt werden, indem alle möglichen Permutationen in die Ergebnismenge inkludiert werden. Vielmehr gilt hier, dass ein $R = 1$ umso besser ist, je kleiner die Ergebnismenge wird.

Wie kann eine ganzheitliche Systemevaluierung gestaltet sein, die zeitgleich minimalst bzw. gar nicht im Sinne der *Circumvention* getäuscht werden kann? In der *Künstlichen Intel-*

ligenz führte 1950 *Alan Turing* den nach ihm benannten *Turing-Test* ein (vgl. [Turing 1950]). Der theoretische Versuchsaufbau sieht vor, dass ein Mensch eine Eingabe an ein System schickt und nicht weiß, ob sich auf der anderen Seite ein Mensch oder ein Computer befindet. Rein auf Basis der Antwort wird diese Differenzierung unterschieden.

Eine ähnliche Anpassung wäre sicher auch für den *Historical Text Re-use* denkbar. Jedoch würde es sich als schwierig erweisen, diesen Versuchsaufbau für eine große Menge Daten direkt zu übersetzen. Aus der Signal- und Satellitentechnik hingegen kann eine stark vereinfachte Form eines solchen Tests abgeleitet werden, welches die Signalstärke im Vergleich zum Rauschen misst. Im Kontext des *Turing-Tests* entspräche der Mensch dem Signal und der Computer dem Rauschen.

Das *Signal-Noise-Ratio* (SNR) setzt beide Größen, wie in Formel 4.20, in ein Verhältnis. Hierbei kann P_{signal} als eine Quantifizierung der Strukturen sowie P_{noise} als ein ungewolltes Signal, dem Rauschen, verstanden werden.

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{4.20}$$

Da zu erwarten ist, dass das Signal deutlich größer als das Rauschen ist, wird das *Signal-Noise-Ratio* auch als logarithmisches Maß, wie in Formel 4.21, beschrieben. Die Einheit dieses logarithmierten *Signal-Noise-Ratio* ist *Bel*. Da aufgrund der logarithmischen Skalierung sehr kleine Werte berechnet werden, wird der skalierte *Score* mit 10 multipliziert und die Einheit ist dann dementsprechend Dezibel (kurz dB).

$$SNR_{db} = 10 \cdot \log_{10}\left(\frac{P_{signal}}{P_{noise}}\right) \tag{4.21}$$

In der konkreten Umsetzung, die *Mining*-Fähigkeit einer Methode bzw. eines Sprachmodells messen zu können, kommt unmittelbar die Frage auf, wie P_{Signal} und P_{Noise} bestimmt werden können.

Hierzu wird die *Mining*-Analyse, wie bereits im Abschnitt 1.6 einführend skizziert, in Shannon's *Noisy Channel Theorem* eingebettet. Eine *Digital Library* wird hierbei als ein diskretes Ausgangssignal $\mathcal{S}$ aufgefasst. Das stellt somit eine Modifikation von Shannon's Experiment dar, welchem ein akustisches, kontinuierliches Signal zugrunde liegt. Einerseits wird das Ausgangssignal über einen natürlichen *Noisy Channel* $\mathcal{S} \oplus \mathcal{N}$ übertragen, welcher lediglich ein natürliches Rauschen $\mathcal{N}$, wie bspw. Sprachevolution, beinhaltet. Zeitgleich wird das gleiche Signal $\mathcal{S}$ über einen zweiten *Noisy Channel* $\mathcal{S} \oplus \mathcal{N} \oplus \mathcal{A}$ gesendet, welcher zusätzlich ein künstliches Störsignal $\mathcal{A}$ enthält (vgl. Abb. 4.3). Im Ergebnis empfangen die *Receiver* aus dem Versuchsaufbau (vgl. Abb. 4.3) zwei Signale, wobei eines davon ein natürliches und das zweite ein stark verrauschtes bzw. zufälliges Signal ist.

Im weiteren Versuchsaufbau (vgl. Abb. 4.3) wird nun auf beide übertragenen Signale $\mathcal{S}'$ und $\mathcal{S}''$ das gleiche *Mining*-Verfahren mit exakt den selben Parameter angewendet. Aus dem in Abschnitt 4.2 dargestellten Problemen folgt direkt, dass sowohl die Ergebnismenge des natürlichen als auch die des zufälligen Signals nicht leer sein werden. Die beiden Ergebnismengen können als P_{Signal} und P_{Noise} verstanden werden.

Die *Artificial Noise Source* aus Abb. 4.3 kann als eine *Randomisierung* einer natürlichsprachlichen *Digital Library* verstanden werden. Es können verschiedene Randomisierungstechniken eingesetzt werden, welche in Abschnitt 4.5 im Detail erklärt werden. Für das weitere Verständnis in diesem Abschnitt sei als Randomisierung ein einfaches *Word Shuffling* angenommen, welches Wort für Wort eine *Digital Library* so verändert, dass eine zufällige Wort-Position bestimmt und das aktuelle Wort mit dem Wort an der zufällig bestimmten Position ausgetauscht wird, wodurch eine *Randomised Digital Library*, wie in Definition 19, generiert wird.

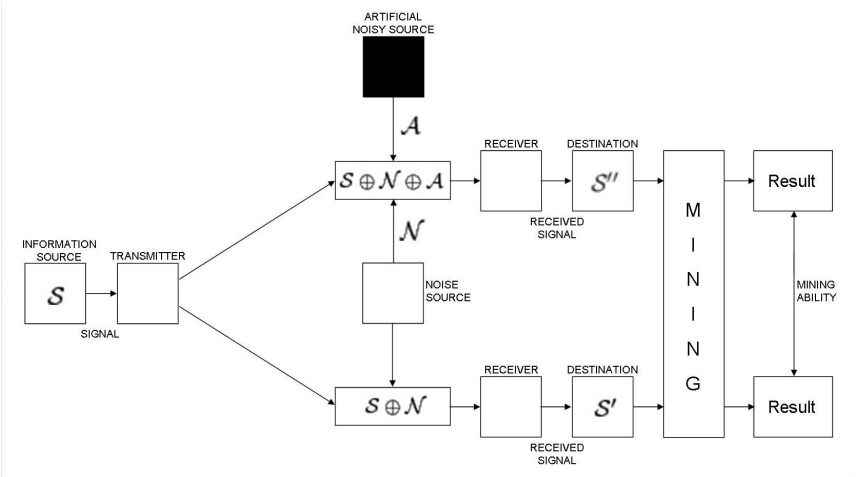


Abbildung 4.3: Versuchsaufbau der *Noisy Channel Evaluation*. Bei dieser Form der quantitativen Evaluierung werden die Ergebnisse sowohl eines natürlichen als auch eines zufälligen Signales verglichen und formen den *Score* der *Mining Ability* einer quantitativen Evaluierung.

Definition 19 (Randomised Digital Library)**.** *Sei ein Digital Library D_S einer Sprache S gegeben. Als eine Randomised Digital Library D_S^m wird die durch eine künstliche Randomisierung der Methode m veränderte Digital Library bezeichnet.*

Aus Abb. 4.3 ist ersichtlich, dass eine *Mining*-Methode sowohl auf die *Digital Library* D_S als auch auf die bezüglich einer Randomisierungsmethode m veränderten *Randomised Digital Library* D_S^m angewendet wird. Wie zur Formel 4.21 bereits reflektiert wurde, kann das *Signal-Noise-Ratio* als das logarithmisch skalierte Verhältnis der Ergebnisse eines *Mining*-Verfahrens auf einer *Digital Library* sowie einer *Randomised Digital Library* verstanden werden. Hierbei wird das Ergebnis aus D_S als das Signal P_{signal} bezeichnet. Das Ergebnis aus D_S^m hingegen wird *Quantitative Noise* P_{noise}^m genannt (vgl. Definition 19).

Definition 20 (Quantitative Noise)**.** *Sei eine Randomised Digital Library D_S^m aus Definition 19 gegeben. Als Quantitative Noise wird die Ergebnismenge einer Mining-Methode auf D_S^m definiert.*

Definition 20 zeigt im Kontext von Abschnitt 4.2 auf, dass selbst einfache Sprachmodelle im niederfrequenten Bereich systematische Probleme aufwerfen, so dass die *Quantitative Noise* im Idealfall zwar als 0 anzusehen ist, jedoch praktisch nur durch sehr hoch gewählte Schwellwerte erreicht werden kann. In Abschnitt 4.6 werden hierzu die *Bigram*- und *Co-occurrence*-Analysen aus Abschnitt 4.2 mit einer *Noisy Channel Evaluation* wiederholt.

Aus Abbildung 4.3 ist ebenfalls ersichtlich, dass das künstliche Störsignal $\mathcal{N}$ als ein additives Rauschen verstanden werden kann. Aus der Signaltechnik, wie der Tontechnik (vgl. [Henle 2001]), können sowohl das *White Noise* als konstantes Rauschen über dem Frequenzspektrum, das *Pink Noise*, dem $\frac{1}{f}$-Rauschen, als auch das *Brownian Noise*[11], dem $\frac{1}{f^2}$-Rauschen, für die Evaluation einer *Mining*-Methode adaptiert werden, wodurch drei verschiedene Arten von *Quantitative Noise* aus Definition 20 möglich sind. Nachfolgend bleiben jedoch sowohl das *Pink* als auch das *Brownian Noise* unberücksichtigt, so dass von einem *Additive White Noise* ausgegangen werden kann.

[11] Alternativ wird das *Brownian Noise* auch *Red Noise* genannt.

Die *Noisy Channel Evaluation* kann demnach als eine ganzheitliche und rein quantitativ evaluierende Sicht eine *Mining*-Methode verstanden werden, welche einerseits eine möglichst große Treffermenge innerhalb einer *Digital Library* sowie andererseits jedoch eine möglichst geringe Ergebnismenge auf einer *Randomised Digital Library* generieren soll. Durch dieses Verhältnis kann die Eigenschaft *Mining Ability* beschrieben werden. Die *Mining Ability* (vgl. Definition 21) ist genau dann groß, wenn entweder ein Verfahren besonders gut funktioniert oder aber auch wenn eine *Digital Library* viele Strukturen enthält, die durch eine Methode sehr gut extrahiert werden können.

Definition 21 (Mining Ability)**.** *Sei ein Signal-Noise Ratio SNR einer Digital Library D_S und einer Randomised Digital Library D_S^m gegeben. Als Mining Ability wird das Signal-Noise-Ratio einer Mining-Methode während der Noisy Channel Evaluation bezeichnet.*

Die *Mining Ability* aus Definition 21 kann demnach als ein Leistungsmaß verstanden werden, welches die von Zufall abweichenden Strukturen natürlicher Sprachen quantifiziert, so dass diese Kenngröße mit anderen *Digital Libraries* als auch Verfahren ganzheitlich verglichen bzw. als quantitative Evaluierungsgröße aufgefasst werden kann. Die *Mining Ability* zu einem gegebenen Parameterraum Θ wird mit $\mathcal{L}_{Quant}(\Theta)$, wie in Formel 4.22, abgekürzt.

Während das Bestimmen von P_{signal} und P_{noise} einfach als die Menge extrahierter Strukturen verstanden werden kann, entspricht $\mathcal{L}_{Quant}(\Theta)$ bei einer Generierung eines *Graphen* $G = (V, E)$ der Menge der Kanten E (vgl. Formel 4.22). Das gilt sowohl für die *Bigram*- und *Co-occurrence*-Graphen aus Abschnitt 4.2 als auch den *Text Re-use Graphen* zu dieser Arbeit. Für einen *Text Re-use Graph*, welcher durch einen *Text Re-use* ϕ_Θ generiert worden ist, ergibt sich somit $\mathcal{L}_{Quant}(\Theta)$ aus Formel 4.22[12].

$$\mathcal{L}_{Quant}(\Theta) = 10 \cdot \log_{10} \frac{|E_{D_S,\phi_\Theta}|}{\max(1, |E_{D_S^m,\phi_\Theta}|)} dB \tag{4.22}$$

Speziell im Kontext des *Historical Text Re-use* wurde in Abschnitt 2.6 bereits eingeführt, dass eine *Text Re-use Analysis* aus einem zusammengesetzten *Hybrid Text Re-use* besteht. Formel 4.23 berechnet $\mathcal{L}^H_{Quant}(\Theta)$ für eine hybride Analyse einer *Digital Library* gegenüber einer *Randomised Digital Library* bestehend aus insgesamt n generierten *Text Re-use Graphen*, welche aus n verschiedenen *Text Re-use Analysis* resultieren.

$$\mathcal{L}^H_{Quant}(\Theta) = 10 \cdot \log_{10} \frac{|\bigcup_{i=1}^{i\leq n} E^i_{D_S,\phi_\Theta}|}{\max(1, |\bigcup_{i=1}^{i\leq n} E^i_{D_S^m,\phi_\Theta}|)} dB \tag{4.23}$$

Da $\mathcal{L}_{Quant}$, und natürlich auch $\mathcal{L}^H_{Quant}$, als ein Leistungsmaß für eine *Mining*-Methode verstanden werden kann, ist in Abhängigkeit von Θ bzw. unterschiedlichen Größen einer *Digital Library* von Interesse, sowohl die *Minimal*- als auch die *Maximalleistung* zu bestimmen und zu welchem Parameter bzw. bei welcher Textgröße diese Extrema erreicht werden (vgl. Abschnitt 4.6). Während die *Minimalleistung* $\mathcal{L}_{min}(\Theta)$ definiert ist als

$$\mathcal{L}_{min}(\Theta) = \min(\mathcal{L}_{Quant}(\Theta)) \tag{4.24}$$

kann die *Maximalleistung* als

$$\mathcal{L}_{max}(\Theta) = \max(\mathcal{L}_{Quant}(\Theta)) \tag{4.25}$$

verstanden werden. In Analogie zu $\mathcal{L}_{min}(\Theta)$ aus Formel 4.24 und $\mathcal{L}_{max}(\Theta)$ aus Formel 4.25 können auch $\mathcal{L}^H_{min}(\Theta)$ und $\mathcal{L}^H_{max}(\Theta)$ bestimmt werden.

[12] Der *max*-Operator in Formel 4.22 verhindert, dass der Nenner nicht 0 werden kann, da andernfalls $\mathcal{L}_{Quant}(\Theta)$ nicht immer definiert wäre.

Sowohl die eingeführten Größen der quantitativen *Mining Ability* $\mathcal{L}_{Quant}(\Theta)$ als auch die Minimalleistung $\mathcal{L}_{min}(\Theta)$ und die Maximalleistung $\mathcal{L}_{max}(\Theta)$ beziehen sich hierbei auf die quantitative Leistung eines *Text Re-use* ϕ_Θ und nicht auf eine etwaige intellektuelle Leistungsfähigkeit, wie einer Fehlalignierung zweier *Re-use Units*, die nach fachwissenschaftlichen Aspekten kein *Text Re-use* sind. Jedes Verfahren aus Kapitel 3 folgt dem darunterliegenden objektiven Modell. Dadurch können *Re-use Units* aufeinander gelinkt werden, denen zumindest kein absichtlicher *Re-use* zugerechnet werden kann. Dies soll an zwei Beispielen kurz verdeutlicht werden.

Einerseits sei der bereits erwähnte Ausspruch *Sein oder nicht sein, das ist hier die Frage* von William Shakespeare und andererseits die Phrase *im Namen unseres Herren Jesus Christus* hierfür ausgewählt. Aus informationstechnischer Sicht folgen beide Zeichenketten festen syntaktischen Regeln, so dass jedes syntaktische Verfahren alle *Re-use Units*, die diese beiden Wortfolgen enthalten, jeweils in einem *Re-use Graph G* enthalten sein werden. Auch wenn der Shakespeare-Ausspruch vom Fachwissenschaftler eher gewollt ist, kann die christliche Phrase weniger von geisteswissenschaftlichem Interesse aufgefasst werden. Das Dilemma wird unter Annahme eines semantischen Verfahrens (vgl. Abb. 3.5 auf Seite 102) noch deutlicher. Während im Shakespeare-Ausspruch lediglich *Frage* als diskriminierendes Wort bestimmt werden kann, sind es in der christlich geprägten Phrase gleich vier: *Namen*, *Herren*, *Jesus* sowie *Christus*. Wird ein etwaiger *Scoring*-Schwellwerte so gewählt, dass entsprechende Kanten in E die Phrase nicht im *Re-use Graph G* enthalten, dann ist automatisch auch der Shakespeare-Ausspruch mit entfernt worden. Wird auf der anderen Seite der *Threshold* so niedrig gewählt, dass auch der Ausspruch von Shakespeare im *Re-use Graph G* enthalten bleibt, dann ist auch die Phrase wieder enthalten.

Zu Beginn dieses Abschnittes wurde einführend dargestellt, dass die beiden Extrema einer *Precision* und eines *Recalls* mit $P = 1$ und $R = 1$ relativ einfach erreicht werden können. Für die *Mining Ability* ist die Maximalleistung $\mathcal{L}_{max}(\Theta)$ wesentlich schwerer absichtlich als auch unabsichtlich zu manipulieren. Ein $R = 1$ würde wie im dazugehörigen Beispiel bedeuten, dass $\mathcal{L}_{Quant}(\Theta) = 0$ ist. Auch die *Precision* $P = 1$ von nur einem richtigen Datensatz zieht ein $\mathcal{L}_{Quant}(\Theta) = 0$ nach sich. Vielmehr steigt die *Mining Ability* bei wachsender Anzahl von 10, 100, 1000 und mehr relevanten Datensätzen ebenfalls sukzessive, so dass zu einer gegebenen *Precision* die Größe des positiven *Signals* die *Mining Ability* dementsprechend beeinflusst.

Da die *Mining Ability* aufgrund der logarithmischen Skalierung nicht beliebig groß werden kann, sei eine allgemeine obere Abschätzung gemacht. Hierzu sei für $E_{D_S^m,\phi_\Theta}$ die theoretische Annahme aus Formel 4.26 mit $\varepsilon \in [0, 1]$ angenommen.

$$|E_{D_S^m,\phi_\Theta}| = \varepsilon \cdot |E_{D_S,\phi_\Theta}| \tag{4.26}$$

Durch diese Annahme für $E_{D_S^m,\phi_\Theta}$ kann die Formel 4.22 wie folgt vereinfacht werden.

$$\mathcal{L}_{Quant} = 10 \cdot \log_{10} \frac{1}{\varepsilon} dB \tag{4.27}$$

ε kann hierbei als Fehler der *Randomised Digital Library* verstanden werden. Tabelle 4.4 reflektiert den Verlauf der *Mining Ability* in Abhängigkeit des Fehlers ε.

ε	10^{-4}	10^{-3}	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
$\mathcal{L}_{Quant}$	40	30	20	17	15.2	14	13	12.2	11.6	11	10.5	10

Tabelle 4.4: Zusammenhang zwischen dem Fehler ε und der *Mining Ability* $\mathcal{L}_{Quant}$.

Tabelle 4.4 zeigt auf, dass bei einem Fehler von $\varepsilon = 0.1$, welcher einem Verhältnis des *Signals* und dem *Quantitative Noise* von *10:1* entspricht, eine *Mining Ability* von $\mathcal{L}_{Quant} = 10$ bestimmt wird. Ist das Verhältnis der Ergebnisse von *Signal* zum *Quantitative Noise* $100 : 1$, dann beträgt die *Mining Ability* gerade einmal $\mathcal{L}_{Quant} = 20$. Wenn der Fehler mit $\varepsilon = 10^{-4}$ angenommen sei, entspricht das einer *Mining Ability* von $\mathcal{L}_{Quant} = 40$. Hierbei können für jede fälschlicherweise erkannte Struktur der *Randomised Digital Library* D_S^m 10000 Strukturen der regulären *Digital Library* D_S aufgedeckt werden. Es kann sich weiterführend überlegt werden, dass eine *Mining Ability* von $\mathcal{L}_{Quant} > 100$ dB nahezu unwahrscheinlich ist. Tabelle 4.5 kann daher als eine Grobklassifikation von sowohl manuellen als auch automatischen *Mining*-Methoden verstanden werden. Während für $\mathcal{L}_{max}(\Theta)$ in der Regel ein Wert zwischen 10 dB $< \mathcal{L}(\Theta) \leq 40$ angesetzt werden kann, sollte es das Ziel sein, $\mathcal{L}_{min}(\Theta)$ möglichst nicht im negativen Bereich bestimmen zu müssen.

Leistungsbereich	**Wertung**
$\mathcal{L}(\Theta) > 20$ dB	*sehr hohe Leistung* durch ein qualitatives *Crowd Sourcing*, sehr gute Verfahren mit optimiertem Parameterraum Θ
10 dB $< \mathcal{L}(\Theta) \leq 20$ dB	*hohe Leistung* durch qualitative und manuelle Arbeit eines Einzelforschers bzw. Verfahren mit einem guten Parameterraum Θ
0 dB $< \mathcal{L}(\Theta) \leq 10$ dB	*niedrige Leistung* durch entweder ein relativ schlechtes Verfahren, einen suboptimalen Parameterraum Θ oder auch dem Fakt, dass die *Digital Library* entgegen der Annahme des Forschers nur eine geringe Menge an Strukturen enthält
$\mathcal{L}(\Theta) = 0$ dB	*Zufall.* Bei einem $\mathcal{L}(\Theta) = 0$ liegt eine Balance aus *Signal* und *Quantitative Noise* vor. Das wird immer genau dann erreicht, wenn es sich um *Zufall* und dementsprechend *keine Struktur* handelt.
-10 dB $\leq \mathcal{L}(\Theta) < 0$ dB	*schwaches Rauschen* durch sehr schlechte Verfahren mit sehr schlecht bis keinen restriktiven Parametern in Θ
$\mathcal{L}(\Theta) < -10$ dB	*starkes Rauschen*: die *Mining*-Methode kann auf diese *Digital Library* nicht angewendet werden

Tabelle 4.5: Grobklassifikation von *Methoden* und deren *Mining Ability* $\mathcal{L}(\Theta)$ inklusive einer Wertung. Der angegebene Leistungsbereich und dessen Wertung basiert auf Erfahrungen im Umgang mit der *Mining Ability.*

Neben einer möglichst hohen maximalen *Mining Ability* $\mathcal{L}_{min}(\Theta)$ kann es als ein *Qualitätskriterium* verstanden werden, wenn die Differenz $\Delta(\Theta)$ zwischen $\mathcal{L}_{max}(\Theta)$ und $\mathcal{L}_{min}(\Theta)$, wie in Formel 4.28, möglichst gering ist (vgl. Abschnitt 4.6).

$$\Delta(\Theta) = \mathcal{L}_{max}(\Theta) - \mathcal{L}_{min}(\Theta) \tag{4.28}$$

Eine kleine Differenz $\Delta(\Theta)$ bedeutet, dass ein Verfahren gegenüber dem Parameterraum Θ oder der Größe der *Digital Library* aufgrund des schmalen Evaluierungskorridors möglichst resistent ist[13].

[13] Auch wenn bereits erwähnt, sei dennoch noch einmal darauf hingewiesen, dass eine kleine Differenz nur dann positiv zu werten ist, wenn $\mathcal{L}_{max}(\Theta) \geqq 20$ ist.

Des Weiteren kann, wie eingangs bereits erwähnt, die *Mining Ability* auch zum Vergleich und der Evaluierung einer manuellen, dem *Qualitative Re-use*, als auch einer automatischen Methode, dem *Quantitative Re-use*, verwendet werden. Für letztere Evaluierungsform wurde bereits $\mathcal{L}_{Quant}(\Theta)$ eingeführt. Für $\mathcal{L}_{Qual}(\Theta)$ kann zwar prinzipiell im Sinne der besseren Vergleichbarkeit das gleiche Vorgehen wie für $\mathcal{L}_{Quant}(\Theta)$ gewählt werden. Jedoch bietet dieses Vorgehen den Nachteil, dass keine *Acceptance* der Fachwissenschaftler zu erwarten ist, auf einer *Randomised Digital Library* nur für Evaluierungszwecke zu arbeiten. Aus diesem Grund kann für eine manuelle Methode, wie sie durch *Crowd Sourcing* gegeben ist, P_{noise} auch als die Menge der Fehlannotationen verstanden werden. Im Kontext des *Historical Text Re-use* entspricht diese Menge denjenigen Kanten eines *Re-use Graph*, welche kein *Text Re-use* sind. Eine solche Entscheidung kann bspw. durch einen *Reviewer* oder innerhalb der Ausbildung durch einen *Lehrenden* entschieden werden. Durch ein solches Review wird die Menge der Kanten E eines *Re-use Graph* $G = (V, E)$ in zwei Mengen E_{signal} und E_{noise} mit der Eigenschaft $E_{signal} \cap E_{noise} = \emptyset$ aufgesplittet, so dass $\mathcal{L}_{Qual}(\Theta)$, wie in Formel 4.29, bestimmt werden kann.

$$\mathcal{L}_{Qual}(\Theta) = 10 \cdot \log_{10} \frac{|E_{signal}|}{\max(1, |E_{noise}|)} dB \tag{4.29}$$

Tabelle 4.6 stellt den *Qualitative Re-use* und den *Quantitative Re-use* bzgl. ihrer Vor- und Nachteile sowie der Messmethode und einer zu erwartenden Leistungsfähigkeit beschrieben durch die *Mining Ability* $\mathcal{L}_{Quant}(\Theta)$ und $\mathcal{L}_{Qual}(\Theta)$ dar. In der Regel kann erwartet werden, dass $\mathcal{L}_{Qual}(\Theta) \geq \mathcal{L}_{Quant}(\Theta)$ gilt. Jedoch ist die Leistung, beschrieben durch die *Mining Ability* $\mathcal{L}_{Qual}(\Theta)$, im Maximum in den meisten Fällen sehr stark beschränkt. In Tabelle 4.6 wird als zu erwartende Leistung eine *Mining Ability* $\mathcal{L}_{Qual}(\Theta)$ von 30 - 40 dB angegeben. Das entspricht in Anlehnung an Formel 4.27 und Tabelle 4.4 einem Fehler auf 1000 bzw. 10000 Datensätzen. Auch wenn ein sehr kleiner Fehler anzunehmen ist, so ist es per Definition eine grundlegende Eigenschaft der *Mining Ability*, dass auch die Menge einen großen Einfluss auf den *Score* hat. So stellen selbst bei nur einem Fehler für die meisten *Crowd Sourcing*-Umgebungen 10000 Datensätze bereits eine Herausforderung dar. Das bedeutet demnach, dass zwar der Fehler minimal ist, jedoch durch die geringe Menge der manuell gesammelten Daten, der *Score* nur sehr selten größer als 40 dB liegen wird. Für $\mathcal{L}_{Quant}(\Theta)$ kann eine Methode bereits als gut angesehen werden, wenn sie 10 dB überschreitet. In der Regel wird ein $\mathcal{L}_{Max}(\Theta)$ von 35 dB für den *Quantitative Re-use* nicht überschritten. Jedoch zeigen bereits einfache Experimente in Abschnitt 4.6, dass auch eine Leistung von 70 - 80 dB für $\mathcal{L}_{Quant}(\Theta)$ beobachtbar sind.

	Qualitative Re-use	**Quantitative Re-use**
Vorteile	Qualität	Objektivität, nahezu beliebig große Textmengen können verarbeitet werden
Nachteile	Subjektivität, geringe Menge	geringe Qualität
Messmethode	$\mathcal{L}_{Qual}(\Theta)$	$\mathcal{L}_{min}(\Theta)$, $\mathcal{L}_{Quant}(\Theta)$, $\mathcal{L}_{max}(\Theta)$
Zielleistung	30 - 40 dB	10 - 80 dB

Tabelle 4.6: Vergleich des qualitativen und quantitativen *Re-use*.

Insbesondere beim *Qualitative Re-use* ist eine rein binäre Entscheidung auf die beiden Mengen E_{signal} und E_{noise} nicht nur äußerst schwierig sondern auch oftmals nicht gesichert

umsetzbar. Das resultiert im Kontext von historischen Sprachen, wie dem Lateinischen oder dem Altgriechischen, einfach daraus, dass es keine Muttersprachler für diese Sprache mehr gibt. Als notwendige Konsequenz daraus kann eine Aussage immer nur bis zu einem gewissen Grad als sicher gelten. Aus diesem Grund sei abschließend noch die *Weighted Mining Ability* eingeführt.

Bei der *Weighted Mining Ability* wird schließlich jedem Element j eines *Signals* ein Gewicht $\gamma_j \in [0,1]$ sowie jedem Element des *Quantitative Noise* ein Gegengewicht $1-\gamma_j$ zugeordnet. Durch die *Gewichtungsfunktion* γ kann somit die Formel 4.30 für die *Weighted Mining Ability* abgeleitet werden. Da für den *Quantitative Re-use* jedes Element des *Signals* durch $\gamma_j \in [0,1]$ gewichtet wird, muss das Restgewicht $1-\gamma_j$ dem *Quantitative Noise* zugerechnet werden. Deshalb wird das *Quantitative Noise* aus Formel 4.30 sowohl auf E_{D_S,ϕ_Θ} als auch $E_{D_S^m,\phi_\Theta}$ berechnet. Die entsprechende Mengenvereinigung wird $E'_{\phi_\Theta} = E_{D_S,\phi_\Theta} \cup E_{D_S^m,\phi_\Theta}$ genannt.

$$\mathcal{L}_{Quant}(\Theta) = 10 \cdot \log_{10} \frac{\sum_{e_j \in E_{D_S,\phi_\Theta}} \gamma_j}{\sum_{e_j \in E'_{\phi_\Theta}} 1-\gamma_j} dB \tag{4.30}$$

Für den *Qualitative Re-use* kann die Formel 4.30 insofern vereinfacht werden, als dass für alle $e \in E$ die γ_j sowie die $1-\gamma_j$ jeweils aufsummiert und als P_{signal} und P_{noise} aus Formel 4.20 aufgefasst werden. Ganz im Sinne eines *Philological Crowd Sourcing* kann γ_j als die *Interrater-Reliabilität* aufgefasst werden, welche ein hohes Gewicht genau dann bestimmt, wenn möglichst viele *Rater* bzw. Personen das gleiche Urteil abgeben (vgl. Formel 4.31). Für eine *Multirater*-Umgebung, wie sie durch das *Philological Crowd Sourcing* gegeben ist, kann das *Fleiss'* κ (vgl. [Fleiss 2004]) bzw. nur ein Teil wie in Formel 4.31 des κ, hierfür eingesetzt werden. Während das *Fleiss* κ eine Gesamtreliabilität berechnet, ist für das hier genannte Szenario lediglich die Berechnung des *Scores* $\gamma_j \in [0,1]$ von Interesse, welcher das von insgesamt R Ratern zugeordnete Verhältnis zu c_i Kategorien bzw. Entscheidungen bestimmt.

$$\gamma_j = \frac{1}{R(R-1)} \left(\sum_{i=1}^{n} c_i^2 - c_i \right) \tag{4.31}$$

Aus Formel 4.31 ist ersichtlich, dass γ_j nicht kleiner als 0 sowie nicht größer als 1 werden kann. Das Maximum wird genau dann erreicht, wenn alle R *Rater* die gleiche Entscheidung c_1 angeben. Für diesen Fall ist $n = 1$ und die Summe entspricht $R^2 - R$. Für γ_j gilt somit $\gamma_j = 1$. Das Minimum der *Interrater-Reliabilität* kann genau dann berechnet werden, wenn jeder *Rater* eine andere Entscheidung trifft. In Formel 4.31 gilt somit sowohl $n = R$ als auch $c_i = 1$, wodurch $c_i^2 - c_i = 0$ ist und somit die R-mal 0 aufsummiert wird. Als Konsequenz dessen wird $\gamma_j = 0$. Für jede andere Entscheidung gilt dementsprechend $\gamma_j \in [0,1]$.

Die gleiche Methode einer *Weighted Mining Ability* kann auch für den *Quantitative Re-use* eingesetzt werden, so dass anstelle der *Rater* verschiedene *Mining*-Verfahren getestet werden. Je mehr automatische Methoden einen *Text Re-use* erkennen, desto größer ist die *Interrater-Reliabilität* γ_j für diese einzelne Information bzw. Struktur. Auf der anderen Seite kann auch das in Abschnitt 4.2 bereits genannte asymptotische Verhalten des *Log-Likelihood Ratio* zum χ^2-Test dazu eingesetzt werden, aus der Gegenwahrscheinlichkeit zur kumulativen Wahrscheinlichkeitsverteilung den statistischen Fehler zu berechnen. Auch wenn das die einzige Möglichkeit ist, einen statistischen Fehler zu berechnen, so bleiben dennoch die in Abschnitt 4.2 genannten statistischen Probleme. Aus diesem pragmatischen Grund wird im Rahmen dieser Arbeit für den *Quantitative Re-use* die ungewichtete *Mining Ability* aus Formel 4.22 bzw. für den *Hybrid Text Re-use* aus Formel 4.23 eingesetzt.

4.5 Arten einer *Randomised Digital Library*

Die *Mining Ability* ist in Abschnitt 4.4 als eine rein quantitative Evaluierungstechnik vorgestellt worden, welche ohne einen qualitativ zugrunde liegenden *Gold Standard* auskommen kann. Anstatt gegen eine solche Evaluierungsbasis zu testen, die oftmals nur teilweise zu einer *Digital Library* passt, wird das Verteilungsverhalten der Ergebnisse zu einer *Randomised Digital Library* verglichen. In Abschnitt 4.4 wurde hierzu bereits kurz das *Word Shuffling* eingeführt, wodurch die Reihenfolge der Wörter verändert wird.

Randomised Digital Library ist ein Oberbegriff für durch eine Randomisierung veränderte Texte. Jede Technik, die entweder Syntax, Kontext aber auch eine Verteilung, wie die Satzlängenverteilung, modifiziert, kann als eine Randomisierung aufgefasst werden. Zu diesen Techniken gehört das bereits erwähnte *Shuffling*. Dementgegen kann ein *Power-Law*-Generator dazu eingesetzt werden, eine andere Wortverteilung zu generieren, um den Einfluss von sprachlicher Vielfalt zu untersuchen. Weiterhin kann auch eine *Gleichverteilung* der Wörter mit einer Wortfrequenz von $\frac{1}{R_{TTR}}$ (vgl. Abschnitt 3.9) angenommen werden, hierbei bleibt die Anzahl der *Tokens* exakt identisch, jedoch gibt es keine *Stopp-* und seltene Wörter wie sie in natürlicher Sprache zu beobachten sind. Eine *Randomised Digital Library* kann demnach als eine Form der Randomisierung verstanden werden, bei welcher das Schaffen von *Willkür* systematisch vom Forscher bestimmt wird, um eine *Mining*-Methode bzgl. bestimmter Eigenschaften untersuchen zu können. Ganz im Sinne eines Turing-Tests kann eine *Digital Library* auch durch eine *Text Synthese* generiert werden, so dass durch das Verketten von *Ngrams* natürliche Sprache möglichst genau approximiert wird. Die *Mining Ability* ist in einem solchen Fall dementsprechend klein.

Tabelle 4.7 zeigt insgesamt fünf Klassen $T1$ bis $T5$ auf, die im Kontext des Kapiteltitels *Zufall und Struktur* bzw. des Kilgariff's Kapitelleitspruch: *Language is never ever random!* als ein *Turingtest* mit fünf verschiedenen Schwierigkeitsklassen aufgefasst werden können. Während die Klasse $T1$ einfach zu bestimmen ist, da alle Charakteristika einer Sprache nachhaltig durch die *Randomisierung* verändert worden sind, stellen die Klassen $T4$ und $T5$ aufgrund einer künstlichen Sprachproduktion (vgl. [Fucks 1968, Küpfmüller 1974]), der *Text Synthese*, die schwierigste aller Klassen für die *Mining Ability* $\mathcal{L}(\Theta)$ dar.

Die Klasse $T1$ zeichnet sich dadurch aus, dass willkürlich Wörter durch das Aneinanderreihen von Buchstaben generiert werden. Neben dem in Tabelle 4.7 genannten Beispiel kann zu dieser Klasse einer *Randomisierung* auch Newman's *Power-Law*-Generator verstanden werden (vgl. [Newman 2005]). Ausgehend von einer Zufallszahl z mit $z \in [0, 1]$ wird durch *Logarithmic Binning* eine *Power-Law*-Verteilung generiert. Interessanterweise werden nur Zahlen anstelle von Wörtern generiert, wobei 1 am häufigsten generiert wird, so dass die *Zipf*-Verteilung mit kleineren Abweichungen im Exponenten α erhalten bleibt.

Beim *Word Shuffling* aus der Klasse $T2$ geht ein *Randomisierer* Wort für Wort durch eine *Digital Library*. Für jedes *Token* wird innerhalb einer Fenstergröße, wie einem *Satz*, *Absatz*, *Dokument* oder der gesamte *Digital Library*, eine neue Position zufällig bestimmt. Dementsprechend werden das aktuelle *Token* sowie das *Token* an der zufällig ausgewählten Position paarweise ausgetauscht. Auf diese Weise wird im Durchschnitt jedes Wort pro Durchlauf gesichert mindestens einmal und im Durchschnitt zweimal an eine neue Position geschrieben. Die Wahl der Fenstergröße hängt von der Forschungsfrage ab. Während ein kleines Fenster, wie *Satz* und *Absatz*, die quantitative Semantik nach Wittgenstein erhält[14], werden die quantitativen semantischen Zusammenhänge zwischen Wörtern durch die maximale Fenstergröße, der gesamten *Digital Library*, neben der Syntax ebenfalls aufgelöst.

[14] Bedeutung wird nach Wittgenstein als *the meaning of a word is its use in the language* bezeichnet und ist somit eine grundlegende Definition für die quantitative Semantik.

Die $T3$-Klasse[15] aus Tabelle 4.7 kann als eine Form der *Randomisierung* verstanden werden, bei welcher explizit die Satzlängenverteilung verändert wird. Das kann einerseits durch das künstliche Verändern der durchschnittlichen Satzlänge geschehen. Auf der anderen Seite kann auch ein Satz auf eine bestimmte Satzlänge normalisiert werden. Das bedeutet im Detail, dass die Satzlängenverteilung von einer *negativen Binomialverteilung* (vgl. [Kelih 2005]) auf genau eine Satzlänge verändert wird. Insbesondere für die semantischen Verfahren bzw. dem semantischen *Featuring* aus Abschnitt 3.4 kann auf diese Weise ein eingeschränkter Kontext untersucht werden bzw. die Abhängigkeit einer *Text Re-use* Technik auf die Fenstergröße bestimmt werden.

Klasse	Kurzbeschreibung
T5	**Artificial Modification**: Bei dieser Klasse von Randomisierungen werden entweder auf Buchstaben- oder Wortebene künstliche Einfügungen, Ersetzungen oder Löschungen vorgenommen.
T4	**Text Synthese**: *Random Text Generation* oder *Text Synthese* kann durch *Markov Chains* erreicht werden. Ausgehend von einem Anfang wird mit oder ohne *Smoothing* eine Sequenz von Wörter generiert, welche auf Basis von in einer *Digital Library* beobachteten *NGrams* am wahrscheinlichsten erscheint.
T3	**Sentence Length Creator**: Diese Klasse von Verfahren verändert die Satzlänge. Das kann u. a. durch eine konstante Satzlänge für alle Sätze aber auch durch eine Anpassung von Parameter der zugrunde liegenden negativen Binomialverteilung geschehen.
T2	**Word Shuffling**: Diese Klasse von Techniken randomisiert Text so, dass durch paarweises Tauschen der Positionen zweier Wörter, der Text in seiner Syntax und Semantik aufgelöst wird. Jedoch bleiben sowohl die Anzahl der *Tokens* als auch die Verteilung der Wörter exakt identisch.
T1	**Random Word Generator**: Ausgehend von einer Buchstabenverteilung bzw. einem *Power-Law*-Generator wird eine Sequenz von Buchstaben zufällig generiert, bei welcher mit der Wahrscheinlichkeit $p_{whitespace}$ ein Leerzeichen folgt.

Tabelle 4.7: Grobklassifikation von *Randomisierungsmethoden*. Die erste Spalte kann als eine *Turing-Test*-Klassifikation verstanden werden, wobei $T1$ im Sinne eines *Turing-Tests* als einfach zu erkennen und $T5$ als ein schwieriger *Turing-Test* anzusehen ist. Je nach *Turing-Test*-Klassifikation sinkt oder steigt die eingeführte *Mining Ability* $\mathcal{L}(\Theta)$.

Die Klasse $T4$ entspricht einer Menge von synthetischen Textgeneratoren. Hierbei werden innerhalb einer *Digital Library Ngrams* gelernt. Zu einem gegebenen Anfang, wie bspw. einem Satzanfang, wird zu einem festen Gedächtnis bestehend aus $n-1$ zuvorstehenden Wörtern eine Entscheidung über das folgende Wort getroffen. Diese Methode ist von zwei Parametern abhängig. Einerseits ist die Wahl der Größe n eines *Ngrams* und andererseits die Entscheidungsstrategie für die *Noisy Channel Evaluation* von Bedeutung. Aus Tabelle 4.2 auf Seite 134 ist ersichtlich, dass aus rein statistischer Sicht eine *Ngram*-Größe von $n \in [2, 5]$ sinnvoll erscheint. Werden größere *Ngrams* gewählt, so werden in der *Randomised*

[15] Diese Klasse ist nur im Falle einer satz- oder absatzbasierten *Segmentation* (vgl. Abschnitt 3.2) von Interesse. Bei einer Segmentierung durch ein *Moving Window* haben jegliche Veränderungen der Satzlängenverteilung keine Auswirkungen.

Digital Library mehr oder weniger die gleichen Sätze wie in der *Digital Library* generiert, wodurch die *Mining Ability* $\mathcal{L}(\Theta)$ nahe null zu erwarten ist. Bezüglich des Entscheidungskriteriums können zwei Techniken eingesetzt werden, um eine Vorhersage für $P(w_i|w_{i-2}w_{i-1})$ treffen zu können, wobei $w_{i-2}w_{i-1}$ als das Gedächtnis der Markov-Ketten bezeichnet wird. Einerseits ist es möglich, durch das Bestimmen des am wahrscheinlich folgendsten Wortes mit $\arg\max P(w_i|w_{i-2}w_{i-1})$ eine Vorhersage zu treffen. Andererseits sind die so generierten Texte sehr monoton und einfach. Deshalb ist die Entscheidung für das wahrscheinlichste Wort oftmals nicht zielführend. Vielmehr reflektiert die Wahl eines der möglichen vorhersagbaren Wörter als Kandidat auf Basis von Beobachtungen innerhalb der *Digital Library* besser die natürliche Vielfalt. Die Wahl des Kandidaten hängt von der entsprechenden Wahrscheinlichkeit unter dem gegebenen Gedächtnis $P(w_i|w_{i-2}w_{i-1})$ ab. Wahrscheinlichere Vorhersagen werden hierbei öfter und unwahrscheinliche Wörter nur selten ausgewählt. Die daraus resultierenden Texte entsprechen im Vergleich zum *arg max*-Ansatz deutlich besser der natürlichen Sprache und ihrer Vielfalt.

Die Klasse $T5$ zeichnet sich dadurch aus, dass zumindest nicht systematisch die Eigenschaften einer Sprache der zugrunde liegenden *Digital Library* verändert werden. Vielmehr zeichnet diese Klasse aus, dass semantische oder seltenere Varianten eines *Tokens* ersetzt werden. Diese Form einer $T5$-Randomisierung ermöglicht im Kontext des *Historical Text Re-use* zu untersuchen, wie gut ein *Mining*-Verfahren mit historischen Varianten umgehen kann bzw. wie robust es gegen sprachliche Vielfalt ist.

Abschließend sei noch erwähnt, dass für die *Mining Ability* $\mathcal{L}(\Theta)$ neben den in diesem Abschnitt vorgestellten textverändernden Operationen, auch die *Random Graph Models* bzw. *Random Graph Generator* (vgl. [Chakrabarti 2010]) bei *Graph*-Ergebnissen, wie einem *Text Re-use Graph*, für p_{noise} eingesetzt werden können. Etablierte Modelle hierfür sind der Ansatz von Erdős & Renyi (vgl. [Erdős 1959, Erdős 1960]), die Methode von Watts & Strogatz (vgl. [Watts 1998]) oder der Generator von Barabási & Albert (vgl. [Barabási 1999])[16]. Die Wahl des *Graph Models* hängt in erster Linie davon ab, welche Eigenschaften eines Graphen für die Evaluierung wichtig sind und welche nicht. Im Rahmen dieser Arbeit sind die *Graph Models* nur erwähnt und werden nicht weiter berücksichtigt. Das ist in erster Linie darin begründet, dass es sehr aufwendig ist, aus einer gegebenen *Power-Law*-Verteilung eines *Re-use Graphen* einen *Random Graph* zu generieren, so dass zumindest der Exponent α gleich ist. Ist dies nicht gegeben, führt eine Abweichung der *Mining Ability* $\mathcal{L}(\Theta)$ auch auf einen oftmals nur minimalen Unterschied im Exponenten α zurück.

Alle in diesem Abschnitt genannten Randomisierungsmethoden sind in dem *RanCor*-Modul[17] zusammengefasst, welches als eine Erweiterung von *Medusa* angesehen werden kann (vgl. [Büchler 2006b]).

4.6 Eigenschaften der *Noisy Channel Evaluation*

Das ganzheitliche Evaluieren eines *Mining*-Ergebnisses hat sich als schwierig herausgestellt. In Abschnitt 4.2 wurde ein möglicher statistischer Weg aufgezeigt, welcher jedoch aufgrund der vielen seltenen Wörter als statistisch schwierig anzusehen ist. In Abschnitt 4.3 ist die weit verbreitete Evaluierung gegen einen *Gold Standard* mit seinen Vor- und Nachteilen erklärt worden. Insbesondere muss die Aussagekraft eines Ergebnisses durch zu viele mögliche Störimpulse in Frage gestellt werden. In Abschnitt 4.4 wurde daher die *Noisy Channel*

[16] Eine gute Referenz für einen Einstieg in *Graph Models* ist das Buch *Linked* von A. Barabási (vgl. [Barabási 2003]). Auch wenn *Linked* ein populärwissenschaftliches Buch darstellt, so ist es insbesondere für den Einstieg sehr zu empfehlen.

[17] *RanCor* = *Random Corpora*.

Evaluation eingeführt, welche anhand der *Mining Ability* $\mathcal{L}(\Theta)$ misst, wie stark eine *Mining*-Methode zwischen einer natürlichsprachlichen Struktur sowie einem definierten Rauschen (vgl. Abschnitt 4.5) unterscheiden kann.

Die *Noisy Channel Evaluation* aus Abb. 4.3 auf Seite 143 ist eine Methode, die sowohl einfachste Sprachmodelle, wie *Bigram*- oder *Co-occurrence*-basierte Analysen, bis hin zu komplexen Modellen in gleichem Maße evaluieren kann. Eine komplexere Evaluation für den *Historical Text Re-use* auf Basis der *7-Level-Architektur* (vgl. Abschnitt 3.1) folgt in Kapitel 5. Um die grundlegenden Eigenschaften zeigen zu können, werden die Experiment aus Abschnitt 4.2 auf Basis von *Bigrams* und *Co-occurrences* fortgesetzt. Einerseits können *Bigrams* und *Co-occurrences* als der kleinstmögliche bzw. technisch machbare *Text Re-use* verstanden werden. Andererseits sind diese Methoden so einfach, dass nur die *Größe der Digital Library* sowie das *Signifikanzmaß* als relevante Dimensionen anzusehen sind.

Als Randomisierungsmethode wird die $T2$-Methode *Word Shuffling* (vgl. Tabelle 4.7 aus Abschnitt 4.5) mit einer Fenstergröße der gesamten *Digital Library* ausgewählt. Diese Methode ist sowohl einfach, als auch sehr mächtig. Einerseits bleiben grundlegende Eigenschaften, wie die Satz- und Worthäufigkeitsverteilung und damit sowohl die *negative Binomialverteilung* der Satzlängen als auch das *Power Law* des *Zipfschen Gesetzes* sowie real existierende Wörter, erhalten. Bei einer Methode der $T1$-Klasse würde zumindest die letzte Eigenschaft verloren gehen. Andererseits hat die $T2$-Methode des *Word Shuffling* den Vorteil, dass sowohl die Satzsyntax, was relevant für die folgende *Bigram*-Analyse ist, als auch die quantitative Semantik nach Wittgenstein, die *Co-occurrence*-Analyse, aufgelöst werden.

Bevor jedoch die Datenanalyse erfolgt, wird zunächst die Zufälligkeit dieser Methode getestet. Beim *Word Shuffling* werden immer zwei Wörter paarweise vertauscht, wobei ein Wort dem fortlaufenden *Token* entspricht und das zweite *Token* zufällig gewählt wird[18]. Als Zufallsgenerator bzw. Randomisierer wurde hierfür die *Linear Congruential Method* (vgl. [Lehmer 1951]) eingesetzt, welche in Formel 4.32 beschrieben ist (nach [Knuth 1997b], Seite 10, Abschnitt 3.2.1). X_i entspricht der i-ten Zufallszahl beginnend mit einem *Seed* X_0. Sowohl a als auch c werden idealerweise als Primzahlen gewählt.

$$X_{n+1} = (aX_n + c) \mod m \tag{4.32}$$

Beim *Word Shuffling* sind durch die Implementierung in Java[19] sowohl a als auch c gegeben. Als *Seed* wird immer $X_0 = 0$ gewählt, so dass jede Folge von generierten Zufallszahlen stets gleich beginnt. m aus Formel 4.32 entspricht der Anzahl der *Tokens* N, so dass stets genau eine zufällige Position aus $X_i \in [0, N-1]$ bestimmt werden kann. Wie bereits in Abschnitt 4.5 beschrieben, wird bei einem einmaligen Durchlaufen der *Digital Library* jede Position im Durchschnitt zweimal, aber mindestens einmal, paarweise vertauscht.

Die grundlegende Frage ist nun, wie zufällig eine solche Sequenz von Zufallszahlen bzw. wie statistisch zufällig das paarweise Tauschen der Position zweier Wörter ist. Es gibt dementsprechend keinen mathematisch perfekten Zufall, sondern nur eine bestmögliche Approximation bzw. Simulation von Zufall. Aus diesem Grund werden Zufallsgeneratoren durch einen *Entropie-Test* auf ihre Zufälligkeit überprüft (vgl. [L'Ecuyer 1997]). Da ein Randomisierer Zufall simulieren soll, kann bei einem *Entropie-Test* angenommen werden, dass alle möglichen X_i gleich wahrscheinlich und damit gleich verteilt sind. Durch diese Verteilungseigenschaft eignet sich die *Entropie* H (vgl. Formel 4.34) besonders für einen solchen Test, da die *maximale Entropie* H_{max} (vgl. Formel 4.33) bei N verschiedenen Ausgängen genau dann erreicht wird, wenn die Einzelwahrscheinlichkeiten der N Ereignisse

[18] Es wurde als Implementierung für den Zufallsgenerator die Java-Klasse `java.math.Random` eingesetzt.
[19] `java.math.Random`

genau $1/N$ beträgt. Dementsprechend kann ein Zufallsgenerator genau dann als perfekt angenommen werden, wenn jedes $X_i \in [0, N-1]$ mit der gleichen Wahrscheinlichkeit von exakt $1/N$ generiert wird. Die *maximal mögliche Entropie* H_{max} ist somit, wie in Formel 4.33, definiert. Bei einem einmaligen Durchlauf einer *Digital Library*, also Wort für Wort, wird jedes *Token* im Durchschnitt genau zweimal verändert. Die Anzahl aller *Positionsveränderungen* beträgt $2N$, da sowohl die Position des fortlaufenden als auch des *zufällig* ausgewählten *Tokens* pro paarweisem Positionstausch verändert werden. Somit kann für die Berechnung der *maximalen Entropie* nicht nur von einer Gleichverteilung, sondern auch von einer *Wahrscheinlichkeit* $2/2N = 1/N$ ausgegangen werden.

$$H_{max} = -\sum_{i=1}^{N}(1/N)\log_2(1/N) = \log_2(N) \tag{4.33}$$

Auf der anderen Seite kann die real erreichte *Entropie* des Zufallsgenerators, wie in Formel 4.34 beschrieben, durch das Aufsummieren der *Einzelentropien* bestimmt werden.

$$H = -\sum_{i=1}^{N} p_i \log_2(p_i) \tag{4.34}$$

Bei einer *Randomisierung* in nur einem Durchlauf entsteht nun das Problem, dass die Erwartung bei genau 2 Positionswechseln für jedes X_i liegt. Es werden für X_i Positionsverschiebungen von 1 bis 14 beobachtet. Gemäß dem *Gesetz der großen Zahlen* ist eine Erwartung von 2 zu klein, um eine statistische Aussage treffen zu können. Aus diesem Grund kann die Anzahl der Durchläufe n erhöht werden und führt somit zu einer *Entropie* H^n, welche unabhängig von n den Erwartungswert $E(X) = 1/N$ besitzt, jedoch sich die Varianz $V(X)$ durch die Erhöhung der Iterationen sukzessive verringert. Die *maximale Entropie* H_{max} bleibt durch $H_{max} = n/nN = 1/N$ von der Anzahl der Iterationen unberührt.

Der *Entropie-Test* ist die Differenz aus Formel 4.35 zwischen der *maximalen Entropie* und der real gemessenen *Entropie*, welche gemäß der *Entropie* die Einheit *Bit* hat.

$$\Delta H^n = H_{max} - H^n \tag{4.35}$$

Ein *Entropie-Test* ΔH^n kann dann als erfolgreich angesehen werden, wenn $\Delta H^n \leq 1$ Bit gilt. Die Tabelle 4.8 reflektiert die Ergebnisse des *Entropie-Tests* ΔH^n für den $T2$-Randomisierer *Word Shuffling* mit $n \in [1, 10, 100]$ Iterationen sowie in Anlehnung an Abschnitt 4.2 auf die Normgrößen von *100*, *10k*, *1M* sowie *100M* Sätzen.

	100	*10k*	*1M*	*100M*
$n = 1$	$1.904 \cdot 10^{-1}$	$1.678 \cdot 10^{-1}$	$1.693 \cdot 10^{-1}$	$1.692 \cdot 10^{-1}$
$n = 10$	$1.957 \cdot 10^{-2}$	$1.808 \cdot 10^{-2}$	$1.796 \cdot 10^{-2}$	$1.795 \cdot 10^{-2}$
$n = 100$	$1.734 \cdot 10^{-3}$	$1.845 \cdot 10^{-3}$	$1.803 \cdot 10^{-3}$	$1.803 \cdot 10^{-3}$

Tabelle 4.8: *Entropie-Test* ΔH^n für verschiedene Anzahl der Iterationen (Zeilen) sowie auf unterschiedlichen Größen der *Digital Library* (Spalten).

Tabelle 4.8 zeigt zwei Aspekte auf. Erstens, kann spaltenweise festgestellt werden, dass sich durch das Erhöhen der Anzahl von Iterationen n um eine Dezimalstelle der *Entropie-Test* ΔH^n um etwa eine Nachkommastelle verringert. Zweitens, ist bei einer zeilenweisen Betrachtung ebenfalls ersichtlich, dass ab einer Normgröße von *10k* Sätzen der *Entropie-Test* ΔH^n bei gleicher Anzahl von Iterationen n nahezu identische Ergebnisse liefert. Somit

kann der Test ab einer bestimmten Mindestgröße als repräsentativ für alle beliebig größeren *Digital Libraries* angenommen werden. Zusammenfassend ist festzustellen, dass für kleine *Digital Libraries* bzw. im Kontext des *Historical Text Re-use* auch für werkweise Vergleiche (vgl. [Lee 2007]) die Anzahl der Iterationen größer zu wählen ist, so dass nicht nur die Syntax als auch die Quantitative Semantik aufgelöst werden, sondern die statistische Zufälligkeit im größtmöglichen Maße hergestellt wird.

Bei einer Anzahl von $n = 100$ Iterationen kann somit bei einem *Entropie-Test* von $\Delta H^n < 2 \cdot 10^{-3}$ ausgegangen werden. Das entspricht nicht nur deutlich der Testbedingung $\Delta H^n \leq 1$ Bit, sondern garantiert zeitgleich ein Höchstmaß an Präzision, so dass ein statistischer Nebeneffekt bedingt durch den Randomisierer ausgeschlossen werden kann. Da für diese Arbeit ein Höchstmaß an Präzision nötig ist, wurde die Testbedingung von $\Delta H^n \leq 1$ Bit auf $\Delta H^n \leq 10^{-4}$ Bit verschärft, was ab einer Größe von *10k* Sätzen durch $n \in [180, 182]$ Iterationen erreicht wird.

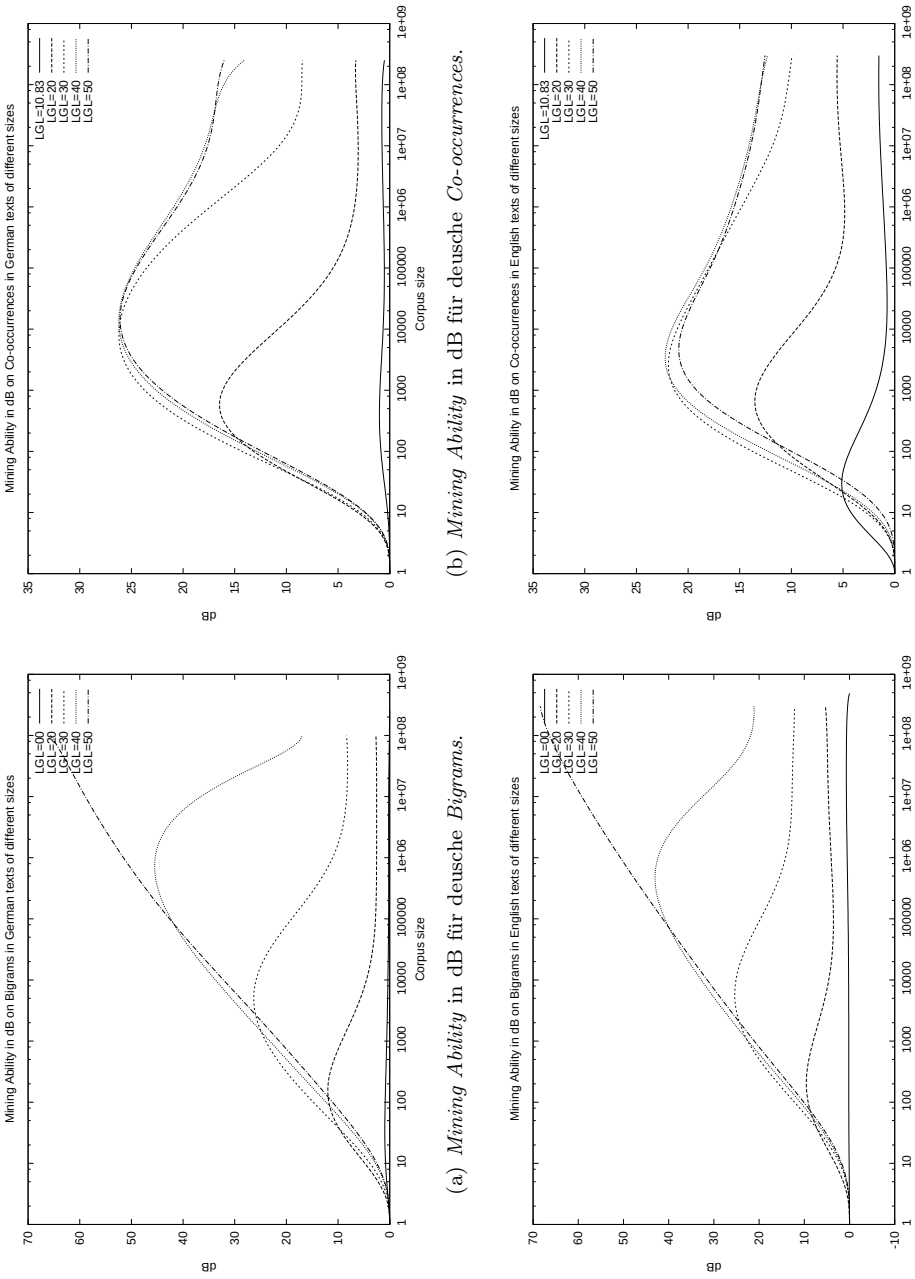
Wie bereits einführend zu diesem Abschnitt dargestellt, sollen aus Gründen der Vereinfachung einige Eigenschaften der *Mining Ability* $\mathcal{L}(\Theta)$ anhand des kleinstmöglichen *Text Re-use*, dem *Bigram* sowie der *Co-occurrence*, aufgezeigt werden. Hierbei wird die *Mining Ability* $\mathcal{L}(\Theta)$ von zwei abhängigen Variablen betrachtet. In Abschnitt 4.6.1 wird das Verhalten der *Mining Ability* bei unterschiedlichen Größen einer *Digital Library* durch verschiedene Normgrößen analysiert (vgl. [Biemann 2007a, Goldhahn 2012]). In Abschnitt 4.6.2 hingegen wird die Größe der *Digital Library* als konstant und nur der Parameterraum Θ, hier das *Log-Likelihood-Ratio* -2λ (vgl. Formel 4.11), als variabel angenommen. In Abschnitt 4.6.3 hingegen werden beide Dimensionen als variabel angenommen und die minimale als auch die maximale *Mining Ability* $\mathcal{L}_{min}(\Theta)$ und $\mathcal{L}_{max}(\Theta)$ betrachtet.

4.6.1 *Mining Ability* in Abhängigkeit von der Größe einer *Digital Library* bei konstantem Parameterraum Θ

In einem ersten Experiment soll die *Mining Ability* $\mathcal{L}(\Theta)$ in Abhängigkeit von der Größe der *Digital Library* durch verschiedene *Normkorpora* untersucht werden (vgl. Abschnitt 4.2). Neben einem deutschen wird auch ein englisches Korpus mit 488 Millionen Sätze auf die Normgrößen[20] *1*, *3*, *10*, *30*, *100*, *300*, *1k*, *3k*, *10k*, *30k*, *100k*, *300k*, *1M*, *3M*, *10M*, *30M*, *100M* sowie *300M* Sätze normalisiert. Für jede dieser *Digital Libraries* mit den genannten Normgrößen wird durch den bereits eingangs genannten $T2$-Randomisier *Word Shuffling* eine *Randomised Digital Library* erstellt. Dies gilt sowohl für die englischen als auch die deutschen *Normgrößen*. Auf Basis der *Normgrößen* wird anschließend für beide Sprachen eine *Bigram*- als auch eine *Co-occurrence*-Analysis durchgeführt, für welche die *Mining Ability* $\mathcal{L}(\Theta)$ bestimmt wird. Die Fragestellung, die mit dieser Analyse einhergeht, ist, ob genau wie beim Menschen eine *Lernkurve* (vgl. [Ebbinghaus 1885]) beobachtet werden kann. Hierzu wird mit zunehmendem Lernaufwand der Lernerfolg zunächst größer, bis er sich auf einem *Plateau* stabilisiert, wodurch die *Lernkurve* auch als S-Kurve bezeichnet wird. Damit einhergehend muss die Frage gestellt werden, ab welcher Normgröße das *Plateau* erreicht wird, das heißt, ab wann der Lernerfolg gemessen am Aufwand in keiner Relation mehr zueinander steht.

Abbildung 4.4 reflektiert die Ergebnisse dieser Analyse für die englisch- als auch deutschsprachigen *Bigram*- und *Co-occurrence*-Berechnungen. In allen vier Teilbildern 4.4(a) bis 4.4(d) sind jeweils fünf Plots für unterschiedliche *Signifikanzschwellwerte* angegeben. Als Signifikanzmaß wird das *Log-Likelihood-Ratio* -2λ (vgl. Formel 4.11) eingesetzt. *LGL* in den Abb. 4.4(a) bis 4.4(d) bedeutet, dass für die *Mining Ability* $\mathcal{L}(\Theta)$ in Abhängigkeit von

[20] k ist die Abkürzung für 1000 und M für eine Millionen.



(a) *Mining Ability* in dB für deusche *Bigrams*.

(b) *Mining Ability* in dB für deusche *Co-occurrences*.

(c) *Mining Ability* in dB für englische *Bigrams*.

(d) *Mining Ability* in dB für englische *Co-occurrences*.

Abbildung 4.4: *Mining Ability* $\mathcal{L}(\Theta)$ in dB für deutsche und englische *Bigrams* und *Co-occurrences* in Abhängigkeit von der Größe der *Digital Library* (Normkorpora der *Leipzig Linguistic Collection* vgl. [Biemann 2007a, Goldhahn 2012]).

der *Normgröße* der *Digital Library* all diejenigen *Bigrams* und *Co-occurrences* ignoriert bleiben, welche den *Schwellwert* nicht überschreiten.

Aus allen vier Teilbildern 4.4(a) bis 4.4(d) kann übereinstimmend festgehalten werden, dass die *Mining Ability* mit zunehmender Normgröße der *Digital Library* bis zu einem Punkt sukzessive ansteigt, sich auf einem *Plateau* stabilisiert und danach wieder abfällt. Des Weiteren steigt $\mathcal{L}_{max}(\Theta)$, wenn auch insbesondere bei den *Co-occurrences* nur marginal, mit zunehmendem Schwellwert LGL. Jedoch verhält sich die *Mining Ability* $\mathcal{L}(\Theta)$ bei einer *Bigram*- und einer *Co-occurrence*-Analyse unterschiedlich. Während bei der *Bigram*-Analyse (vgl. Abbildungen 4.4(a) und 4.4(c)) die maximale *Mining Ability* $\mathcal{L}_{max}(\Theta)$ mit der *Normgröße* der *Digital Library* mitskaliert, ist $\mathcal{L}_{max}(\Theta)$ bei der *Co-occurrence*-Analyse für alle Schwellwerte LGL in den Abbildungen 4.4(b) und 4.4(d) auf einer Normgröße von maximal 10000 Sätzen zu beobachten und kann damit als unabhängig von der Größe der *Digital Library* angesehen werden. Interessanterweise schließen die entsprechenden *Maxima* einen *Normgrößen*-Bereich von $[30, 10000]$ Sätzen ein, was wiederum etwa dem typischen Umfang bzw. der Größe von Dokumenten entspricht (vgl. hierzu auch weiterführend Abschnitt 6.2).

Werden die Analysen der *Bigrams* und *Co-occurrences* direkt verglichen, kann die *Mining Ability* für *Bigrams* sowohl auf deutschen als auch englischen Texten im Maximum fast 70 dB bei einem allerdings auch sehr hohen Schwellwert von $-2\lambda = 50$ erreicht werden. Die maximale *Mining Ability* $\mathcal{L}_{max}(\Theta)$ bei einer zugrunde liegenden *Co-occurrence*-Analyse erreicht auf deutschen Texten jedoch nur etwa 25 dB und auf englischen Texten gar nur knapp 21 dB. Der Unterschied von 4 dB zwischen deutschen und englischen Texten kann durch die sprachliche Vielfalt begründet werden. Beide Texte sind nicht lemmatisiert, wodurch die sprachliche Vielfalt der deutschen Texte aufgrund eines ausgeprägteren Kasus sowie einer umfangreicheren Morphologie größer ist. Wörter treten jedoch nicht als Konzepte, wie in einem Schlagwortverzeichnis oder inhaltlichen *Tags*, in einem Text auf, sondern in deren morphologischen Variante. Dies bedeutet, dass in einer Sprache mit wenig Vielfalt, wie dem Englischen, die Wortformen sehr häufig sind, was einerseits der Sprachstatistik zugutekommt, aber zeitgleich auch die Wortformen in einer *Randomised Digital Library* durch deren größere Wortwahrscheinlichkeit häufiger zusammen auftreten, wodurch auch die Menge der *Co-occurrences* stärker mit ansteigt und die *Mining Ability* $\mathcal{L}_{max}(\Theta)$ dementsprechend 4 dB kleiner ist. Dieser Effekt ist für den *Historical Text Re-use* insofern von großer Bedeutung, als dass ein nicht notwendiges *Preprocessing*, wie die *Lemmatisation*, mitunter das Ergebnis nicht nur verbessern, sondern gar verschlechtern kann. Auch wenn ganz klar festgehalten werden kann, dass ein aktiveres *Preprocessing* nicht notwendiger Weise schlecht sein muss, so steht mit der *Mining Ability* ein Instrument zur Verfügung, etwaige Vorverarbeitungschritte zu testen und deren Einzelnutzen zum *Mining*-Ergebnis zu evaluieren.

Eingangs wurde die Frage nach der Möglichkeit einer Reproduktion einer Lernkurve (vgl. [Ebbinghaus 1885]) durch die *Mining Ability* $\mathcal{L}(\Theta)$ mit zunehmender Größe der *Digital Library* gestellt. Hierbei repräsentiert das Anwachsen der *Digital Library* den ansteigenden Lernaufwand und die *Mining Ability* kann als der Lernerfolg aufgefasst werden. Während der Lernerfolg, gemessen durch die *Mining Ability*, zunächst ansteigt, fällt er nach einem globalen *Maximum* wieder ab. In der Interpretation bedeutet ein Abfallen der *Mining Ability*, dass die Menge der *Bigrams* und *Co-occurrences* auf der *Randomised Digital Library* bei zunehmender *Normgröße* schneller steigt als auf der natürlichsprachlichen *Digital Library*. Dieser Effekt kann in Anlehnung an das in Abschnitt 4.2 genannte und durch das *Zipfsche Gesetz* induzierte statistische Problem begründet werden. Die meisten *Bigrams* und *Co-occurrences* sind, wie in Abschnitt 4.2, dargestellt, selbst beim einmaligen Beobachten innerhalb einer *Digital Library* bereits statistisch signifikant. Durch das Vergrößern der *Digital Library* nimmt nicht nur die Menge dieser Ereignisse zu, sondern sie werden selbst auf einer *Randomised Digital Library* sukzessive häufiger als einmal beobachtet, so dass die

Signifikanz korrelierend mit ansteigt. Selbst bei einem steigenden Signifikanzschwellwert, als in der wissenschaftlichen Community bereits unüblichen Werten von $-2\lambda \geq 10.83$, ist dieses Phänomen in den Abbildungen 4.4(a) bis 4.4(d) feststellbar. Die Interpretation dieses Phänomens bedeutet demnach, dass mit zunehmender Größe der *Digital Library*, die Fehler zunehmen und dementsprechend auch aus Fehlern gelernt wird. Während ein Erhöhen des Schwellwertes im Sinne einer größeren Sicherheit bei einer *Bigram*-Analyse (vgl. Abbildungen 4.4(a) und 4.4(c)) auch größer werdende *Digital Libraries* rechtfertigt, kann dieser Effekt bei der *Co-occurrence*-Analyse nicht festgestellt werden. Vielmehr macht diese *quantitative Evaluierung* nicht deutlich, dass die reale *Mining Ability* bei gleichem Verhalten noch wesentlich kleiner ist. Dies sei an einem einfachen Beispiel dargestellt: *Die schwarze Katze jagt die kleine Maus*. Die Assoziation zwischen *Katze* und *Maus* ist sowohl statistisch signifikant als auch relevant. Ebenso beschreibt *schwarze* ein typisches Attribut von *Katze* sowie *kleine* ein typische Eigenschaft von *Maus*. All diese drei *Co-occurrences* werden innerhalb einer passenden *Digital Library* hinreichend oft beobachtet. Zwangsläufig kommt es zu einer *statistischen Interferenz*, wodurch auch *schwarze* und *kleine* als zumindest schwach signifikant auch ohne jedwede Probleme durch das *Power Law* der Wortverteilungen gemessen werden kann. Insbesondere bei größeren Fenstern als nur einem Satz tritt dieser Effekt verstärkt auf. Auch wenn für diese Arbeit nicht weiter relevant, sei dennoch kurz darauf verwiesen, dass aus diesem Grund seit einigen Jahren im Forschungsbereich der quantitativen Semantik syntaktische *Parse-Bäume*, wie die aus einer *Treebank* (vgl. [Bamman 2008]), dazu eingesetzt werden, Abhängigkeiten zwischen Wörtern zu bestimmen. Auf Basis des Abstandes im *Dependency Graph* wird berechnet, ob eine linguistische Abhängigkeit besteht, so dass eine *Co-occurrence* bei einem zu großen Abstand abgelehnt wird. In Abschnitt 3.4 wurde bereits darauf hingewiesen, dass diese syntaktischen Relationen zukünftig eine stärkere Rolle in der *Text Re-use* Forschung spielen werden. Jedoch ist die Qualität der *Parse*- bzw. *Dependency*-Graphen noch zu schlecht. Pragmatischer wird in diesem Zusammenhang in [Büchler 2006a] beschrieben, dass durch *Distance based Bigrams* auf *PoS*-getaggten Texten solche Strukturen im Sinne eines flachen *Parsings* dazu eingesetzt werden können, um syntaktische Abhängigkeiten zwischen verschiedenen *PoS*-Tags zu bestimmen. Wird nun wie im obigen Beispiel ein *Co-occurrence Candidate* wie *schwarze* und *kleine* im Text beobachtet, kann anhand der *PoS*-Tags festgestellt werden, dass ein Wortabstand von 4 zwischen zwei Adjektiven nicht als signifikant anzusehen ist. Einerseits ist der Abstand zu groß. Zweitens sind signifikante Abstände (1 und 3) ungerade, was dadurch bedingt ist, dass die Adjektive zu einem Substantiv entweder durch ein Komma oder eine Konjunktion, wie *und* bzw. *oder*, getrennt sind. Dadurch wird nicht nur die Qualität innerhalb der *Digital Library* verbessert, sondern es werden auch die meisten der sehr seltenen *Co-occurrences* in der *Randomised Digital Library* als irrelevant herausgefiltert, wodurch sich die *Mining Ability* deutlich verbessert.

Weiterhin zeigen alle vier Teilbilder 4.4(a) bis 4.4(d) auf, dass insbesondere bei niedrigeren Schwellwerten zwischen einer *Normgröße* von *10k* und *100k* ein *lokales Minimum* systematisch beobachtet werden kann. Bei größer werdenden *Digital Libraries* steigt die *Mining Ability* wieder um 1 bis 2 dB an. Aufgrund dessen, dass sich dieser Effekt nur sehr schwach darstellt und sich auf knapp 10 *Normgrößen* verteilt, ist es nicht möglich gewesen, ihn sachlich zu begründen. Vielmehr wäre jede Interpretation reine Spekulation. Deshalb sei dieser Effekt des *temporären lokalen Minimums* an dieser Stelle nur erwähnt.

Zusammenfassend kann festgehalten werden, dass das in Abschnitt 4.2 aufgezeigte statistische Problem der Sprachstatistik grundlegende Auswirkungen auf *Mining*-Analysen hat. Da die *Mining Ability* $\mathcal{L}(\Theta)$ bzgl. einer abhängigen Variablen, wie der Normgröße, das Verhältnis des Zuwachses der gemessenen Daten sowohl innerhalb einer *Digital Library* als auch einer *Randomised Digital Library* bestimmt, kann das Abfallen der *Mining Ability*

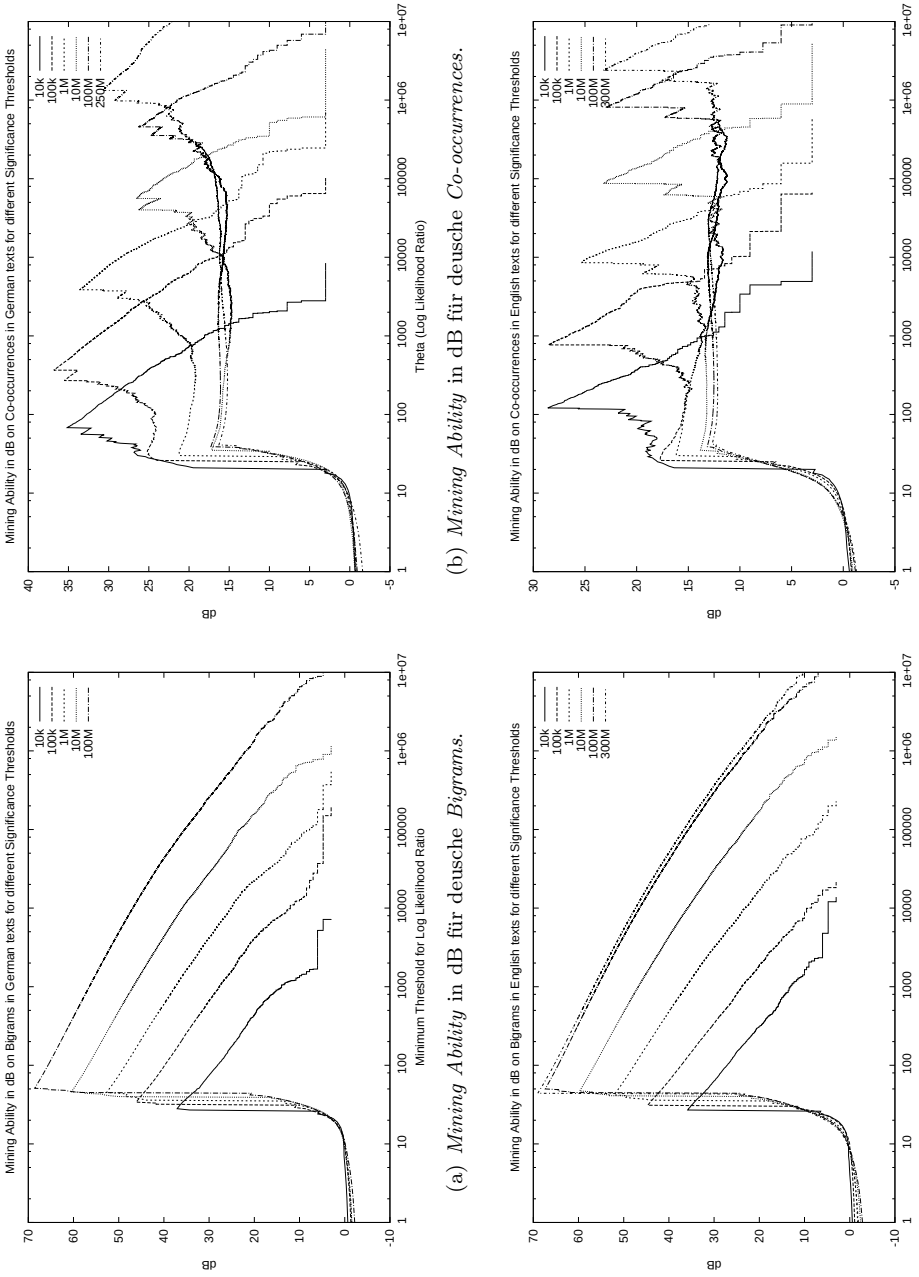
nach dem *globalen Maximum* so gedeutet werden, dass die vermeintlich gefundenen und bzgl. eines statistischen Schwellwertes signifikanten *Co-occurrences* und *Bigrams* innerhalb der *Randomised Digital Library* stärker anwachsen als im natürlichsprachlichen Textbestand. Auch wenn dieser Effekt definitiv nicht gewollt ist, so ist er dennoch beobachtbar. Weiterhin wurden mit den syntaktischen *Parse*-Bäumen bzw. den *Distance based Bigrams* zwei Lösungen vorgeschlagen, welche die *Mining Ability* $\mathcal{L}_{max}(\Theta)$ deutlich verbessern. Ziel muss es sein, die *Mining Ability* als ein Instrument zu verstehen, welche das systematische Rauschen einer *Mining*-Methode bestimmt. Liefert ein Verfahren eine zu geringe *Mining Ability* oder verursacht, wie in diesem Abschnitt dargestellt, ein negatives Verhalten bei zu nehmender Größe der *Digital Library*, so ist das Risiko gegeben, dass aus Fehlern bzw. dem systematischen Rauschen der Methode in komplexeren Verfahren, welche auf diese Technik aufsetzen, gelernt und somit die Gesamtqualität deutlich reduziert wird.

Auch wenn die in diesem Abschnitt dargestellten Ergebnisse nicht den natürlichen Erwartungen entsprechen, so kann jedoch die Mächtigkeit dieser neuen und rein quantitativen Methode durch eine ideale bzw. hypothetische Analyse verdeutlicht werden. In Anlehnung an die Lernkurve (vgl. [Ebbinghaus 1885]) ist im Idealfall ein Anstieg der *Mining Ability* mit zunehmender Größe der *Digital Library* zu erwarten. Da kein Verfahren fehlerfrei ist, kann es als Idealfall angesehen werden, wenn eine *Mining*-Methode bei zunehmender Größe der *Digital Library* entweder ein konstantes oder ein leicht sinkendes *systematisches Grundrauschen* auf einer *Randomised Digital Library* besitzt. Auf der anderen Seite jedoch wäre zu erwarten, dass das gemessene Signal bzw. die extrahierten Strukturen in der natürlichsprachlichen *Digital Library* bei zunehmender Größe systematisch degressiv ähnlich der Logarithmus-Funktion wachsen. Daraus wäre im skizzierten Idealfall zu erwarten, dass die maschinelle Lernkurve gemessen durch die *Mining Ability* der menschlichen Lernkurve sehr ähnlich ist und zu einem *Plateau* des Lernens führt. Da in dem Experiment zu den Abbildungen 4.4(a) bis 4.4(d) davon ausgegangen werden kann, dass mit zunehmender Größe der *Digital Library* auch neue Informationen gemäß dieser Erwartung aufgedeckt werden können, ist nahe liegend, dass sowohl die *Co-occurrence*- als auch die *Bigram*-Analyse durch das in Abschnitt 4.2 aufgezeigte statistische Problem der Sprachstatistik bezüglich der wachsenden Datenmenge kein konstantes systematisches Rauschen haben kann, sondern das Rauschen sogar schneller wächst als die neu gefundenen Strukturen.

4.6.2 *Mining Ability* in Abhängigkeit vom Parameterraum Θ bei konstanter Größe einer *Digital Library*

In einer zweiten Analyse werden die gleichen Daten wie aus dem vorhergehenden Abschnitt dazu verwendet, die *Mining Ability* $\mathcal{L}(\Theta)$ bezogen auf ihr Verhalten in Abhängigkeit von einem Schwellwert zu untersuchen. Hierzu wird das Ergebnis einer *Co-occurrence*- als auch einer *Bigram*-Analyse auf englischen sowie deutschen Normgrößen einer *Digital Library* und der daraus resultierenden *Randomised Digital Library* so verglichen, dass immer nur diejenigen Datensätze für die Bestimmung der *Mining Ability* als relevant erachtet werden, welche auch einen bestimmten Schwellwert t zum *Log-Likelihood-Ratio* -2λ erfüllen. Die Abbildungen 4.5(a) bis 4.5(d) reflektieren die Ergebnisse dieser Analyse für einige ausgewählte Normgrößen ([Biemann 2007a, Goldhahn 2012]).

Die natürliche Erwartung der *Mining Ability* $\mathcal{L}(\Theta)$ in Abhängigkeit von einem Schwellwert t ist, dass mit zunehmenden t die *Mining Ability* ansteigt. Es ist offensichtlich, dass für das *Log Likelihood Ratio* -2λ innerhalb einer *Digital Library* deutlich größere *Signifikanzen* als in einer *Randomised Digital Library* berechnet werden können. Aus diesem Grund gilt bei einem höheren Schwellwert schneller $|E_{D_S^m,\phi_\Theta}| = 0$ als bei $|E_{D_S,\phi_\Theta}|$ auf der natürlichsprachlichen *Digital Library* (vgl. Formel 4.22 auf Seite 144). Deshalb ist zu er-



(a) *Mining Ability* in dB für deusche *Bigrams*.

(b) *Mining Ability* in dB für deusche *Co-occurrences*.

(c) *Mining Ability* in dB für englische *Bigrams*.

(d) *Mining Ability* in dB für englische *Co-occurrences*.

Abbildung 4.5: *Mining Ability* $\mathcal{L}(\Theta)$ in dB für deutsche und englische *Bigrams* und *Co-occurrences* in Abhängigkeit vom Schwellwert des *Log-Likelihood-Ratio* -2λ.

warten, dass nach einem Maximum und spätestens der Bedingung $|E_{D_S^m,\phi_\Theta}| = 0$ gilt, dass bei einem weiteren Erhöhen des Schwellwertes die *Mining Ability* nun wieder sinken muss, da auch $|E_{D_S,\phi_\Theta}|$ sukzessive kleiner wird. Genau dieses erwartbare Verhalten kann in allen vier Plots aus Abb. 4.5 beobachtet werden, auch wenn sich, wie im vorigen Abschnitt, das prinzipielle Verhalten zwischen der *Bigram*- und der *Co-occurrence*-Analyse unterscheidet. Des Weiteren können fünf weitere Auffälligkeiten den Abbildungen 4.4(a) bis 4.4(d) sowohl sprach- als auch methodenunabhängig entnommen werden.

Erstens, alle vier Abbildungen 4.4(a) bis 4.4(d) zeigen auf, dass für jeden Schwellwert $t \leq 16$ gilt, dass $\mathcal{L}(\Theta) < 0$ ist. Im Detail bedeutet dies, dass sowohl für deutsche als auch englische *Bigrams* und *Co-occurrence* für niedrige Schwellwerte eine negative *Mining Ability* beobachtbar ist. Dies bedeutet letztlich, dass bei geringen *Schwellwerten* mehr Daten zum Schwellwert t auf der *Randomised Digital Library*, dem *Quantitative Noise*, extrahiert werden als Strukturen aus der natürlichsprachlichen *Digital Library*. Einerseits ist es ein natürliches und auch erwartbares Verhalten, da sich innerhalb einer *Digital Library* Strukturen in der Tiefe herausbilden, während auf einer *Randomised Digital Library* vergleichbarer Größe mehr Daten in der Breite gebildet werden. Bei einem kleinen Schwellwert ist die Selektion zu gering, wodurch sich dieses Verhalten in einer negativen *Mining Ability* auswirkt. Ernst zu nehmen ist jedoch der Fakt, dass ein *Score* für das *Log Likelihood Ratio* von $-2\lambda = 16$ bei einem Freiheitsgrad bereits einem statistischen Fehler von weniger als 10^{-4} entspricht (vgl. Tabelle A.1 auf Seite 234). Das ist insofern unerwartet, als dass in Fachpublikationen meist ein deutlich niedrigerer Schwellwert von 6.63 (vgl. u. a. [Biemann 2007b, Bordag 2007]) als guter Schwellwert angenommen wird, welcher einem statistischen α-Fehler von 0.01 entspricht[21]. Gemessen durch die *Mining Ability* $\mathcal{L}(\Theta)$ kann jedoch gezeigt werden, dass für einen Schwellwert t deutlich unter 20 faktisch kein Unterschied zwischen den zu einem Schwellwert t extrahierten Daten einer natürlichsprachlichen *Digital Library* sowie einer *Randomised Digital Library* gemacht werden kann. Auch wenn die mathematischen Modelle zweifelsfrei richtig sind, wird an dieser Stelle einmal mehr die Auswirkung des statistischen Problems bedingt durch die *Power-Law*-Verteilung von natürlichsprachlichen Wörtern deutlich (vgl. Abschnitt 4.2).

Zweitens, alle vier Abbildungen 4.4(a) bis 4.4(d) zeigen ebenfalls auf, dass unabhängig von der Größe der *Digital Library*[22] für $20 \leq t \leq 30$ ein signifikanter Anstieg der *Mining Ability* $\mathcal{L}(\Theta)$ zu verzeichnen ist. Entgegen einem mathematisch motivierten Schwellwertes von 6.63 bezogen auf die statistische Einzelentscheidung (vgl. Abschnitt 4.2) muss in der Gesamtevaluierung durch die *Mining Ability* $\mathcal{L}(\Theta)$ davon ausgegangen werden, dass ein Hypothesentest, wie es das *Log Likelihood Ratio* -2λ ist, erst ab einem drei- bis vierfachen Schwellwert t einen ganzheitlichen Unterschied zwischen Struktur und Zufall machen kann.

Drittens, wenn auch aus den Abbildungen 4.4(a) bis 4.4(d) nicht direkt ersichtlich, so kann auf Basis der Daten ebenfalls festgehalten werden, dass $\mathcal{L}_{max}(\Theta)$ immer genau dann erreicht wird, wenn in Formel 4.22 nicht mehr $|E_{D_S^m,\phi_\Theta}| = \max(1, |E_{D_S^m,\phi_\Theta}|)$ gilt, sondern $1 = \max(1, |E_{D_S^m,\phi_\Theta}|)$.

Viertens, das Verhalten von $\mathcal{L}_{max}(\Theta)$ bzgl. unterschiedlicher Normgrößen aus Abb. 4.5 unterscheidet sich insofern zwischen *Bigrams* und *Co-occurrences*, als dass alle *Maxima* der verschiedenen *Bigram*-Analysen (vgl. Abb. 4.5(a) und 4.5(c)) sich in einem *Schwellwert*-Bereich $20 \leq t \leq 30$ wiederfinden. Bei der *Co-occurrence*-Analyse hingegen (vgl. Abb. 4.5(b) und 4.5(d)) skaliert der Schwellwert, bei welchem $\mathcal{L}_{max}(\Theta)$ erreicht wird, mit der Normgröße der *Digital Library*.

Fünftens, genau wie im vorigen Abschnitt 4.6.1 ist zwischen dem ersten lokalen Ma-

[21] Es gibt kein wirklich gutes Kriterium dafür, wie groß ein α-Fehler sein darf. Es wird jedoch oftmals auf die α-Fehler von 0.05 ($-2\lambda = 3.84$) und 0.01 ($-2\lambda = 6.63$) zurückgegriffen.

[22] Jede der genannten Abbildungen enthält mehrere Plots für unterschiedliche Normgrößen.

ximum bei $20 \leq t \leq 30$ und dem globalen Maximum der *Mining Ability* $\mathcal{L}_{max}(\Theta)$ ein schwach ausgeprägtes *lokales Minimum* bei sowohl den deutschen als auch den englischen *Co-occurrences* beobachtbar. Besonders gut mit etwas mehr als 2 dB Unterschied kann dies in den Abbildungen 4.5(b) und 4.5(d) bei der Normgröße von $1M$ Sätzen aufgezeigt werden. Auch hier gilt, dass das *lokale Minimum* zu schwach ausgeprägt ist, so dass letztlich nur über deren Ursachen spekuliert werden kann. Aufgrund der Charakteristik in allen Plots für *Co-occurrences* sei zumindest auf diesen Effekt ebenfalls hingewiesen.

In diesem Abschnitt wurde gezeigt, dass das grundsätzlich erwartbare Verhalten der *Mining Ability* $\mathcal{L}(\Theta)$ auch beobachtet werden kann. Was ist jedoch ein guter Schwellwert? Dies bleibt insbesondere bei mitskalierenden $\mathcal{L}_{max}(\Theta)$, wie es bei den *Co-occurrences* der Fall ist, kein triviales Problem. Es kann in jedem Fall festgestellt werden, dass es keinen statischen, sondern einen von der Größe der *Digital Library* abhängigen Schwellwert geben muss. Für die Bestimmung eines geeigneten Schwellwertes t können jedoch Abbildungen, wie 4.4(a) bis 4.4(d), dabei helfen, diese Entscheidung in Abhängigkeit von den Eigenschaften einer *Digital Library* zu treffen.

4.6.3 Minimale und maximale *Mining Ability* $\mathcal{L}_{min}(\Theta)$ und $\mathcal{L}_{max}(\Theta)$ bei einem dynamischen Parameterraum Θ sowie unterschiedlich großen *Digital Libraries*

In den beiden Analysen aus den Abschnitten 4.6.1 und 4.6.2 wurde entweder die Größe der *Digital Library* oder der Schwellwert t des *Log Likelihood Ratio* -2λ als konstant angenommen. In diesem Abschnitt soll das Verhalten untersucht werden, wenn keiner der Parameter als konstant angenommen wird. In Anbetracht von $\Delta(\Theta)$ aus Formel 4.28 wird die Dimension des Schwellwertes auf $\mathcal{L}_{min}(\Theta)$ sowie $\mathcal{L}_{max}(\Theta)$ reduziert. Die vier Abbildungen 4.6(a) bis 4.6(d) reflektieren diese Analysen.

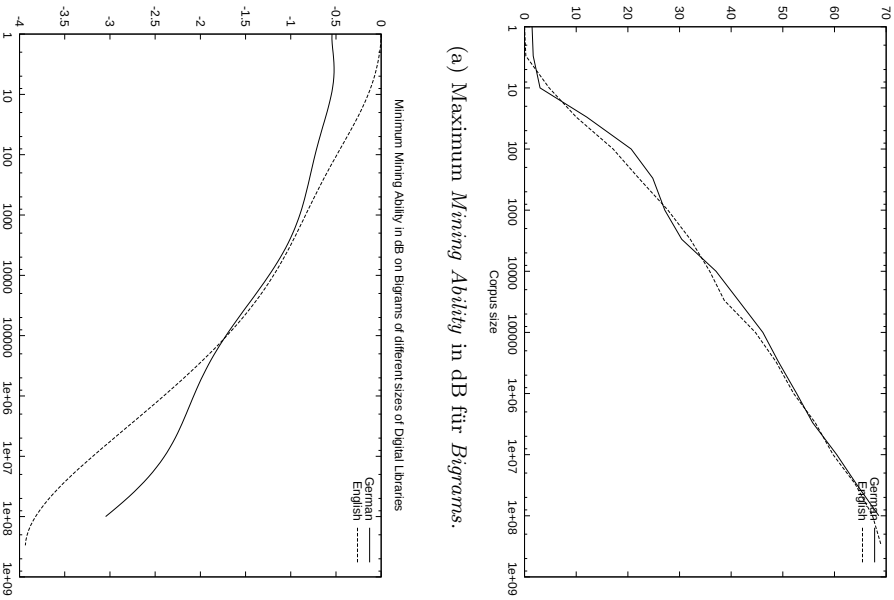
Erstens, es kann aus den Abbildungen 4.6(c) und 4.6(d) entnommen werden, dass sowohl für *Bigrams* und für *Co-occurrences* auf deutschen als auch englischen Texten die minimale *Mining Ability* $\mathcal{L}_{min}(\Theta)$ mit zunehmender Größe der *Digital Library* fällt.

Zweitens, sowohl auf deutschen als auch englischen Texten kann beobachtet werden, dass die maximale *Mining Ability* $\mathcal{L}_{max}(\Theta)$ für *Bigrams* sukzessive mit der Größe der *Digital Library* mitskaliert (vgl. Abb. 4.6(a)). Dies bedeutet, dass bei größer werdender *Digital Library* ebenfalls $\mathcal{L}_{max}(\Theta)$ linear mitsteigt.

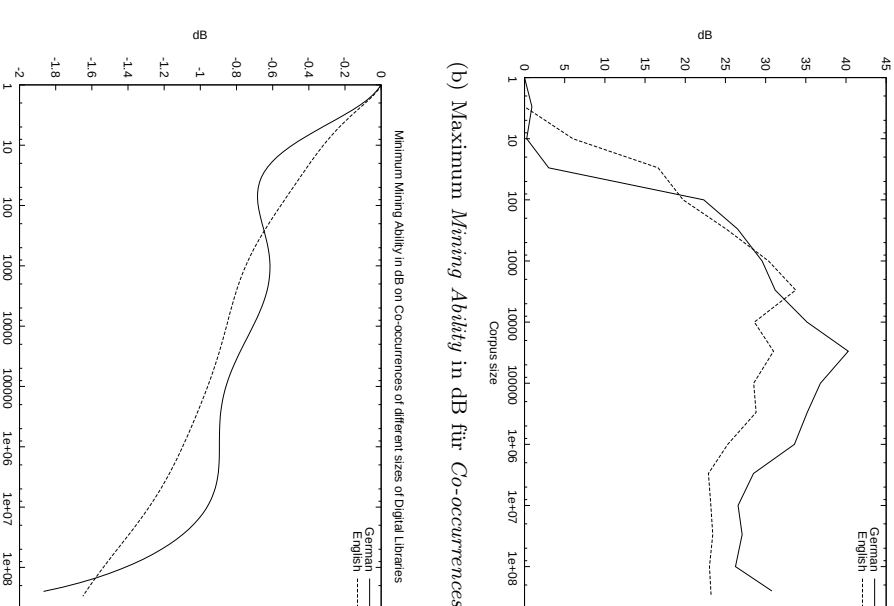
Drittens, die maximale *Mining Ability* $\mathcal{L}_{max}(\Theta)$ für *Co-occurrences* erreicht ein *globales Maximum* für eine Normgröße der *Digital Library* zwischen *3k* und *30k*. Danach fällt $\mathcal{L}_{max}(\Theta)$ um mehr als 10 dB ab und stabilisiert sich auf einem *Plateau*.

Viertens, sowohl für *Bigrams* als auch für *Co-occurrences* steigt $\Delta(\Theta)$ aus Formel 4.28 mit zunehmender Größe der *Digital Library* ebenfalls. Einerseits könnte das als positiv aufgefasst werden, da eine Methode in Abhängigkeit von der Größe der *Digital Library* einen steigenden Leistungsbereich hat. Jedoch bedeutet dieser wachsende Leistungsbereich auch, dass durch einen Schwellwert t die Chance steigt, eine schlechte bzw. suboptimale Wahl bzgl. der Parameter des Modells zu treffen. Wäre $\Delta(\Theta)$ hingegen klein, so ist das Risiko, vom Optimum abzuweichen, deutlich geringer. Insofern muss das von der Größe der *Digital Library* abhängige und mitsteigende $\Delta(\Theta)$ eher als nachteilig sowohl bei der *Bigram*- als auch der *Co-occurrence*-Analyse betrachtet werden.

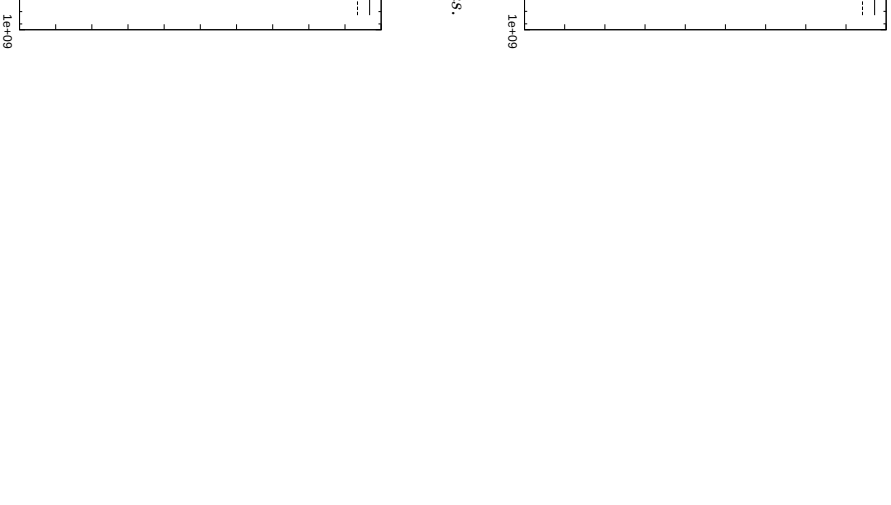
Fünftens, es können keine nennenswerten Unterschiede zwischen dem Deutschen und dem Englischen für $\mathcal{L}_{min}(\Theta)$ ausgemacht werden (vgl. Abbildungen 4.6(c) und 4.6(d)). Das Verhalten zwischen beiden Sprachen kann als nahezu identisch angesehen werden. Die gleiche Aussage kann auch für $\mathcal{L}_{max}(\Theta)$ bei deutschen und englischen *Bigrams* getroffen werden (vgl. Abb.4.6(a)). Für $\mathcal{L}_{max}(\Theta)$ auf *Co-occurrences* können jedoch signifikante Unterschiede



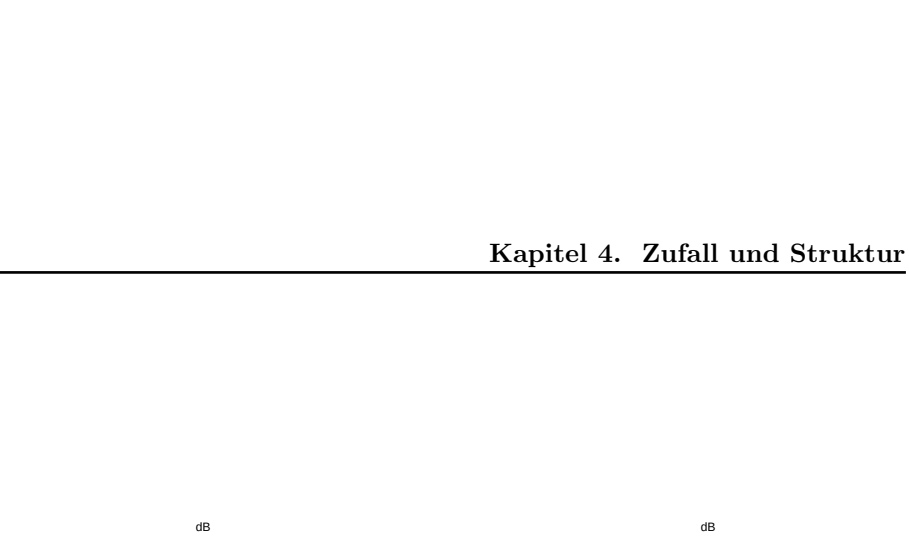
(a) Maximum *Mining Ability* in dB für *Bigrams*.



(b) Maximum *Mining Ability* in dB für *Co-occurrences*.



(c) Minimum *Mining Ability* in dB für *Bigrams*.



(d) Minimum *Mining Ability* in dB für *Co-occurrences*.

Abbildung 4.6: *Minimum* und *Maximum Mining Ability* $\mathcal{L}_{min}(\Theta)$ und $\mathcal{L}_{max}(\Theta)$ für deutsche sowie englische *Bigrams* und *Co-occurrences*.

in Abb. 4.6(b) ausgemacht werden. Auch wenn das grundlegende Verhalten zwar identisch ist, so ist ab einer *Normgröße* von etwa *10k* Sätzen eine Unterschied von 7 - 8 dB feststellbar, was wiederum bezogen auf die Datenmenge einen signifikanten Unterschied impliziert.

4.7 Einbettung dieses Kapitels in die gesamte Arbeit

Language is never ever random ist nicht nur Kilgariff's Maxime für dieses Kapitel, sondern ist auch Vorsatz für die *Noisy Channel Evaluation* mit dem in diesem Kapitel eingeführten *Score* der *Mining Ability* $\mathcal{L}(\Theta)$. Eine Sprache besteht aus verschiedenen Formen von Strukturen, die sowohl morphologischer, syntaktische oder auch semantischer Art sein können. Durch ein gezieltes Auflösen dieser Strukturen hin zu einer *Randomised Digital Library* kann eine neuartige Form der Evaluierung geschaffen werden, welche die *Mining*-Stärke eines Verfahrens bzgl. dieser Strukturen messen kann.

Wichtig ist hierbei, die Grenzen der *Noisy Channel Evaluation* präzise zu verstehen. Die *Noisy Channel Evaluation* kann im Kontext der *eHumanities* niemals eine qualitative Evaluierung ersetzen. Die *Noisy Channel Evaluation* ist in der Lage, das strukturelle Grundrauschen eines *Mining*-Verfahrens gemäß der genannten Maxime zu messen. Wie in diesem Kapitel in verschiedenen Abschnitten gezeigt wurde, kann die *Mining Ability* dabei helfen, das Verhalten einer *Mining*-Methode bezogen auf Skalierbarkeit oder bestimmte Parameter zu verstehen. Die *Mining Ability* als *Score* kann jedoch nicht über ein intellektuelles Rauschen entscheiden. So kann die *Noisy Channel Evaluation* angewandt auf den *Historical Text Re-use* entscheiden, dass eine Phrase, wie *im Namen unseres Herren Jesus Christus*, als Struktur erkannt wird. Jedoch ist sie nicht in der Lage, eine qualitative Entscheidung darüber zu treffen, dass an dieser Form von *Text Re-use* in den meisten Fällen kein größeres fachwissenschaftliches Interesse besteht. Vielmehr kann eine solche Phrase oftmals als geisteswissenschaftlicher Spam verstanden werden. Weiterhin kann zwar durch das gezielte Auflösen von Strukturen, die Fähigkeit der *Mining*-Methode gemessen werden, jedoch hat sie signifikante Schwächen, sobald Kognition beim Textverständnis nötig ist. Die *Mining Ability* $\mathcal{L}(\Theta)$ wird bspw. nur sehr schwer die Fähigkeit abbilden können, ob eine Methode die *Re-use Units like will to like* und *Birds of same feather flock together* aufeinander linken kann.

Dennoch ist die *Noisy Channel Evaluation* mit dem *Score* der *Mining Ability* eine Bereicherung bei der Evaluierung von *Mining*-Ergebnisse natürlicher Sprache. Entgegen einer Evaluierung gegen einen *Gold Standard* benötigt die *Noisy Channel Evaluation* keine Testgrundlage außer dem Text selber. Eines der Grundprobleme der Evaluierung gegen einen *Gold Standard* ist, dass oftmals Unklarheit besteht, wie gut eine Testbasis auf die untersuchte *Digital Library* passt. Eine niedrige *Precision*, *Recall* oder *F-Measure* kann vieles bedeuten. Einerseits kann es ausdrücken, dass eine *Mining*-Methode schlecht funktioniert, was defacto in den meisten Fällen angenommen wird. Andererseits kann ein niedriger *Score* auch bedeuten, dass es nur eine geringe inhaltliche Überlappung zwischen *Gold Standard* und *Digital Library* gibt, wodurch eine ernsthafte Bewertung einer Evaluierung gegen einen *Gold Standard* oftmals deutlich erschwert und nicht nur durch die Zahlen selber ausgedrückt werden kann.

Die *Noisy Channel Evaluation* hingegen misst die Fähigkeit, aus einer *Digital Library* bestimmte Strukturen zu extrahieren. Es kann in diesem Kontext die *Randomised Digital Library* als *Gold Standard* verstanden werden, wobei es der wissenschaftlichen Methodik des Forscher obliegt, welche der fünf Klassen *T1* bis *T5* sich für die jeweiligen Experimente am besten eignet. Je höher die Klasse ist, desto geringer fällt die erwartbare *Mining Ability* aus.

Ein weiterer fundamentaler Vorteil wurde in Abschnitt 4.4 aufgezeigt. Evaluierungen gemessen durch klassische Maße wie *Precision*, *Recall* und *F-Measure* beschränken einen Test auf automatische *Mining*-Methoden. Ein manuelles *Mining* durch ein *Philological Crowd Sourcing* wird niemals an die Quantität des *Recalls* einer automatischen Methode herankommen. In Abschnitt 4.3 wurde hierzu aufgezeigt, dass faktisch kein *Crowd Sourcing System* ein höheres *F-Measure* erreichen kann, als selbst die pragmatischsten Algorithmen. Inspiriert durch die Geisteswissenschaften, die ein hohes Maß an *Precision* durch ein manuelles *Mining* erreichen, wurde in diesem Kapitel in Form eines Gedankenexperimentes aufgezeigt, dass selbst ein *Gold Standard*, der als Evaluierungsgrundlage dienen soll, gemessen an dem hypothetisch vollständigen Strukturen, die hätten extrahiert werden können, selber ein sehr geringes *F-Measure* erreichen würde. Das daraus resultierende Paradoxon wird durch die *Noisy Channel Evaluation* dadurch aufgelöst, dass durch die *Randomised Digital Library* immer eine Evaluierungsgrundlage gegeben ist, die exakt die gleiche Größe hat, wodurch in der Terminologie des *Text Re-use* vermieden wird, *Äpfel mit Birnen zu vergleichen*. Es wurde die Grundlage geschaffen, die eine Evaluierung sowohl für automatische Methoden und zeitgleich auch für manuelle fachwissenschaftliche Arbeit ermöglicht. Damit ist erstmals ein direkter Vergleich zwischen beiden Herangehensweisen möglich.

Weiterhin kann ein Unterschied zwischen zwei Werten der *Mining Ability* $\mathcal{L}(\Theta)$ unmittelbar miteinander verglichen werden und hat eine Bedeutung. Einerseits kann eine höhere *Mining Ability* zwischen zwei Verfahren auf der gleichen *Digital Library* ausdrücken, dass eine *Mining*-Methode besser ist als eine andere bzw. dass sich ein bestimmter Parameter für die Fähigkeit, zwischen Zufall und Struktur zu unterscheiden, besser eignet. Andererseits kann eine höhere *Mining Ability* zwischen zwei verschiedenen *Digital Libraries* mit der gleichen Methode auch ausdrücken, dass eine *Digital Library* mehr vom Zufall unterscheidende Strukturen enthält als die andere. Aus diesem Grund wurde eingangs zu diesem Kapitel *Text Mining* als das Auffinden von nicht zufälligen Strukturen definiert.

Die Mächtigkeit der *Noisy Channel Evaluation* wird deutlich, wenn die *Qualitätskriterien* für *Text Mining* aus Abschnitt 2.4 hinzugezogen werden. In Abschnitt 4.3 sind einige absichtliche und unabsichtliche Verfälschungen von Ergebnissen aufgezeigt worden, welche eine klassische Evaluierung gegen einen *Gold Standard* oftmals so verfälschen, dass beim Versuch der Reproduktion von Ergebnissen diese häufig nicht wiederholt werden können. Als ein Beispiel ist das intelligente Abschneiden des *Long Tails* genannt worden, wodurch sich der *Recall* bei nahezu identischer *Precision* deutlich erhöht und somit auch das *F-Measure* positiv beeinflusst. Als ein Qualitätskriterium in dieser Arbeit ist in Abschnitt 4.3 die *Circumvention* genannt worden. Jedes *Mining*-Verfahren und dementsprechend auch deren Evaluierungen darf nicht manipulierbar sein. Bei der Evaluierung gegen einen *Gold Standard* hingegen können durch zu viele kleine Parameter und Heuristiken die Ergebnisse positiv verändert werden. Bei der *Mining Ability* ist jedoch eine *Circumvention* nahezu ausschließbar, da die *Evaluierungsbasis*, die *Randomised Digital Library*, ebenfalls durch einfache und präzise Algorithmen generiert wird.

Neben der in diesem Kapitel zugrunde liegenden *Evaluation* kann die *Noisy Channel Evaluation* auch vielseitig in anderen Szenarien eingesetzt werden. Zwei dieser Einsatzmöglichkeiten sollen an dieser Stelle kurz erklärt werden. Während verschiedener Testläufe hat sich gezeigt, dass eine *T4*-Randomisierung sehr gut dazu geeignet ist, um speziell die drei verschiedene *Meme Idiom*, *Winged Word* sowie *Syntactical Phrase* (vgl. Abschnitt 2.6) vom restlichen *Text Re-use* zu separieren. Diese drei *Meme* können als nicht absichtlicher *Text Re-use* verstanden werden. Eine weitere Gemeinsamkeit liegt darin, dass sie syntaktisch als fest zu bezeichnen sind und vergleichsweise frequentiert auftreten. Genau diese Eigenschaften bewirken bei eine *T4*-Randomisierung (vgl. *Text Synthese* in Tabelle 4.7 auf Seite 150), dass diese drei syntaktisch festen *Meme* auch in der *Randomised Digital Library* generiert

werden können. Als Konsequenz ist es aus der Überlappung einer *Text Re-use Analysis* sowohl auf einer *Digital Library* als auch einer *T4*-randomisierten *Randomised Digital Library* möglich, genau diese *Meme* in beiden Ergebnissen zu beobachten und dementsprechend zu isolieren.

Als ein weiteres Szenario kann die Analyse von *OCR* digitalisierten Texten angesehen werden. Die grundlegende Fragestellung lautet, wie müssen Texte aufbereitet sein, damit darauf sowohl *Information Retrieval*- als auch *Text Mining*-Techniken angewendet werden können. Letztlich können einerseits Texte durch ein starkes *Postprocessing* verbessert werden. Andererseits können *OCR*-Fehler nicht von Rechtschreibfehlern, sprachevolutionären Varianten oder auch nur verschiedenen Dialekten unterschieden werden. Mit der *Mining Ability* $\mathcal{L}(\Theta)$ kann die Qualität von durch *OCR* digitalisierten Texten auf die *Mining*-Fähigkeit untersucht werden. Die im Rahmen dieses Kapitels eingesetzte *T2*-Randomisierung eignet sich besonders gut für solche Tests. Auf diese Weise kann der bestmögliche Kompromiss gefunden werden, der auf der einen Seite Texte durch ein *Postprocessing* nicht zu stark verändert und auf der anderen Seite Texte in eine *Digital Library* mit dem Ziel integriert, dass mit ihnen auch gearbeitet werden kann. Das bedeutet sowohl, dass *Information Retrieval*- und *Text Mining*-Techniken mit den fehlerbehafteten Texten arbeiten können, aber auch, dass die *Digital Edition* im Sinne eines *Postprocessing* nicht zu sehr von sprachlichen Varianten gesäubert wird.

In diesem Kapitel wurde sehr vereinfacht auf die beiden kleinstmöglichen, mehrgliedrigen *Atome* des *Text Re-use*, das *Bigram* und die *Co-occurrence*, eingegangen. Ziel war es, einige Grundeigenschaften in den Abschnitten 4.6.1 bis 4.6.3 aufzuzeigen. Dennoch hat sich in diesen bereits sehr einfachen Szenarien gezeigt, dass das Verhalten von der Größe der *Digital Library* aber auch der Wahl von Parametern stark abhängig ist. Allein die Rechenzeit für die Analysen aus den Abschnitten 4.6.1 bis 4.6.3 beträgt bereits etwa sechs Prozessormonate bei lediglich zwei abhängigen Variablen. Bei der *7-Level-Architektur* (vgl. Abschnitt 3.1) des *Historical Text Re-use* ist der Raum der Abhängigen deutlich größer, so dass dieses Optimierungsproblem durch verschiedene *Hill Climbing*-Techniken (vgl. [Cormen 2001, Foss 2006]) gelöst bzw. bestmöglich gelöst werden muss. In jedem Fall ist auf Basis der bereits in den einfachen Experimenten zu diesem Kapitel gesammelten Erfahrungen zu erwarten, dass eine enorme Menge an *Computational Resource* nötig sein wird, die *Mining Ability* für komplexere *Mining*-Modelle einzusetzen. Insbesondere, wenn nicht nur ein *lokales*, sondern das *globale Maximum* für die *Mining Ability* $\mathcal{L}(\Theta)$ bestimmt werden soll.

Zusammenfassend kann gesagt werden, dass dieses Kapitel die *Noisy Channel Evaluation* mit ihrem *Score* der *Mining Ability* eingeführt hat. Das wurde an zwei sehr einfachen Beispielen gezeigt. Vielmehr wurde ein grundlegendes Problem der Sprachstatistik dargestellt. Durch die *Mining Ability* ist es möglich gewesen, dass teilweise sehr negative Verhalten einer größer werdenden *Digital Library* deutlich zu machen, welches mit herkömmlichen Evaluierungsmethoden nicht erfassbar ist. Des Weiteren wurden die hier angeführten Beispiele gezielt ausgewählt. Einerseits wurde als Motivation für die *Bigram*- und *Co-occurrence*-Analyse immer angegeben, dass es die einfachste Form eines *Text Re-use* ist. Dieser Aspekt ist bei der Konzeption dieses Kapitels immer davon begleitet gewesen, auch zu reflektieren, wie schwierig die Sprachstatistik durch das *Power Law* bzw. der Umgang mit probabilistischen Sprachmodellen letztlich ist. Das in diesem Kapitel nur an *Bigrams* oder *Co-occurrences* aufgezeigte Verhalten kann sehr leicht auf bedingte Wahrscheinlichkeiten sowie jedwede Form von Mathematik mit Wortwahrscheinlichkeiten extrapoliert werden. Die Auswirkung der wieder sinkenden *Mining Ability* bei steigender Größe der *Digital Library* kann aufgrund des *Zipfsche Gesetzes* letztlich überall erwartet werden. Die das Kapitel begleitenden und aufgezeigten sprachstatistischen Probleme sind der Hauptgrund, warum im Kapitel 3 probabilistische Modelle nicht nennenswert vertreten und bestenfalls der Voll-

ständigkeit halber erwähnt worden sind. Ein positiver Aspekt bleibt dennoch. Aus Tabelle 4.2 in Abschnitt 4.2 kann entnommen werden, dass ab einem *Re-use Overlap* von etwa drei Wörtern, und damit nahezu jedem *Re-use Overlap*, die Überlappung immer als statistisch signifikant angesehen werden kann, so dass es keiner mathematischen bzw. probabilistischen Sprachmodelle bedarf.

Kapitel 5

Ergebnisse

Contents

5.1 Einführung . . . 168

5.2 *Text Re-use* in der *Perseus Digital Library* . . . 171

5.2.1 Level 3 - *Featuring*: Bigram Shingling vs. Unigram . . . 172

5.2.2 Level 1 - *Segmentation*: *Sentence*-basierte und nicht überlappende Segmentierung vs. *Moving Window* mit fester Fenstergröße . . . 172

5.2.3 Level 5 - *Scoring*: normalisierte Gewichtung vs. absolute Gewichtung des *Re-use Overlap* . . . 173

5.2.4 Evaluierung auf der *Perseus Digital Library*: *Wieviel Homer steckt in Athenaeus*? . . . 174

5.2.5 Zusammenfassung . . . 175

5.3 System Evaluation . . . 176

5.3.1 Evaluierung durch *Precision & Recall* gegen einen *Gold Standard* . . . 178

5.3.2 Evaluierung durch *Noisy Channel Evaluation* . . . 185

5.3.3 Evaluierung durch *Text Re-use Compression* . . . 189

5.3.4 Zusammenfassung . . . 192

5.4 Component & Aggregated Evaluation . . . 201

5.4.1 Qualität der *Lemmatisation* . . . 202

5.4.2 Qualität im Umgang mit paradigmatischen Relationen . . . 203

5.4.3 Qualität im Umgang mit historischen Varianten . . . 204

5.4.4 Qualität der *Digital Signature* . . . 205

5.4.5 Qualität des *Linking* . . . 209

5.5 *Noisy Channel Mining*: Extraktion paradigmatischer und historischer Schreibweisen . . . 212

5.6 Zusammenfassung . . . 215

Failure is success if we learn from it.

Malcolm Forbes (1919-1990)

In diesem Kapitel wird auf Basis von sieben verschiedenen englischsprachigen Bibelversionen, der *Perseus Digital Library* sowie zwei digitalisierten Büchern deutscher Redewendungen eine *Text Re-use Analysis* evaluiert. Neben der eigentlichen Evaluierung steht insbesondere der Systemvergleich mit der *Text Re-use Compression* sowie der *Noisy Channel Evaluation* im Vordergrund. Um eine *Text Re-use Analysis* verbessern zu können, folgen der *System Evaluation* einzelne Aspekte einer *Component Evaluation*. Abschließend werden Ergebnisse und Perspektiven des *Noisy Channel Mining* aufgezeigt.

5.1 Einführung

In Abschnitt 2.6 wurde der *Historical Text Re-use* in seiner *Data Diversity* vorgestellt. Das Kapitel 3 führt eine entsprechende Implementierung einer *Text Re-use Analysis* ein. Wie wird nun letztlich das Ergebnis einer *Text Re-use Analysis* evaluiert? Im Rahmen dieses einführenden Abschnittes werden verschiedene Evaluierungstechniken systematisiert. In den folgenden Abschnitten sind jene Evaluierungsmethoden in verschiedenen Szenarien in der praktischen Anwendung.

Aus der Biometrie (vgl. [Maltoni 2009, BIMA 2012]) kann ein grundlegender Umgang mit Evaluierungen adaptiert werden. Einerseits gibt es eine *System Evaluation*. Auf der anderen Seite wird für eine Optimierung die *Component Evaluation* in Form von verschiedenen *Error Rates* eingesetzt. Um ein biometrisches Verfahren zu evaluieren, wird hierzu nicht nur das gesamte System getestet, sondern im Sinne einer Optimierung werden verschiedene Teile des Systems einzeln betrachtet. Liefert eine Systemanalyse in einem fiktiven Beispiel eine *Precision* von 0.8, dann ist es oftmals nicht möglich, aus dieser Evaluierung direkt eine Verbesserung ableiten zu können. Erst eine komponentenweise Evaluierung des *Scanners*, der *Feature Extraction*-Komponente oder der *Linking Software* bietet die Möglichkeit an, genau zu analysieren, wo entsprechende Verluste bei der *Precision* zu suchen sind. In der Automatischen Sprachverarbeitung hingegen werden weitestgehend nur *System Evaluation*, wie die Evaluierung gegen einen *Gold Standard*, angewendet. Meist bleibt jedoch offen, wie die *Black Box* des *Mining*-Verfahrens verbessert und evaluiert werden kann.

Ziel dieser Arbeit und insbesondere dieses Kapitels ist es, sowohl die *System* als auch die *Component Evaluation* für den *Historical Text Re-use* ein- sowie dementsprechend dann auch durchzuführen. Abbildung 5.1 fasst eine entsprechende Evaluierungstaxonomie zusammen.

Die *System Evaluation* des *Historical Text Re-use* kann sowohl in die *Qualitative* als auch *Quantitative Evaluation* eingeteilt werden (vgl. Abb. 5.1). Die *Quantitative Evaluation* besteht aus zwei Techniken, die im Rahmen dieser Arbeit bereits ausführlich eingeführt wurden. Die *Text Re-use Compression* ist eine quantitative Methode, welche die durch *Text Re-use* induzierte *Redundancy* bzw. Dopplungen innerhalb einer *Digital Library* dazu nutzt, um sie gezielt zu komprimieren und daraus abzuleiten, wie viel *Text Re-use* aufgedeckt worden ist (vgl. Abschnitt 3.10). Die *Noisy Channel Evaluation* hingegen misst die *Mining Ability* $\mathcal{L}(\Theta)$ einer *Digital Library* gegenüber einer *Randomised Digital Library* (vgl. Abschnitt 4.4).

Die *Qualitative Evaluation* umfasst das Testen gegen einen *Gold Standard* (vgl. Abschnitt 4.3) aber auch den manuelle Abgleich von Ergebnissen, der *Hand operated Evaluation*. Der Vorteil einer Evaluierung gegen einen *Gold Standard* ist, dass die Ergebnisse zumindest theoretisch gegen andere Verfahren einfach verglichen werden können[1]. Während die Datenmenge eines *Gold Standards* für einen solchen Test spricht, bietet die *Hand operated Evaluation* auch bei einer vergleichsweise kleinen Menge an überprüften Daten systematische Vorteile. Einerseits ergeben sich keine Verzerrungen durch verschiedene Überlappungen zwischen *Gold Standard* und der *Digital Library* (vgl. Abschnitt 4.3). Andererseits wird bei einem Test durch einen *Gold Standard* nur gegen die Daten der Evaluierungsbasis getestet, während die *Hand operated Evaluation*, also das Betrachten von Ergebnislisten, auch nicht im *Gold Standard* enthaltene Daten berücksichtigt. Die Auswirkung für die Interpretation eines solchen Tests kann an einem einfachen Beispiel veranschaulicht werden. Wie in Ab-

[1] Es sei an dieser Stelle vermerkt (vgl. Abschnitt 4.3), dass die Evaluierung gegen einen *Gold Standard* systematische Nachteile hat. Im hier genannten Kontext sei insbesondere auf die etwaige ungleiche Überlappung zwischen *Gold Standard* und der *Digital Library* hingewiesen, welche systematische Verzerrungen der Ergebnisse verursachen kann.

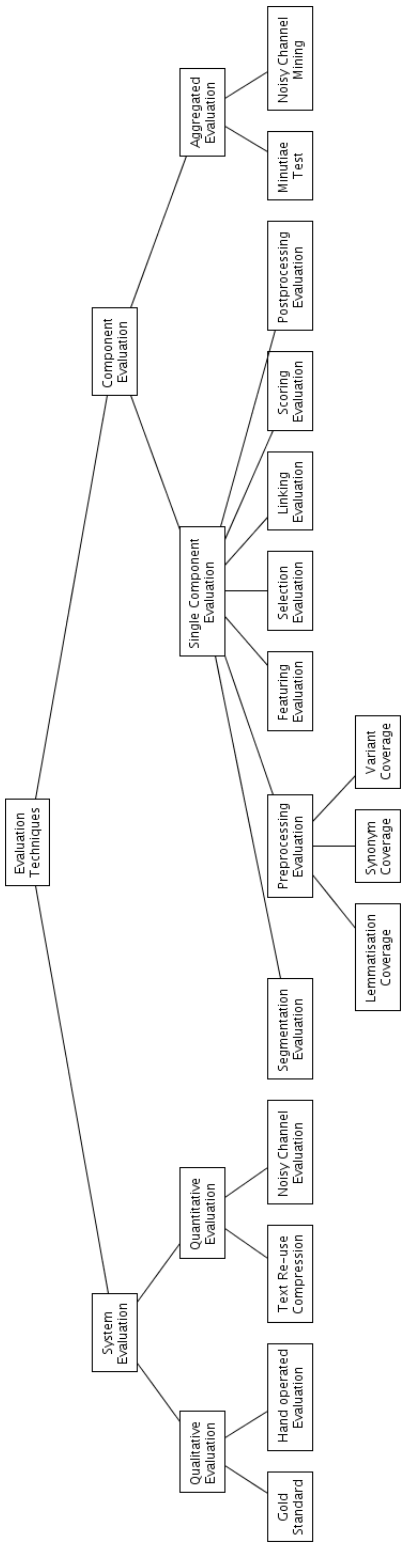


Abbildung 5.1: Taxonomie von Evaluierungstechniken für den *Historical Text Re-use*. Aus der *Biometrie* (vgl. [Maltoni 2009, BIMA 2012]) kann sowohl die *System* als auch die *Component Evaluation* adaptiert werden, um ein Verfahren bzw. ein Teil eines Verfahrens zu evaluieren.

schnitt 4.3 bereits eingeführt wurde, ist die Größe des *Gold Standards* deutlich kleiner als die hypothetisch maximale Menge an Daten, die zu messen wäre. Je nach Parametern liegt das Ergebnis einer *Mining*-Analyse in der Regel zwischen diesen beiden Extrema. Es sei nun der Einfachheit halber angenommen, dass das *Mining*-Ergebnis nur 1000-mal größer als der *Gold Standard* ist. Es ist nun ein Leichtes sich zu überlegen, dass das Evaluieren gegen einen *Gold Standard* eine zu geringe Datendichte aufweist. So können zwar gegebenenfalls eine gute *Precision* und ein guter *Recall* berechnet sein, jedoch bleibt die Frage, ob die 999 von 1000 anderen Ergebnisse genau diese Evaluierung ebenfalls reflektieren würden. Auch wenn bei einer *Hand operated Evaluation* meist deutlich weniger Testdaten zugrunde liegen, so kann genau dieser Frage der fehlenden Repräsentierbarkeit von nicht im *Gold Standard* enthaltenen Daten Genüge getan werden. Oftmals ist die somit verbundene *Precision* sehr viel geringer und reflektiert die wirkliche Qualität deutlich besser.

Die *Component Evaluation* kann in eine *Single Component Evaluation* sowie eine *Aggregated Evaluation* aufgeteilt werden. Die *Single Component Analysis* testet jedes einzelne Level separat auf dessen Qualität. So kann eine *Preprocessing Evaluation* derart gestaltet sein, dass die Qualität sowohl für die Lemmatisierung, den Umgang mit historischen Varianten als auch Effekte durch eine Stringähnlichkeitsanalyse ein gezieltes *Lemmatisation Coverage*, *Synonym Coverage* oder auch eine *Variant Coverage* (für alle vgl. Abb. 5.1) bestimmt wird (Details in Abschnitt 5.4). Ziel des *Coverage Score* ist es, die Abdeckung der *Tokens* zu bestimmen, für die bspw. eine Grundform vorliegt. Auf Basis eines solchen *Scores* kann dann wiederum entschieden werden, ob mitunter nicht durch die *Text Re-use Analysis* als solche ein Ergebnis schlecht ausgefallen ist, sondern lediglich die Lemmatisierung suboptimal funktioniert hat. Insbesondere ist es bei historischen Dokumenten nicht selten beobachtbar, dass über Jahrzehnte oder gar Jahrhunderte hinweg verschiedene Schreibweisen eines Konzeptes in Texten enthalten sind. Fast kein Morphologie- bzw. Lemmatisierungstool kann in der Gegenwart alle historischen Schreibweisen verarbeiten (vgl. Abschnitt 5.4.1).

Die *Aggregated Evaluation* testet nicht einzelne Level der *7-Level-Architektur*, wie es bei der *Single Component Evaluation* der Fall ist, sondern zwei oder drei kombinierte Level, so dass nie das gesamte System, sondern ein Teilsystem aus mehreren Komponenten evaluiert wird. Der *Minutiae Test* legt hierbei den Fokus auf die zwei Level *Featuring* und *Selection*. Der *Minutiae Test* misst vielmehr die Überlappung aus der digital bestimmten *Re-use Signature* und dem theoretisch optimalen *Re-use Nucleus*. Das *Noisy Channel Mining* (vgl. Abschnitt 5.5) kann hingegen als eine Aggregation der Level *Linking*, *Scoring* und *Postprocessing* verstanden werden, welches speziell auf Basis der gelinkten *Re-use Units* paradigmatisch benutzte Relationen extrahiert.

Da im Rahmen dieser Arbeit aus Gründen des Umfanges nicht alle Tests umgesetzt und für diese im Detail Ergebnisse generiert werden können, soll eine Beschränkung auf die wichtigsten Tests erfolgen. Als Datengrundlage für diese Form der Evaluierungen dienen drei *Digital Libraries*. Dem Abschnitt 5.2 liegt die *Perseus Digital Library* zugrunde. In den Abschnitten 5.3 bis 5.5 werden insgesamt sieben englischsprachige Bibelversionen analysiert. Im Abschnitt 5.4.4 werden zwei Bücher von deutschsprachigen *Idioms* aus dem Mittelalter bzw. biblische Redewendungen verwendet (vgl. [Wagner 2011a, Wagner 2011b]).

In diesem Kapitel sind alle vier in Abb. 5.1 genannten *System Evaluation* eingesetzt. Die *Hand operated Evaluation* ist die Grundlage für den Abschnitt 5.2. Das Evaluieren gegen einen *Gold Standard* wird sowohl im Abschnitt 5.2 auf altgriechischen Texten der *Perseus Digital Library* als auch im Abschnitt 5.3.1 auf englische Bibelversionen verwendet. In den Abschnitten 5.3.2 und 5.3.3 wird die Evaluierung gegen einen *Gold Standard* aus Abschnitt 5.3.1 mit der *Noisy Channel Evaluation* sowie der *Text Re-use Compression* verglichen.

Tabelle 5.1 reflektiert die sechs im Rahmen dieses Kapitels ausgewählten Tests der *Component Evaluation*-Klassifizierung. Insgesamt werden sechs der sieben Level der *7-Level-Architektur* des *Historical Text Re-use* (vgl. Kapitel 3) evaluiert.

Evaluierung	**Abschnitt**	**L1**	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**
Lemmatisation Coverage	5.4.1		X					
Synonym Coverage	5.4.2		X					
Variant Coverage	5.4.3		X					
Minutiae Test	5.4.4			X	X			
Linking Evaluation	5.4.5					X		
Noisy Channel Mining	5.5					X	X	X

Tabelle 5.1: Testsystematik dieses Kapitels bezogen auf die *7-Level-Architektur* des *Historical Text Re-use*. Aufgelistet sind insgesamt sechs verschiedene Tests (1. Spalte), die in den folgenden Abschnitten (2. Spalte) durchgeführt werden. *L1* bis *L7* entsprechen den sieben Level *Segmentation*, *Preprocessing*, *Featuring*, *Selection*, *Linking*, *Scoring*, *Postprocessing*.

Der *Minutiae Test* wird auf Datensätzen von 24 Personen zu jeweils 402 deutschsprachigen Redewendungen aus dem biblischen Kontext sowie dem Mittelalter durchgeführt (vgl. [Wagner 2011a, Wagner 2011b]). Alle weiteren Tests aus Tabelle 5.1 basieren auf den bereits genannten englischsprachigen Bibelversionen. Bei beiden *Digital Libraries* sind die *Re-use Units* bereits klar definiert. Einerseits sind die Redewendungen bereits aus [Wagner 2011a, Wagner 2011b] segmentiert. Auf der anderen Seite ist bei einer Analyse der Bibel der Vers als *Re-use Unit* nahe liegend. Aus diesem Grund ist im Rahmen dieses Kapitels auf eine ausführlichere *Segmentation Evaluation* (vgl. Tabelle 5.1) verzichtet worden. Dennoch wird im folgenden Abschnitt auch auf die Erfahrungen und Ergebnisse bei der Wahl der richtigen *Segmentation* eingegangen (vgl. insbesondere Abschnitt 5.2.2).

5.2 *Text Re-use* in der *Perseus Digital Library*

In diesem Abschnitt wird die *Complexity* einer *Text Re-use Analysis* auf der *Perseus Digital Library* reflektiert[2]. Die *Perseus Digital Library* umfasst zehn Millionen Wörter in Altgriechisch sowie sieben Millionen Wörter in Latein. Dieser Abschnitt enthält vier Teile. Ausgehend von einer initialen Methode aus der Plagiarismusforschung wird in den drei ersten Abschnitten jeweils eine nennenswerte Zäsur beschrieben. Der letzte Abschnitt umfasst letztlich die Evaluierung und die Ergebnisse.

Die initiale Methode, wie sie in der Plagiarismusforschung meistens eingesetzt wird, umfasst eine Satzsegmentierung. Während des *Preprocessing* wird die *Perseus Digital Library* sowohl lemmatisiert, alle diakritischen Zeichen entfernt als auch die Groß- und Kleinschreibung vereinheitlicht. Es wird das *Bigram Shingling* als *Featuring* eingesetzt. Weiterhin werden frequente *Features* aus der *Digital Signature* entfernt. Als *Scoring*-Metrik wird Broder's *Resemblance* benutzt (vgl. Abschnitt 3.7).

Mit diesen Einstellungen wird zwar ein *Text Re-use* festgestellt, jedoch kann nur ein identischer bzw. nahezu wortwörtlicher *Text Re-use* aufgedeckt werden. Ausgehend von der hohen *Precision* wird nachfolgend dargestellt, wie der *Recall* optimiert wird, ohne dass dies nur durch ein Heruntersetzen der Parameter geschieht. Die Verbesserung des *Recalls*

[2] Die Ergebnisse dieses Abschnitts sind in [Büchler 2012c] auf der Konferenz *Theory and Practice of Digital Libraries 2012* vorgestellt worden.

wird allein dadurch generiert, dass das Sprachmodell auf den in der *Perseus Digital Library* enthaltenen *Text Re-use* angepasst wird. Die nachfolgenden Experimente fassen die Erfahrungen aus dem Zeitraum zwischen Juni 2011 und März 2012 zusammen[3].

5.2.1 Level 3 - *Featuring*: Bigram Shingling vs. Unigram

Ausgehend von dem initialen Setup konnten insgesamt 553285 Links zwischen den *Re-use Units* der *Perseus Digital Library* generiert werden. 99.7% der Links wurden jedoch nicht zwischen verschiedenen Werken aufgedeckt, sondern reflektieren den *Text Re-use* innerhalb eines Werkes. Im Detail bedeutet dies, dass 18 der 20 Werke mit dem höchsten *Text Re-use* nur einen *Self Text Re-use* enthalten. Das Werk mit dem stärksten *Self Text Re-use* ist *Elements* von *Euclid*[4], welches zahlreiche mathematische und sich wiederholende Ausdrücke enthält, wie bspw. Streckenbezeichnungen $\overline{AB}$, so dass *Re-use Units* sich ähnlicher erscheinen als sie kognitiv in der Interpretation der Formeln sind. Letztlich kann dieser *Self Text Re-use* auf eine bereits sehr formelhafte und formale Sprache in der frühen Mathematik zurückgeführt werden. Weiterhin sind Werke von Platon und Homer in den Top 20 der Werke mit dem meisten *Text Re-use* vertreten. Hierbei liegt der Grund nach Prüfung der Texte in autor- bzw. genrespezifischen Ausdrücken, so dass das initiale Sprachmodell eher autor- bzw. textsortenspezifische Merkmale gemessen hat.

Aus diesem Grund wurde das *Featuring* von einem *Bigram Shingling* zu einem *Word based Featuring* modifiziert, wodurch fast 19% mehr Ergebnisse gefunden werden konnten.

5.2.2 Level 1 - *Segmentation*: *Sentence*-basierte und nicht überlappende Segmentierung vs. *Moving Window* mit fester Fenstergröße

Sätze, wie sie in modernen Editionen historischer bzw. antiker Quellen vorgefunden werden, können in den Manuskripten und Inschriften nur sehr selten beobachtet werden. Vielmehr sind sie in der Gegenwart durch Editoren zum besseren Verständnis bzw. Lesen hinzugefügt worden. Da es keine festen Satzregeln in der Antike gab, können die Satzendemarkierer eines Editors nicht als gesichert angenommen werden, da letztlich jeder Editor nach seinen Regeln und bestem Gewissen Zeichen für das Satzende eingefügt hat. Aufgrund dessen kann ein atypisches Verhalten der Satzlängenverteilung beobachtet werden.

Weiterhin kann in der *Perseus Digital Library* ein *Text Re-use* von nur sehr wenigen Worten ausgemacht werden. Anhand von 353[5] manuell gesammelten *Text Re-use Edges* kann die Verteilung, wie in Abb. 5.2 dargestellt, gezeigt werden. Im Detail haben über 80% der manuell gesammelten *Re-use Units* eine Länge von sieben Wörtern und weniger. Aus diesem Grund wird durch die satzweise Segmentierung und dem daraus resultierenden satzweisen Vergleich sehr viel *Text Re-use* nicht erkannt, da selbst fünf gemeinsame Wörter im *Re-use Overlap* bei einer durchschnittlichen Satzlänge von etwas mehr als 20 nicht einmal ansatzweise an den *Scoring*-Schwellwert t von 0.7 herankommt. Das Herabsetzen des

[3] Die folgenden Unterabschnitte reflektieren das systematische Verbessern der Ergebnisse auf der *Perseus Digital Library*. In erster Linie wird auch das Vor und Zurück wiedergegeben, um den *Text Re-use* innerhalb einer *Digital Library* zu verstehen. Für eine geisteswissenschaftliche Expertise während der Evaluierung konnte auf Prof. Dr. Gregory Crane, auch Koautor in der zugrunde liegenden wissenschaftlichen Publikation (vgl. [Büchler 2012c]), zurückgegriffen werden. Auch wenn nachfolgend nicht immer explizit erwähnt, so spiegeln die Ergebnisse die Interaktion zwischen Prof. Crane und dem Autor wider.

[4] vgl. `http://de.wikipedia.org/wiki/Euklid`

[5] Die Evaluierungsgrundlage wurde im Rahmen einer geisteswissenschaftlichen Lehrveranstaltung an der *Tufts University* in Boston, USA unter Leitung von Prof. Dr. Gregory Crane von Studenten erstellt.

Schwellwertes erhöht zwar den *Recall* lässt aber durch sehr viel Rauschen auch die *Precision* inakzeptabel abfallen.

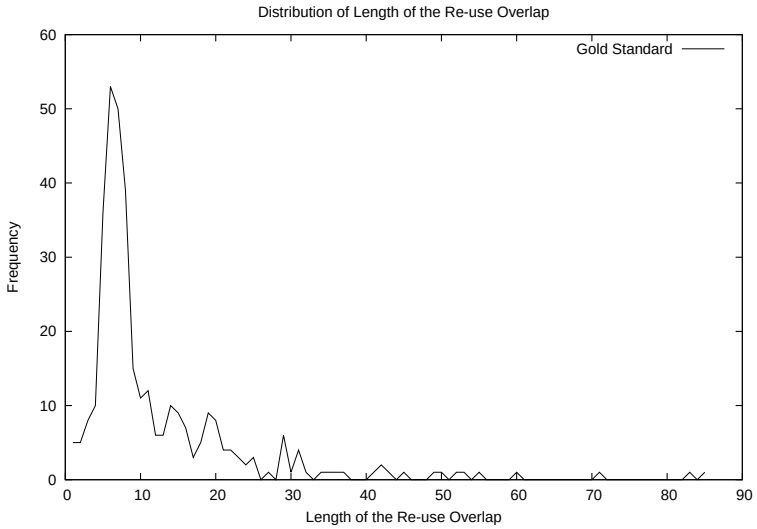


Abbildung 5.2: Verteilung der Länge des *Re-use Overlaps* gemessen an 353 manuell gesammelten Datensätzen im Rahmen einer fachwissenschaftlichen Lehrveranstaltung.

Das genannte Rauschen würde bei einer satzweisen Segmentierung und dem Herabsetzen des Schwellwertes für den *Similarity Score* dadurch induziert werden, dass sich Wörter eines kleineren *Re-use Overlaps* über den ganzen Satz verteilen und somit eher als ein zufälliger *Re-use Overlap* angesehen werden müssen. Ferner wird hierdurch die Eigenschaft der *Locality* des *Text Re-use* (vgl. Abschnitt 2.5) verletzt.

Da für die *Perseus Digital Library* sowohl die Zeichen für ein Satzende nicht eindeutig definiert als auch der enthaltene *Text Re-use* in seiner Länge sehr kurz sind, wurden die Experimente von einer satzweisen Segmentierung zu einem *Moving Window*-Ansatz mit einer festen Fenstergröße von 5 verlegt. Die Wahl der Fenstergröße hat sich in verschiedenen Experimenten mit unterschiedlichen Größen als der beste Kompromiss aus *Locality* und Rauschen herausgestellt. Je größer das Fenster gewählt wird, desto mehr Rauschen kann bei einem sehr kurzen *Text Re-use* von zwei oder drei Wörtern beobachtet werden. Die Fenstergröße von 5 ist rein experimentell bestimmt.

5.2.3 Level 5 - *Scoring*: normalisierte Gewichtung vs. absolute Gewichtung des *Re-use Overlap*

Im initialen Setup wurde für das *Scoring* Broder's Resemblance $\theta^R_\Theta(s_i, s_j)$ mit einem Schwellwert von 0.7 eingesetzt (vgl. [Broder 1997a]), welches die Ähnlichkeit zweier *Re-use Units* s_i und s_j auf einen Bereich von $\theta^R_\Theta(s_i, s_j) \in [0, 1]$ normalisiert (vgl. Abschnitt 3.7). Dieses Maß wird u. a. in der Plagiarismusforschung oder auch bei einer *Text Re-use Analysis* auf Bibelversen (vgl. Abschnitt 5.3) eingesetzt, wobei die durchschnittliche Länge des *Re-use Overlap* deutlich größer als 10 ist (vgl. mit dem kurzen *Text Re-use*, welcher in Abschnitt 5.2.2 beschrieben worden ist). Auch wenn Stoppwörter explizit entfernt worden sind, so gibt es dennoch eine signifikante Menge von *Text Re-use*, der zwar den Schwellwert für den *Similarity Score* erfüllt, jedoch der *Re-use Overlap* zu allgemein ist, so dass nicht von einem *Text Re-use* ausgegangen werden kann. Aus diesem Grund wurde ein normalisiertes

Scoring nicht weiter verfolgt. Da es ferner nicht nur auf die Menge der Wörter im *Re-use Overlap* ankommt, sondern auch auf deren Qualität im Sinne von Semantik, wurde das *Scoring* durch das *Word Class Weighting* (vgl. Abb. 3.9(b)) ausgetauscht.

Bedingt durch den kurzen *Text Re-use* in der *Perseus Digital Library* können zahlreiche *Meme* (vgl. Abschnitt 2.6) beobachtet werden. Während sich bei einer satzweisen Segmentierung zahlreiche Wörter im *Re-use Overlap* befinden, muss bei einem kurzen *Text Re-use* eine Unterscheidung zwischen *Meme*, wie *Multi Word Unit* oder *Idiom*, gemacht werden. *Multi Word Units*, wie *King Alexander the Great*, sind sowohl syntaktisch fest, als auch mit zwei Substantiven sowie einem Namenszusatz bereits drei der vier Wörter semantisch im Sinne des *Word Class Weighting* gewichtig. In den untersuchten Daten kann hierzu festgestellt werden, dass in mehr als 90% des aufgedeckten *Text Re-use* mindestens ein Verb enthalten ist, wie z. B. in *jemandem auf das Dach steigen* oder auch *sein oder nicht sein, das ist hier die Frage*. Durch dieses Kriterium können einfach *Multi Word Units* separiert werden, welche sonst zumindest die quantitative Interpretation der *Text Re-use Analysis* als auch die Evaluierung verfälschen.

5.2.4 Evaluierung auf der *Perseus Digital Library*: *Wieviel Homer steckt in Athenaeus?*

Um die Fähigkeit *Text Re-use* auf der *Perseus Digital Library* zu bestimmen, wurden im Rahmen einer Lehrveranstaltung insgesamt 353 Instanzen von *Text Re-use* manuell in einer digitalen Edition von Athenaeus' *Deipnosophistai Re-use Styles* von den Typen *Quote*, *Allusion* als auch *Paraphrase* markiert. Als Quelle des markierten *Text Re-use* dienen Homer's *Iliad* und *Odyssey*.

Ausgehend von der digitalen Edition konnten drei verschiedene Typen von Markierungen des *Text Re-use* festgestellt werden, welche in Abb. 5.3[6] dargestellt sind.

Cit-quote-bibl	blockquote	bibl without quote
<cit> <quote>du/o ku/nes a)rgoi\ ei (/ponto</quote> <bibl n="Hom. Od. 2.11">Od. 2.11</bibl> </cit>	<quote rend="blockquote">> - <line> a)gxou= d' i(stame/nh e)/pea ptero/enta proshu/da <bibl n="Hom. Il. 4.92">Il. 4.92</bibl> </line> - <line> a)ll' a)/ge nu=n ma/stiga kai\ h (ni/a sigalo/enta <bibl n="Hom. Il. 5.226">Il. 5.226</bibl> </line> </quote>	 [...]a)nti\ tou= proe/pinon. kuri/ws ga/r e)sti tou=to propi/nein, to\ e (te/rw| pro\ e(autou= dou=nai piei=n. kai\ o(*)odusseu\s de\ para\ tw=| * (omh/rw| <bibl n="Hom. Od. 13.57">Od. 13.57</bibl> [...]

Abbildung 5.3: Drei verschiedene Formen der Annotation für den *Historical Text Re-use*. Je nachdem wie gesichert ein *Text Re-use* angenommen werden kann, ist es möglich, durch das *quote*-Tag sowohl die Grenzen als auch die genaue Position des *Text Re-use* in *XML* festzuhalten.

Alle drei Arten von Annotationen (vgl. Abb. 5.3) werden separat von einander evaluiert, da mit den Annotationsformen unterschiedliche Grade der Sicherheit einhergehen. Das kann mit der Benutzung des *quote*-Tags (vgl. Abb. 5.3) begründet werden. Während in der linken Abbildung das *Tag* präzise an den Grenzen des *Text Re-use* gesetzt werden konnte, fehlt im rechten Bild das *quote*-Tag vollständig. Es gibt vielmehr lediglich ein *bibl*-Tag, welches

[6] Die Extraktion der Daten wurde im Rahmen der Publikation [Büchler 2012c] von Maria Moritz durchgeführt, welche auch diese Abbildung generiert hat.

den *Text Re-use* anzeigt. Der Grund für die unterschiedlichen Annotationsformen liegt im *Re-use Style* begründet. Während bei einem wortwörtlichen *Quote* die Grenzen des *Text Re-use* eindeutig bestimmt werden können, sind die linke und rechte Grenze des *Text Re-use* insbesondere bei *Re-use Styles*, wie der *Allusion* oder der *Paraphrase*, nicht eindeutig bestimmbar und kann von Editor zu Editor auch variieren (vgl. auch Abb. 2.2 auf Seite 60).

		Odyssey	Iliad	
found	Cit-quote-bibl	84	80	
	blockquote	34	50	
	bibl without quote	40	43	331
not found	Cit-quote-bibl	1	1	
	blockquote	11	7	
	bibl without quote	2	0	22
		172	181	353

Abbildung 5.4: Ergebnis der Evaluierung. Es wird sowohl nach Quelle, die *Odyssey* und die *Iliad* beide von Homer, als auch den drei Annotationsformen (vgl. Abb. 5.3) unterschieden.

Ausgehend vom initialen Setup können durch die in den Abschnitten 5.2.1 bis 5.2.3 gemachten Modifikationen am Modell insgesamt 331 von 353 Datensätzen der Evaluierungsgrundlage erkannt werden, was einem *Recall* von $R = 0.938$ entspricht. Die *Precision* liegt bei $P = 0.73$[7]. Eine *Precision* von $P = 1$ kann erreicht werden, wenn angenommen wird, dass mindestens fünf Wörter den *Re-use Overlap* bilden. Wird hingegen eine Größe des *Re-use Overlap* von mindestens drei Wörtern angenommen, liegt die *Precision* bei $P = 0.81$. Überlappungen von zwei Wörtern hingegen, wie *non olet*, sind nur schwer von *Bigrams* oder *Co-occurrences* zu unterscheiden, so dass das größere Rauschen durch sie in die Evaluierung einfließt.

Wie in Abb. 5.4 dargestellt, können insgesamt 22 Stellen für einen *Text Re-use* nicht gefunden werden. Fünf der 22 nicht aufgedeckten Testdatensätze bestehen nur aus einem einzigen Wort. Für *Text Re-use* wurde jedoch angenommen, dass mindestens zwei Wörter den *Re-use Overlap* formen müssen. Das ist nicht nur eine technische Einschränkung, sondern auch aus wissenschaftlicher Sicht nötig, um *Text Re-use* von *Topic Detection and Tracking* (vgl. [Allan 2002]) zu unterscheiden. Wenn demnach diese Restriktion nicht gemacht wird, könnte faktisch jedes Wort im Sinne seiner wiederholten Verwendung innerhalb einer *Digital Library* als *Text Re-use* aufgefasst werden. Auch wenn durch diese Einschränkung ein Teil des *Text Re-use* unerkannt bleibt, so ist die Menge im Verhältnis akzeptabel. Ferner besteht in einer fachwissenschaftlichen Anwendung, wie dem *Philological Crowd Sourcing*, immer die Möglichkeit, den entsprechenden auf ein Wort bezogenen *Text Re-use* auch nachträglich der Vollständigkeit halber hinzuzufügen. Zwei weitere Datensätze sind sehr stark paraphrasiert, so dass kein entsprechender Treffer gefunden werden konnte.

5.2.5 Zusammenfassung

Ziel dieses Abschnittes ist es, als Einführung in das Ergebnis-Kapitel dieser Arbeit aufzuzeigen, dass es bei der *Historical Text Re-use Analysis* keine Konstante gibt. Letztlich kann immer ausgehend vom Text nur eine Näherung des enthaltenen und somit maximal aufdeckbaren *Text Re-use* erfolgen, der oftmals in seiner Form als unbekannt angenommen werden muss. Als Erfahrungswert kann eine Dauer von 6 – 9 Monaten angegeben werden, um einen Überblick zum *Text Re-use* innerhalb einer *Digital Library* zu bekommen.

[7] Die *Precision* ist manuell von Prof. Dr. Gregory Crane bestimmt worden.

5.3 System Evaluation

In Abschnitt 2.6 sind bereits sowohl verschiedene *Meme* als auch unterschiedliche *Re-use Styles* vorgestellt worden. Auch wenn sie für die Systematik und das Gesamtverständnis wichtig sind, so ist es nahezu unmöglich, jede einzelne Permutation im Rahmen dieser Arbeit zu analysieren. Aus diesem Grund wird komplementär zu dem eher kurzen *Text Re-use* in der *Perseus Digital Library* (vgl. Abschnitt 5.2) in diesem Abschnitt der Fokus mit dem *Meme Verse* auf einem größeren *Text Re-use* liegen. Als Kantentypen können verschiedene *Re-use Edges* von *Verbatim* bis hin zur *Paraphrase* angenommen werden (vgl. Abb. 2.3 auf Seite 77), wobei letztere als dominant angesehen werden kann.

Für die *System Evaluation* wird auf sieben verschiedene englischsprachige Bibelversionen zurückgegriffen. Die Motivation hierfür ist dadurch gegeben, dass von einem *Archetyp* ausgehend alle anderen Bibeln und Übersetzungen abgeleitet sind. Sie haben gemeinsam, dass bspw. *Buch Genesis, Kapitel 1, Vers 1* (vgl. Tabelle 5.2) in jeder Bibelversion die gleiche Semantik hat. Der Unterschied ist jedoch, dass jede Bibelversion mit einem bestimmten Interesse bzw. aus einer Notwendigkeit erstellt worden ist. Die Änderungen können einerseits weitere Bücher und Texte in den einzelnen Editionen sein. Andererseits liegt den in diesem Abschnitt eingesetzten Bibelversionen zugrunde, dass sie in erster Linie unterschiedliche Intentionen bzgl. der eingesetzten Sprache haben.

Die älteste Version der Bibel im Rahmen dieser Arbeit ist *die King James Version* (KJV) aus dem 16. Jahrhundert, welche eine große Vielfalt von orthografischen sowie verschiedenen Sprachvarianten des *Early Modern English* beinhaltet. Dem gegenüber ist die *Bible in Basic English* (BBE) in einem so einfach wie möglichen Englisch der Gegenwart formuliert. Die *Young Literal Translation* (YLT) hingegen benutzt zwar auf der einen Seite ein qualifiziertes Englisch, jedoch ist sie auf der anderen Seite nahe an der hebräischen Satzsyntax. Bei dieser Übersetzung war es das Ziel, auch syntaktische Phänomene untersuchen zu können, wenn ein Forscher nicht des Hebräischen mächtig ist. Der *Webster Bible* (WBS) hingegen lag die *King James Version* zugrunde. Ziel der Revision von Webster[8] war es weniger, eine neue Bibelversion zu generieren, sondern in erster Linie die Bibel im 19. Jahrhundert von inzwischen unüblichen Schreibweisen zu bereinigen. Die *Darby Bible* wurde im 19. Jahrhundert aus hebräischen und griechischen Quellen ins Englische übersetzt und war anschließend Grundlage für verschiedene andere Bibelversionen. Weiterhin werden die *American Standard Version* (ASV) aus dem 20. Jahrhundert sowie die *World English Bible* (WEB) aus dem 21. Jahrhundert für die *System Evaluation* als Datenbasis genutzt. Alle sieben Bibelversionen sind über *Believer's Resource* (vgl. [Believer's Resource 2011]) als XML-Dateien erreichbar[9]. In diesem Sinne kann eine *Paraphrase* von einem einfachen *Rewording* bis hin zum Austauschen von synonymen Wörtern eines Verses verstanden werden.

Die hier genannten Unterschiede seien an einem der bekanntesten Verse der Bibel, *Buch Genesis, Kapitel 1, Vers 1* (vgl. Tabelle 5.2), verdeutlicht. Auch wenn dieser Vers in den verschiedenen Versionen im Vergleich zu anderen Beispielen noch relativ ähnlich bleibt, so sind dennoch die wesentlichen Merkmale bereits erkennbar. In der *BBE* wird *at the first* anstelle des temporalen Ausdruckes *in the beginning* benutzt. Auf der anderen Seite treten *create*, *make* sowie *prepare* in einer paradigmatischen Relation auf. Der Himmel kann sowohl im Singular *heaven* als auch im Plural *heavens* beobachtet werden.

[8] vgl. `http://en.wikipedia.org/wiki/Webster's_Revision`

[9] Die Daten wurden im Rahmen der Arbeit von Daniel Müller (vgl. [Müller 2011]) extrahiert und aufbereitet. Das schließt auch die noch nachfolgend erwähnte Erstellung der reduzierten Bibelversionen ein, die der *Text Re-use Analysis* dieses Abschnittes zugrunde liegen.

ASV	In the beginning God created the heavens and the earth.
BBE	At the first God made the heaven and the earth.
DBY	In the beginning God created the heavens and the earth.
KJV	In the beginning God created the heaven and the earth.
Webster	In the beginning God created the heaven and the earth.
WEB	In the beginning God created the heavens and the earth.
YLT	In the beginning of God‘s preparing the heavens and the earth.

Tabelle 5.2: Sieben verschiedene Versversionen aus *Buch Genesis, Kapitel 1, Vers 1.*

Nicht in jeder Bibelversion sind alle Übersetzungen und Verse enthalten. So sind Übersetzungen aus *Genesis*, *Mark* oder *Luke* in allen Fassungen enthalten, während Texte, wie *Baruch* und *Ecclesiasticus*, lediglich in bestimmten Versionen eingefügt wurden. Aus diesem Grund werden für die nachfolgende *System Evaluation* lediglich diejenigen Verse betrachtet, welche in allen sieben Bibelversionen enthalten sind. Einerseits schränkt das die Anzahl der Verse auf 28632 pro Bibel ein. Andererseits kann somit jedoch auch eine gute Evaluierungsbasis generiert werden. Hierbei wird angenommen, dass in jedem Vers, wie *Buch Genesis, Kapitel 1, Vers 1*, auch der gleiche Inhalt geschrieben steht, so dass alle sieben Bibelversionen in 21 paarweisen Vergleichen anhand der 28632 gemeinsamen Verse evaluiert werden können.

Tabelle 5.3 reflektiert einige grundlegende sprachstatistische Eigenschaften der sieben eingesetzten Bibelversionen. Die *BBE* stellt hierbei als einzige Fassung einen Ausreißer gegenüber den anderen Versionen dar. Zum einen enthalten die 28632 Verse knapp 50000 *Tokens* mehr als die anderen Bibelversionen. Dem steht ein Gesamtvokabular gegenüber, das in etwa nur die Hälfte der Wörter im Vergleich zu den anderen Fassungen enthält. Die durchschnittliche Satzlänge ist mit etwa 1.6 Wörtern nur leicht erhöht.

	# Tokens	**# Types**	**TTR**	**Avg. verse length**
ASV	741267	13485	54.97	25.89
BBE	791367	7350	107.67	27.64
DBY	732928	14971	48.96	25.60
KJV	746746	13466	55.45	26.08
Webster	722817	13556	53.32	25.25
WEB	744137	13655	54.50	25.99
YLT	745422	13973	53.35	26.03

Tabelle 5.3: Grundlegende sprachstatistische Kennzahlen für die 28632 gemeinsamen Verse der sieben eingesetzten Bibelversionen.

Um den *Text Re-use* zu bestimmen, sind fünf der sieben Level der *7-Level-Architektur* des *Text Re-use* (vgl. Kapitel 3) als konstant angenommen und nur für das *Preprocessing* sowie das *Featuring* werden unterschiedliche Ansätze evaluiert.

Bezüglich der *Segmentation* sind alle sieben Bibelversionen versweise zerlegt, so dass insgesamt $200424 = 7 \cdot 28632$ Verse aus sieben reduzierten Fassungen für die Analyse in Betracht gezogen werden können.

Im *Preprocessing* wird immer die Groß- und Kleinschreibung normalisiert, was nachfolgend durch *Base* ξ_1 für *Baseline* gekennzeichnet ist. Durch das Normalisieren der Groß- und Kleinschreibung werden insbesondere editorische Varianten, wie *God* und *GOD*, auf eine gemeinsame Schreibweise umgeformt.

Des Weiteren sind darauf aufbauend drei Vorverarbeitungsschritte unabhängig von einander durchgeführt (vgl. Tabelle 5.4):

- *StringSim* ξ_2: *String Similarity* wird insbesondere bei den historischen Bibelversionen eingesetzt, um ähnliche Schreibweisen eines Wortes aufzudecken und über eine Heuristik durch die plausibelste Form auszutauschen (vgl. Abschnitt 3.3). Für alle sieben Bibelversionen sind insgesamt 83866 ungerichtete Kanten zwischen ähnlich geschriebenen Wörter berechnet worden.
- *Lemma* ξ_3: Durch die *morphy*-Funktion des *Natural Language Toolkit* (vgl. [Bird 2009]), welches über *WordNet* (vgl. [Miller 1995, Fellbaum 1998]) erreichbar ist, werden für die Wortformen aus allen sieben Bibeln insgesamt 7479 Wort- zu Grundform-Paare bestimmt. Wenn keine Grundform für eine Wortform bekannt ist, dann bleibt die Wortform erhalten.
- *Lemma+Syn* ξ_4: Weiterhin sind aus *WordNet* (vgl. [Miller 1995, Fellbaum 1998]) Daten für Synoyme extrahiert worden, um eine Normalisierung der sprachlichen Vielfalt durchzuführen. Hierfür ist eine grundformreduzierte *Digital Library* notwendig. Aus *WordNet* werden für alle sieben Bibelversionen insgesamt 33074 relevante Synonympaaren extrahiert.

Für das *Featuring* werden drei verschiedene *Atome Word*, *Bigram* sowie *Trigram* ausgewählt. Tabelle 5.2 lässt vermuten, dass alle drei *Atome* gute Ergebnisse liefern. In Abschnitt 4.2 wurde bereits aufgezeigt, dass ein größeres *Ngram* nicht notwendig ist. Für die *Atome Bigram* und *Trigram* wird das *Bigram Shingling* sowie *Trigram Shingling* (beide vgl. Abschnitt 3.4) aufgrund ihrer überlappenden Eigenschaft eingesetzt, die die *Bi-* und *Trigrams* damit resistenter gegenüber Einschüben und Löschungen macht.

Tabelle 5.4 reflektiert die insgesamt zwölf Möglichkeiten S_{ij}, die vier, i, *Preprocessing*- und drei, j, *Featuring*-Techniken zu kombinieren.

		Featuring		
		Trigram	Bigram	Word
Preprocess.	Base	S_{11}	S_{21}	S_{31}
	StringSim	S_{12}	S_{22}	S_{23}
	Lemma	S_{13}	S_{23}	S_{33}
	Lemma+Syn	S_{14}	S_{24}	S_{34}

Tabelle 5.4: S_{ij}-Matrix für i *Preprocessing*- und j *Featuring*-Techniken, die im Rahmen der Evaluierung untersucht werden.

Für das *Feature Selection* wird ein *Maximum Pruning* (vgl. Abb. 3.6(b)) auf Seite 109 mit *Local Selection Usage* (vgl. Abschnitt 3.5) eingesetzt. In erster Linie soll diese *Selection*-Strategie dazu dienen, die Rechenzeit durch das Entfernen von häufigen *Features* zu reduzieren. Die *Feature Density* (vgl. Definition 15) ist mit $\mathcal{F} = 0.8$ angesetzt. Für das *Scoring* wird Broder's *Resemblance* (vgl. Abschnitt 3.7) mit einem Schwellwert von 0.6 ausgewählt.

5.3.1 Evaluierung durch *Precision & Recall* gegen einen *Gold Standard*

Für das Evaluieren verschiedener Techniken werden die paarweise verlinkten 28632 Verse jeder reduzierten Bibelversion benutzt, wodurch im *Recall* somit insgesamt max. 28632

Datensätze der Evaluierungsbasis pro Vergleich zweier Bibelversionen gefunden werden können. Für die gesamte *Digital Library* aus sieben Bibelversionen ergeben sich somit $601272 = 21 \cdot 28632$ symmetrische Kanten zwischen den Bibelstellen[10].

Die in den Tabellen 5.5 bis 5.7 abgebildeten Ergebnisse reflektieren eine Wiederholung der Experimente in [Büchler 2011c]. Der Versuchsaufbau ist bis auf einen Punkt identisch. Während in [Büchler 2011c] ein *Scoring*-Schwellwert t für Broder's *Resemblance* von $t = 0.7$ gewählt wurde, gilt im obigen Versuchsaufbau $t = 0.6$. Hintergrund und Motivation für das Herabsetzen des *Scoring*-Schwellwertes wird in Abschnitt 5.3.4 gegeben, in welchem gezeigt wird, dass das *F-Measure* bei teilweise deutlich geringen Schwellwerten t als $t = 0.7$ aus dem initialen Experimenten in [Büchler 2011c] maximal wird.

Im Vergleich lassen beide bis auf den Schwellwert t gleichen Experimente ein ähnliches Verhalten erkennen. So zeigt der *Recall* (vgl. Tabelle 5.6) bei allen drei *Featuring*-Techniken auf, dass es einen starken Zusammenhang zwischen den Versionen *KJV* und *WBS* gibt. Wie eingangs zu diesem Abschnitt bereits erwähnt wurde, kann dies damit motiviert und begründet werden, dass sich *Webster's Revision* (*WBS*) auf die *King James Version* bezieht. Vielmehr stellt *WBS* lediglich eine Anpassung des archaischen Englisch des 16. an die sprachlichen und grammatikalischen Gegebenheiten des 19. Jahrhunderts dar, so dass Änderungen vergleichsweise minimal ausfallen.

Des Weiteren sind vor allem die beiden Versionen *YLT* und *BBE* von Interesse, da in der einen Version eine für das Englische eher unübliche und an das Hebräische angelehnte Satzsyntax benutzt wird und in der anderen Version die Verse gemäß eines breiteren Publikums so paraphrasiert wurde, dass die Bibeltexte einfach und für jedermann verständlich sind. Für beide Varianten kann sowohl für das *Trigram* als auch das *Bigram Shingling* ein zu vernachlässigend niedriger *Recall* gemessen werden (vgl. Tabelle 5.6). Auch wenn die *Precision* (vgl. Tabelle 5.5) für beide Versionen vergleichsweise groß ist, so muss darauf hingewiesen werden, dass gemessen an den maximal 28632 als positiv zu findenden Versen, die *Precision* auf einer vergleichsweise sehr kleinen Menge von Datensätzen der Evaluierungsgrundlage basiert. Auch wenn sich für das *Word based Featuring* die Ergebnisse abheben, so ist insbesondere der *Recall* für *YLT* sowie *BBE* (vgl. Tabelle 5.6) von durchschnittlich weniger als 0.2 dennoch sehr gering. Unter Berücksichtigung der *Conditional Kolmogorov Complexity* (vgl. Abschnitt 2.8, Seiten 83 ff.) kann somit festgehalten werden, dass unabhängig von semantischen als auch syntaktischen Veränderungen *Text Re-use* genau dann schwierig zu erkennen ist, wenn der Grad der Veränderungen hoch ist. Im Kontext der *Conditional Kolmogorov Complexity* kann dies so interpretiert werden, dass das minimalistische Programm $\mathcal{P}_{min}$, welches ein Signal S in ein durch den *Noisy Channel* modifiziertes Signal S' transformiert, nicht zu groß sein darf, da ansonsten ein *Mining*-Verfahren den *Text Re-use* nicht mehr bestimmen kann.

Es gibt jedoch auch wesentliche Unterschiede in der Interpretation des Experimentes zu den Tabellen 5.5 bis 5.7 und dem Ausgangsexperiment in [Büchler 2011c]. Während im Ausgangsexperiment mit einem *Scoring*-Schwellwert $t = 0.7$ Unterschiede der vier eingesetzten *Preprocessing*-Techniken (vgl. Tabelle 5.4) festgestellt werden konnten, zeigen die Tabellen 5.5 bis 5.7 im Rahmen dieser Arbeit mit einem Schwellwert von $t = 0.6$ nahezu keinen Einfluss des *Preprocessing* auf das Ergebnis. In den Ausgangsexperimenten konnten deutliche Abweichungen im *Recall* beobachtet werden, sobald das *Preprocessing* verändert wurde. Je stärker das *Preprocessing* ausgewählt worden ist, desto höher war der *Recall*. Insbesondere ist bei jenen Experimenten auffällig gewesen, dass das *Preprocessing StringSim*,

[10] 21 entspricht der Anzahl der paarweisen Vergleiche aller sieben Bibelversionen mit den jeweils sechs anderen Versionen. Insgesamt werden somit, wie auch in den Tabellen 5.5 bis 5.7 dargestellt, 42 Vergleiche angestellt. Da mit Broder's *Resemblance* ein symmetrisches *Scoring*-Maß eingesetzt wird, ergeben sich 21 verschiedene Vergleiche.

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	0.70	0.70	0.70	0.71	0.75	0.75	0.75	0.72	0.61	0.62	0.63	0.66
ASV vs. DBY	0.74	0.75	0.74	0.75	0.72	0.73	0.73	0.73	0.73	0.73	0.73	0.74
ASV vs. KJV	0.94	0.94	0.94	0.95	0.92	0.92	0.92	0.93	0.71	0.71	0.71	0.71
ASV vs. WEB	0.94	0.94	0.94	0.94	0.92	0.92	0.92	0.93	0.83	0.83	0.82	0.82
ASV vs. WBS	0.93	0.92	0.93	0.94	0.91	0.91	0.92	0.92	0.84	0.84	0.84	0.84
ASV vs. YLT	0.91	0.92	0.93	0.93	0.81	0.83	0.88	0.89	0.44	0.47	0.50	0.52
BBE vs. ASV	0.70	0.70	0.70	0.71	0.75	0.75	0.75	0.72	0.61	0.62	0.63	0.66
BBE vs. DBY	0.88	0.86	0.88	0.88	0.82	0.82	0.83	0.83	0.55	0.56	0.57	0.60
BBE vs. KJV	0.56	0.59	0.56	0.58	0.40	0.42	0.41	0.41	0.35	0.37	0.22	0.24
BBE vs. WEB	0.70	0.71	0.71	0.72	0.76	0.76	0.77	0.78	0.56	0.56	0.59	0.61
BBE vs. WBS	0.18	0.20	0.18	0.20	0.18	0.20	0.19	0.20	0.20	0.21	0.21	0.23
BBE vs. YLT	0.91	0.91	0.92	0.92	0.87	0.88	0.89	0.89	0.66	0.68	0.71	0.79
DBY vs. ASV	0.74	0.75	0.74	0.75	0.72	0.73	0.73	0.73	0.73	0.73	0.73	0.74
DBY vs. BBE	0.88	0.86	0.88	0.88	0.82	0.82	0.83	0.83	0.55	0.56	0.57	0.60
DBY vs. KJV	0.94	0.95	0.94	0.94	0.92	0.92	0.92	0.92	0.83	0.82	0.82	0.83
DBY vs. WEB	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.82	0.82	0.82	0.82
DBY vs. WBS	0.94	0.95	0.94	0.94	0.92	0.92	0.92	0.92	0.67	0.68	0.68	0.68
DBY vs. YLT	0.93	0.93	0.93	0.94	0.87	0.87	0.93	0.93	0.79	0.80	0.83	0.84
KJV vs. ASV	0.94	0.94	0.94	0.95	0.92	0.92	0.92	0.93	0.71	0.71	0.71	0.71
KJV vs. BBE	0.56	0.59	0.56	0.58	0.40	0.42	0.41	0.41	0.35	0.37	0.22	0.24
KJV vs. DBY	0.94	0.95	0.94	0.94	0.92	0.92	0.92	0.92	0.83	0.82	0.82	0.83
KJV vs. WEB	0.96	0.96	0.96	0.96	0.95	0.94	0.94	0.94	0.84	0.84	0.84	0.84
KJV vs. WBS	0.95	0.95	0.95	0.95	0.92	0.91	0.92	0.92	0.76	0.76	0.76	0.76
KJV vs. YLT	0.94	0.94	0.94	0.94	0.83	0.84	0.91	0.91	0.84	0.84	0.88	0.88
WEB vs. ASV	0.94	0.94	0.94	0.94	0.92	0.92	0.92	0.93	0.83	0.83	0.82	0.82
WEB vs. BBE	0.70	0.71	0.71	0.72	0.76	0.76	0.77	0.78	0.56	0.56	0.59	0.61
WEB vs. DBY	0.91	0.91	0.91	0.92	0.91	0.91	0.91	0.91	0.82	0.82	0.82	0.82
WEB vs. KJV	0.96	0.96	0.96	0.96	0.95	0.94	0.94	0.94	0.84	0.84	0.84	0.84
WEB vs. WBS	0.96	0.96	0.96	0.96	0.93	0.93	0.92	0.92	0.79	0.80	0.80	0.80
WEB vs. YLT	0.65	0.70	0.69	0.70	0.72	0.74	0.75	0.76	0.77	0.79	0.82	0.83
WBS vs. ASV	0.93	0.92	0.93	0.94	0.91	0.91	0.92	0.92	0.84	0.84	0.84	0.84
WBS vs. BBE	0.18	0.20	0.18	0.20	0.18	0.20	0.19	0.20	0.20	0.21	0.21	0.23
WBS vs. DBY	0.94	0.95	0.94	0.94	0.92	0.92	0.92	0.92	0.67	0.68	0.68	0.68
WBS vs. KJV	0.95	0.95	0.95	0.95	0.92	0.91	0.92	0.92	0.76	0.76	0.76	0.76
WBS vs. WEB	0.96	0.96	0.96	0.96	0.93	0.93	0.92	0.92	0.79	0.80	0.80	0.80
WBS vs. YLT	0.73	0.75	0.74	0.92	0.74	0.76	0.80	0.91	0.81	0.82	0.84	0.87
YLT vs. ASV	0.91	0.92	0.93	0.93	0.81	0.83	0.88	0.89	0.44	0.47	0.50	0.52
YLT vs. BBE	0.91	0.91	0.92	0.92	0.87	0.88	0.89	0.89	0.66	0.68	0.71	0.79
YLT vs. DBY	0.93	0.93	0.93	0.94	0.87	0.87	0.93	0.93	0.79	0.80	0.83	0.84
YLT vs. KJV	0.94	0.94	0.94	0.94	0.83	0.84	0.91	0.91	0.84	0.84	0.88	0.88
YLT vs. WEB	0.65	0.70	0.69	0.70	0.72	0.74	0.75	0.76	0.77	0.79	0.82	0.83
YLT vs. WBS	0.73	0.75	0.74	0.92	0.74	0.76	0.80	0.91	0.81	0.82	0.84	0.87

Tabelle 5.5: *Precision* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß der berechneten *Precision* zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt.

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.09	0.10	0.11	0.12
ASV vs. DBY	0.16	0.17	0.17	0.17	0.28	0.30	0.30	0.31	0.70	0.72	0.73	0.74
ASV vs. KJV	0.36	0.38	0.37	0.38	0.53	0.56	0.55	0.56	0.86	0.88	0.88	0.88
ASV vs. WEB	0.32	0.34	0.32	0.33	0.46	0.48	0.47	0.47	0.76	0.79	0.77	0.77
ASV vs. WBS	0.27	0.29	0.28	0.29	0.44	0.46	0.46	0.46	0.82	0.84	0.84	0.85
ASV vs. YLT	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.18	0.21	0.25	0.26
BBE vs. ASV	0.02	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.09	0.10	0.11	0.12
BBE vs. DBY	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.07	0.08	0.08	0.10
BBE vs. KJV	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.08	0.09	0.10	0.11
BBE vs. WEB	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.11	0.12	0.13	0.15
BBE vs. WBS	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.10	0.11	0.13
BBE vs. YLT	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.04
DBY vs. ASV	0.16	0.17	0.17	0.17	0.28	0.30	0.30	0.31	0.70	0.72	0.73	0.74
DBY vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.07	0.08	0.08	0.10
DBY vs. KJV	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.62	0.65	0.65	0.66
DBY vs. WEB	0.07	0.08	0.07	0.08	0.14	0.15	0.14	0.15	0.46	0.49	0.49	0.51
DBY vs. WBS	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.64	0.67	0.67	0.68
DBY vs. YLT	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.18	0.21	0.26	0.27
KJV vs. ASV	0.36	0.38	0.37	0.38	0.53	0.56	0.55	0.56	0.86	0.88	0.88	0.88
KJV vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.08	0.09	0.10	0.11
KJV vs. DBY	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.62	0.65	0.65	0.66
KJV vs. WEB	0.10	0.11	0.10	0.10	0.18	0.20	0.19	0.19	0.51	0.55	0.53	0.55
KJV vs. WBS	0.75	0.78	0.76	0.77	0.89	0.91	0.90	0.90	0.99	0.99	0.99	0.99
KJV vs. YLT	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.14	0.16	0.19	0.20
WEB vs. ASV	0.32	0.34	0.32	0.33	0.46	0.48	0.47	0.47	0.76	0.79	0.77	0.77
WEB vs. BBE	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.11	0.12	0.13	0.15
WEB vs. DBY	0.07	0.08	0.07	0.08	0.14	0.15	0.14	0.15	0.46	0.49	0.49	0.51
WEB vs. KJV	0.10	0.11	0.10	0.10	0.18	0.20	0.19	0.19	0.51	0.55	0.53	0.55
WEB vs. WBS	0.11	0.12	0.11	0.12	0.20	0.22	0.21	0.21	0.56	0.60	0.59	0.60
WEB vs. YLT	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.12	0.15	0.16
WBS vs. ASV	0.27	0.29	0.28	0.29	0.44	0.46	0.46	0.46	0.82	0.84	0.84	0.85
WBS vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.10	0.11	0.13
WBS vs. DBY	0.12	0.13	0.12	0.13	0.22	0.24	0.23	0.24	0.64	0.67	0.67	0.68
WBS vs. KJV	0.75	0.78	0.76	0.77	0.89	0.91	0.90	0.90	0.99	0.99	0.99	0.99
WBS vs. WEB	0.11	0.12	0.11	0.12	0.20	0.22	0.21	0.21	0.56	0.60	0.59	0.60
WBS vs. YLT	0.01	0.02	0.02	0.01	0.02	0.03	0.03	0.03	0.15	0.17	0.21	0.22
YLT vs. ASV	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.18	0.21	0.25	0.26
YLT vs. BBE	0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.01	0.03	0.03	0.03	0.04
YLT vs. DBY	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.18	0.21	0.26	0.27
YLT vs. KJV	0.01	0.02	0.01	0.01	0.02	0.02	0.02	0.02	0.14	0.16	0.19	0.20
YLT vs. WEB	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.12	0.15	0.16
YLT vs. WBS	0.01	0.02	0.02	0.01	0.02	0.03	0.03	0.03	0.15	0.17	0.21	0.22

Tabelle 5.6: *Recall* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß des berechneten *Recall* zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt.

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	0.03	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.16	0.17	0.19	0.21
ASV vs. DBY	0.26	0.28	0.27	0.28	0.41	0.43	0.42	0.43	0.71	0.73	0.73	0.74
ASV vs. KJV	0.52	0.54	0.54	0.54	0.68	0.70	0.69	0.70	0.78	0.78	0.78	0.79
ASV vs. WEB	0.48	0.50	0.48	0.49	0.61	0.63	0.62	0.62	0.79	0.81	0.79	0.80
ASV vs. WBS	0.42	0.44	0.43	0.44	0.59	0.61	0.61	0.61	0.83	0.84	0.84	0.84
ASV vs. YLT	0.03	0.03	0.03	0.03	0.05	0.06	0.06	0.06	0.25	0.29	0.33	0.35
BBE vs. ASV	0.03	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.16	0.17	0.19	0.21
BBE vs. DBY	0.02	0.02	0.02	0.02	0.03	0.04	0.03	0.04	0.12	0.14	0.15	0.16
BBE vs. KJV	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.13	0.14	0.14	0.15
BBE vs. WEB	0.03	0.03	0.03	0.03	0.05	0.05	0.05	0.06	0.19	0.20	0.22	0.24
BBE vs. WBS	0.02	0.02	0.02	0.02	0.03	0.04	0.04	0.04	0.13	0.14	0.15	0.17
BBE vs. YLT	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.05	0.06	0.06	0.07
DBY vs. ASV	0.26	0.28	0.27	0.28	0.41	0.43	0.42	0.43	0.71	0.73	0.73	0.74
DBY vs. BBE	0.02	0.02	0.02	0.02	0.03	0.04	0.03	0.04	0.12	0.14	0.15	0.16
DBY vs. KJV	0.21	0.22	0.22	0.22	0.36	0.38	0.37	0.38	0.71	0.72	0.72	0.73
DBY vs. WEB	0.13	0.14	0.14	0.14	0.24	0.25	0.25	0.26	0.59	0.61	0.62	0.63
DBY vs. WBS	0.21	0.22	0.22	0.22	0.36	0.38	0.37	0.38	0.65	0.67	0.67	0.68
DBY vs. YLT	0.03	0.03	0.03	0.03	0.04	0.05	0.06	0.06	0.29	0.34	0.39	0.41
KJV vs. ASV	0.52	0.54	0.54	0.54	0.68	0.70	0.69	0.70	0.78	0.78	0.78	0.79
KJV vs. BBE	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.13	0.14	0.14	0.15
KJV vs. DBY	0.21	0.22	0.22	0.22	0.36	0.38	0.37	0.38	0.71	0.72	0.72	0.73
KJV vs. WEB	0.18	0.19	0.18	0.19	0.30	0.32	0.31	0.32	0.64	0.67	0.65	0.66
KJV vs. WBS	0.84	0.86	0.84	0.85	0.90	0.91	0.91	0.91	0.86	0.86	0.86	0.86
KJV vs. YLT	0.03	0.03	0.03	0.03	0.04	0.05	0.05	0.05	0.23	0.27	0.31	0.33
WEB vs. ASV	0.48	0.50	0.48	0.49	0.61	0.63	0.62	0.62	0.79	0.81	0.79	0.80
WEB vs. BBE	0.03	0.03	0.03	0.03	0.05	0.05	0.05	0.06	0.19	0.20	0.22	0.24
WEB vs. DBY	0.13	0.14	0.14	0.14	0.24	0.25	0.25	0.26	0.59	0.61	0.62	0.63
WEB vs. KJV	0.18	0.19	0.18	0.19	0.30	0.32	0.31	0.32	0.64	0.67	0.65	0.66
WEB vs. WBS	0.19	0.21	0.20	0.21	0.33	0.35	0.34	0.34	0.66	0.68	0.68	0.69
WEB vs. YLT	0.02	0.03	0.02	0.02	0.03	0.04	0.04	0.04	0.17	0.20	0.25	0.27
WBS vs. ASV	0.42	0.44	0.43	0.44	0.59	0.61	0.61	0.61	0.83	0.84	0.84	0.84
WBS vs. BBE	0.02	0.02	0.02	0.02	0.03	0.04	0.04	0.04	0.13	0.14	0.15	0.17
WBS vs. DBY	0.21	0.22	0.22	0.22	0.36	0.38	0.37	0.38	0.65	0.67	0.67	0.68
WBS vs. KJV	0.84	0.86	0.84	0.85	0.90	0.91	0.91	0.91	0.86	0.86	0.86	0.86
WBS vs. WEB	0.19	0.21	0.20	0.21	0.33	0.35	0.34	0.34	0.66	0.68	0.68	0.69
WBS vs. YLT	0.03	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.25	0.29	0.33	0.35
YLT vs. ASV	0.03	0.03	0.03	0.03	0.05	0.06	0.06	0.06	0.25	0.29	0.33	0.35
YLT vs. BBE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.05	0.06	0.06	0.07
YLT vs. DBY	0.03	0.03	0.03	0.03	0.04	0.05	0.06	0.06	0.29	0.34	0.39	0.41
YLT vs. KJV	0.03	0.03	0.03	0.03	0.04	0.05	0.05	0.05	0.23	0.27	0.31	0.33
YLT vs. WEB	0.02	0.03	0.02	0.02	0.03	0.04	0.04	0.04	0.17	0.20	0.25	0.27
YLT vs. WBS	0.03	0.03	0.03	0.03	0.05	0.05	0.05	0.05	0.25	0.29	0.33	0.35

Tabelle 5.7: *F-Measure* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß des berechneten *F-Measure* zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt.

welches ähnlich geschriebene Wörter berechnet und entsprechend durch den wahrscheinlichsten Kandidaten ersetzt, in der Evaluierung genauso gut wie die *Lemmatisation* und teilweise auch besser abgeschnitten hat. Das konnte mit der signifikanten Überlappung zwischen den durch *StringSim* als ähnlich bestimmten Wörtern und der *Lemmatisation* begründet werden. Bis auf irreguläre Verben besteht die Lemmatisierung zu großen Teilen aus dem Anwenden von Wortsuffixregeln, so dass die Wörter im Wortstamm sowie im Präfix gleich bleiben. Des Weiteren erreichte das *StringSim-Preprocessing* auch deswegen gute Ergebnisse, da historische Rechtschreibvarianten, wie *believedst* vs. *believest*, *gibeah* vs. *gibeath*, *galilaeans* vs. *galileans* oder *gishpa* vs. *gispa*, aufeinander abgebildet und während des *Preprocessing* normalisiert werden. Bei einem Schwellwert von $t = 0.6$ hingegen sind bis auf wenige Ausnahmen nur marginale Unterschiede im Vergleich der verschiedenenen *Preprocessing*-Schritte erkennbar (vgl. Tabellen 5.5 bis 5.7).

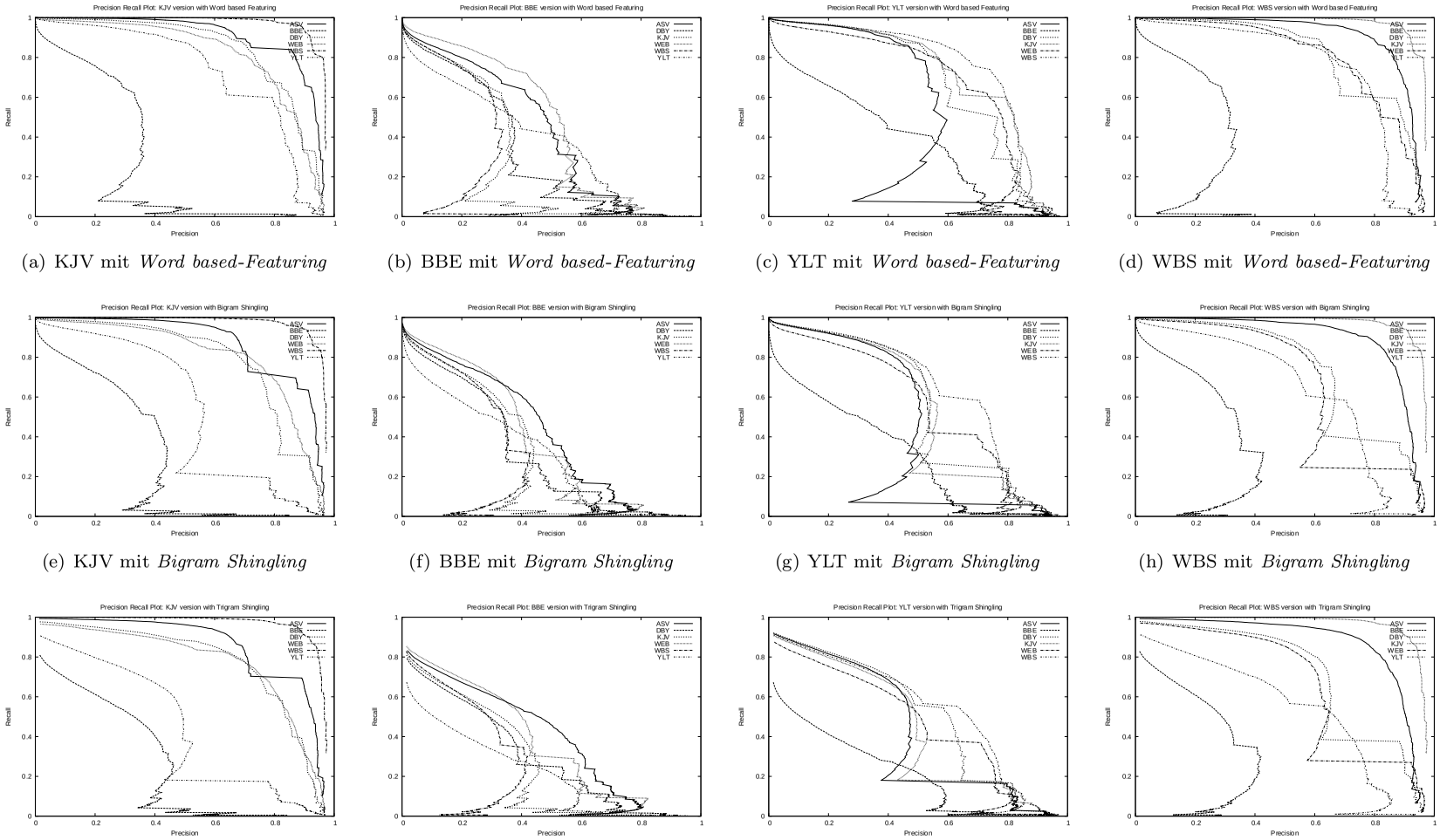
In Anlehnung an den Aspekt der verschiedenen historischen Schreibweisen aus dem Absatz zuvor sei insbesondere auf den *Text Re-use* zwischen den beiden Versionen *KJV* und *BBE* verwiesen (vgl. *Precision* in Tabelle 5.5). Für ξ_1 wird eine *Precision* $P = 0.35$ erreicht. Für die *StringSim*-Methode ξ_2 ist eine leichte Verbesserung auf $P = 0.37$ messbar. Sowohl für die *Lemmatisation* ξ_3 sowie zusätzlich die Normalisierung von Synonymen ξ_4 werden mit $P = 0.22$ sowie $P = 0.24$ schlechtere Evaluierungen erreicht. Vielmehr verschlechtert sich die *Precision* sogar deutlich ausgehend von der Baseline ξ_1 um mehr als 0.1.

Weiterhin kann eine positive Auswirkung unterschiedlicher *Preprocessing*-Techniken im Vergleich mit den anderen Editionen nur für die *YLT*-Version beobachtet werden (vgl. Tabelle 5.5). Gegenüber der *Preprocessing*-Baseline ξ_1 wird die *Precision* im Vergleich mit den anderen Bibelversionen immer um mindestens 0.04 verbessert. Bei einer *Text Re-use Analysis* zwischen den Versionen *YLT* und *BBE* wird sogar ein Unterschied in der *Precision* von 0.13 festgestellt. Jedoch sind diese Unterschiede ausschließlich beim *Word based Featuring* und nicht beim *Bigram* und *Trigram Shingling* feststellbar.

Die farblichen Hintergründe zwischen den Tabellen 5.6 und 5.7 zeigen bereits auf, dass es eine starke Korrelation bei der *Text Re-use Analysis* auf verschiedenen Bibelversionen zwischen dem *Recall* und dem *F-Measure* gibt. Das kann einfach damit begründet werden, dass der *Recall* von Werten $R < 0.01$ bis $R \approx 0.99$ stark streut (vgl. Tabelle 5.6), während die *Precision* mit wenigen Ausnahmen konstant hoch ist und mit Werten um $P \approx 0.9$ eine vergleichsweise geringe bzw. gering zu gewichtende Streuung besitzt.

Das maximale *F-Measure* F_{max} wird mit Ausnahme des Vergleiches zwischen *KJV* und *WBS* immer durch ein *Word based Featuring* erreicht (vgl. zeilenweise in Tabelle 5.7). Im Kontext der *Conditional Kolmogorov Complexity* ist dieses Ergebnis auch schlüssig, da das Paraphrasieren von Texten eben nicht nur aus einfachen Einschüben, Löschungen oder Ersetzungen besteht, sondern aus komplexeren Umformulierungen, so dass das minimalste Programm $\mathcal{P}_{min}$ für den *Edge Type Paraphrase* immer vergleichsweise groß ausfällt. Für den *Text Re-use* zwischen *KJV* und *WBS* wird F für das *Bigram Shingling* mit Werten von $F = 0.9$ und $F = 0.91$ maximal. Dem steht lediglich ein *F-Measure* von $F = 0.86$ für das *Word based Featuring* gegenüber.

Das *F-Measure* wird genau dann maximal (vgl. Formel 4.19 auf Seite 138), wenn sowohl die *Precision* P als auch der *Recall* R für eine *Digital Library* bzw. den Vergleich zweier Werke den bestmöglichen Kompromiss darstellen. Abb. 5.5 reflektiert genau dieses Verhältnis zwischen beiden Evaluierungsgrößen durch einen *Precision-Recall-Plot*. Hierzu wurden die Versionen *KJV* (die älteste Version), *BBE* (die durch einfachere Wörter am stärksten modifizierte Version), *YLT* (die syntaktisch am weitesten entfernte Bibelversion) sowie *WBS* (die Revision von *KJV* mit kleineren Veränderungen) ausgewählt. Weiterhin repräsentieren die drei Zeilen die *Featuring*-Techniken des *Word based Featuring*, des *Bigram Shinglings* sowie des *Trigram Shinglings*. Um einen *Precision-Recall-Plot* zu erhalten,



(a) KJV mit *Word based-Featuring* (b) BBE mit *Word based-Featuring* (c) YLT mit *Word based-Featuring* (d) WBS mit *Word based-Featuring*

(e) KJV mit *Bigram Shingling* (f) BBE mit *Bigram Shingling* (g) YLT mit *Bigram Shingling* (h) WBS mit *Bigram Shingling*

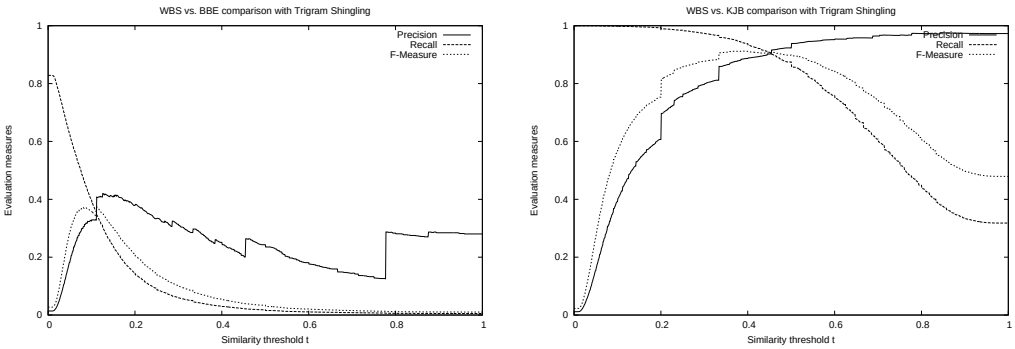
(i) KJV mit *Trigram Shingling* (j) BBE mit *Trigram Shingling* (k) YLT mit *Trigram Shingling* (l) WBS mit *Trigram Shingling*

Abbildung 5.5: *Precision-Recall-Plot* für *KJV*, *BBE*, *YLT* sowie *WBS* (Spalten) und *Word based Featuring*, *Bigram* sowie *Trigram Shingling* (Zeilen). Die Abbildungen reflektieren das unterschiedliche Verhalten von *Precision* und *Recall* beim paarweisen Vergleich von Bibelversionen.

wird für jeden *Scoring*- Schwellwert $t \in [0, 1]$[11] sowohl die *Precision* P als auch der *Recall* R berechnet. Da sich *Precision* und *Recall* umgekehrt proportional zueinander verhalten, ist ein Verlauf zu erwarten, der mit steigender *Precision* P bedingt durch einen sukzessive größer werdenden *Scoring*-Schwellwert t den *Recall* R kleiner werden lässt.

Aus Abb. 5.5 sind teilweise deutlich unterschiedliche Verhaltensmuster ersichtlich, auch wenn das zu erwartende Verhalten in den meisten Fällen feststellbar ist. *KJV* und *WBS* zeigen, einen nahezu optimalen Verlauf im *Precision-Recall-Plot* (vgl. Spalten 1 und 4 in Abb. 5.5). Aus Systemsicht ist jedoch aus der 2. und 3. Spalte deutlich ersichtlich, dass die *Text Re-use Analysis* auf stärker modifizierten Texten, wie der *BBE* und der *YLT*, ungleich schlechter ausfällt. In einem vertikalen bzw. spaltenweisen Vergleich wird deutlich, dass sich für alle vier in Abb. 5.5 ausgewählten Bibelversionen zu einer *Precision* P der *Recall* R von *Trigram Shingling* zu *Bigram Shingling* sowie von *Bigram Shingling* zum *Word based Featuring* sukzessive erhöht.

Außerdem muss auf den atypischen Verlauf von *BBE* und teilweise auch *YLT* in den Abb. 5.5 hingewiesen werden. Insbesondere in den Abbildungen 5.5(d), 5.5(h) und 5.5(l) wird deutlich, dass mit steigendem *Scoring*-Schwellwert t zwar, wie zu erwarten ist, der *Recall* R sinkt, jedoch fällt die *Precision* P nach einem anfänglichen Anstieg wieder ab, was auf einen deutlich geringen *Scoring*-Schwellwerte t hinweist, zu welchem das *F-Measure* maximal wird.



(a) *Trigram Shingling* für *WBS* vs. *BBE* (b) *Trigram Shingling* für *WBS* vs. *KJV*

Abbildung 5.6: Vergleich der Evaluierungsmetriken *Precision*, *Recall* und *F-Measure* in Abhängigkeit vom *Scoring*-Schwellwert t. Es wird der Unterschied in F_{max} bei der Analyse von *WBS* vs. *BBE* sowie *WBS* vs. *KJV* verglichen.

Abb. 5.6 bestätigt deutlich, dass für verschiedene Analysen nicht nur unterschiedliche *Featuring*-Techniken[12], sondern auch unterschiedliche *Scoring*-Schwellwerte t zugrunde gelegt werden müssen. Während bei einer Analyse zwischen *WBS* und *BBE* das maximale *F-Measure* mit $F_{max} = 0.374$ bei $t = 0.112$ erreicht wird, ist das *F-Measure* bei der Analyse zwischen *WBS* und *KJV* bei einem *Scoring*-Schwellwert $t = 0.374$ mit $F_{max} = 0.912$ maximal.

5.3.2 Evaluierung durch *Noisy Channel Evaluation*

Die Evaluierung einer *Text Re-use Analysis* hat sich im vorigen Abschnitt als sehr schwierig herausgestellt, da es viele Faktoren und Einflüsse gibt, die auf die Analyse oftmals einwirken.

[11] Der Schwellwert wird hierbei pro Iteration um 0.01 erhöht.
[12] vgl. F_{max} in Tabelle 5.7. Für *KJV* vs. *WBS* wird F beim *Bigram Shingling* maximal.

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	26.5	26.6	26.5	26.8	28.5	28.6	28.7	29.0	34.3	34.6	34.9	35.5
ASV vs. DBY	36.6	37.0	36.8	37.0	39.1	39.4	39.3	39.4	43.0	43.2	43.2	43.3
ASV vs. KJV	40.2	40.4	40.3	40.4	41.8	42.0	42.0	42.0	43.9	44.0	44.0	44.0
ASV vs. WEB	39.6	39.9	39.7	39.7	41.2	41.4	41.3	41.2	43.4	43.5	43.4	43.4
ASV vs. WBS	38.9	39.2	39.1	39.1	41.0	41.2	41.2	41.2	43.7	43.8	43.8	43.9
ASV vs. YLT	26.2	27.1	26.7	26.8	28.6	29.4	29.4	29.6	37.1	37.8	38.5	38.7
BBE vs. ASV	26.5	26.6	26.5	26.8	28.5	28.6	28.7	29.0	34.3	34.6	34.9	35.5
BBE vs. DBY	24.9	25.5	25.1	25.1	26.9	27.2	27.0	27.3	33.0	33.5	33.8	34.3
BBE vs. KJV	24.1	24.8	24.2	24.5	26.6	27.0	26.7	27.0	33.5	34.0	34.4	34.9
BBE vs. WEB	26.5	26.6	26.5	26.8	28.7	28.8	28.8	29.1	35.2	35.5	35.8	36.4
BBE vs. WBS	24.9	25.4	25.0	25.4	27.4	27.8	27.5	27.8	34.3	34.7	35.1	35.8
BBE vs. YLT	20.8	21.7	21.1	21.1	22.7	23.3	23.1	23.1	28.9	29.4	29.9	30.4
DBY vs. ASV	36.6	37.0	36.8	37.0	39.1	39.4	39.3	39.4	43.0	43.2	43.2	43.3
DBY vs. BBE	24.9	25.5	25.1	25.1	26.9	27.2	27.0	27.3	33.0	33.5	33.8	34.3
DBY vs. KJV	35.2	35.6	35.5	35.6	38.0	38.3	38.2	38.4	42.5	42.7	42.7	42.8
DBY vs. WEB	33.1	33.4	33.2	33.5	35.9	36.2	36.1	36.3	41.2	41.5	41.5	41.7
DBY vs. WBS	35.2	35.6	35.4	35.6	38.0	38.4	38.3	38.4	42.6	42.8	42.8	42.9
DBY vs. YLT	25.9	26.5	26.4	26.5	28.2	28.9	29.1	29.4	37.1	37.8	38.7	39.0
KJV vs. ASV	40.2	40.4	40.3	40.4	41.8	42.0	42.0	42.0	43.9	44.0	44.0	44.0
KJV vs. BBE	24.1	24.8	24.2	24.5	26.6	27.0	26.7	27.0	33.5	34.0	34.4	34.9
KJV vs. DBY	35.2	35.6	35.5	35.6	38.0	38.3	38.2	38.4	42.5	42.7	42.7	42.8
KJV vs. WEB	34.4	34.9	34.6	34.7	37.1	37.5	37.3	37.4	41.7	42.0	41.8	41.9
KJV vs. WBS	43.3	43.5	43.4	43.4	44.1	44.1	44.1	44.1	44.5	44.5	44.5	44.5
KJV vs. YLT	25.7	26.4	25.8	25.8	27.9	28.5	28.4	28.5	35.9	36.6	37.3	37.6
WEB vs. ASV	39.6	39.9	39.7	39.7	41.2	41.4	41.3	41.2	43.4	43.5	43.4	43.4
WEB vs. BBE	26.5	26.6	26.5	26.8	28.7	28.8	28.8	29.1	35.2	35.5	35.8	36.4
WEB vs. DBY	33.1	33.4	33.2	33.5	35.9	36.2	36.1	36.3	41.2	41.5	41.5	41.7
WEB vs. KJV	34.4	34.9	34.6	34.7	37.1	37.5	37.3	37.4	41.7	42.0	41.8	41.9
WEB vs. WBS	34.9	35.4	35.1	35.2	37.5	37.9	37.7	37.8	42.1	42.3	42.3	42.4
WEB vs. YLT	24.6	25.7	25.3	25.5	26.6	27.5	27.4	27.7	34.4	35.2	36.2	36.6
WBS vs. ASV	38.9	39.2	39.1	39.1	41.0	41.2	41.2	41.2	43.7	43.8	43.8	43.9
WBS vs. BBE	24.9	25.4	25.0	25.4	27.4	27.8	27.5	27.8	34.3	34.7	35.1	35.8
WBS vs. DBY	35.2	35.6	35.4	35.6	38.0	38.4	38.3	38.4	42.6	42.8	42.8	42.9
WBS vs. KJV	43.3	43.5	43.4	43.4	44.1	44.1	44.1	44.1	44.5	44.5	44.5	44.5
WBS vs. WEB	34.9	35.4	35.1	35.2	37.5	37.9	37.7	37.8	42.1	42.3	42.3	42.4
WBS vs. YLT	26.2	26.7	26.4	26.3	28.4	29.0	29.0	29.1	36.3	37.0	37.7	38.0
YLT vs. ASV	26.2	27.1	26.7	26.8	28.6	29.4	29.4	29.6	37.1	37.8	38.5	38.7
YLT vs. BBE	20.8	21.7	21.1	21.1	22.7	23.3	23.1	23.1	28.9	29.4	29.9	30.4
YLT vs. DBY	25.9	26.5	26.4	26.5	28.2	28.9	29.1	29.4	37.1	37.8	38.7	39.0
YLT vs. KJV	25.7	26.4	25.8	25.8	27.9	28.5	28.4	28.5	35.9	36.6	37.3	37.6
YLT vs. WEB	24.6	25.7	25.3	25.5	26.6	27.5	27.4	27.7	34.4	35.2	36.2	36.6
YLT vs. WBS	26.2	26.7	26.4	26.3	28.4	29.0	29.0	29.1	36.3	37.0	37.7	38.0

Tabelle 5.8: *Modified Noisy Channel Evaluation* in dB für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß der maximalen *Mining Ability* zwischen 0.0 (weiß) und 44.568 dB(schwarz) festgelegt.

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	28.0	28.1	28.0	28.2	29.8	29.9	29.9	30.4	36.4	36.7	36.9	37.3
ASV vs. DBY	37.9	38.2	38.1	38.2	40.5	40.7	40.7	40.8	44.3	44.5	44.5	44.6
ASV vs. KJV	40.4	40.6	40.6	40.6	42.2	42.4	42.3	42.4	45.4	45.5	45.5	45.5
ASV vs. WEB	39.9	40.2	39.9	40.0	41.6	41.8	41.6	41.6	44.2	44.3	44.3	44.3
ASV vs. WBS	39.2	39.5	39.4	39.4	41.4	41.6	41.5	41.6	44.5	44.6	44.6	44.6
ASV vs. YLT	26.6	27.4	27.1	27.2	29.5	30.2	29.9	30.1	40.7	41.1	41.5	41.6
BBE vs. ASV	28.0	28.1	28.0	28.2	29.8	29.9	29.9	30.4	36.4	36.7	36.9	37.3
BBE vs. DBY	25.5	26.1	25.6	25.6	27.7	28.1	27.9	28.1	35.6	36.0	36.3	36.5
BBE vs. KJV	26.6	27.1	26.7	26.9	30.6	30.8	30.6	30.9	38.1	38.4	40.9	41.1
BBE vs. WEB	28.0	28.1	28.0	28.3	29.9	30.1	30.0	30.2	37.7	38.0	38.1	38.6
BBE vs. WBS	32.4	32.5	32.4	32.5	34.8	34.9	34.7	34.9	41.3	41.4	42.0	42.2
BBE vs. YLT	21.2	22.1	21.5	21.4	23.3	23.8	23.6	23.6	30.7	31.1	31.4	31.4
DBY vs. ASV	37.9	38.2	38.1	38.2	40.5	40.7	40.7	40.8	44.3	44.5	44.5	44.6
DBY vs. BBE	25.5	26.1	25.6	25.6	27.7	28.1	27.9	28.1	35.6	36.0	36.3	36.5
DBY vs. KJV	35.5	35.8	35.8	35.9	38.4	38.7	38.6	38.7	43.3	43.5	43.5	43.6
DBY vs. WEB	33.5	33.8	33.6	33.8	36.3	36.6	36.5	36.7	42.1	42.3	42.4	42.5
DBY vs. WBS	35.5	35.8	35.7	35.9	38.4	38.7	38.6	38.8	44.3	44.5	44.5	44.6
DBY vs. YLT	26.2	26.8	26.7	26.8	28.8	29.5	29.5	29.7	38.1	38.8	39.5	39.7
KJV vs. ASV	40.4	40.6	40.6	40.6	42.2	42.4	42.3	42.4	45.4	45.5	45.5	45.5
KJV vs. BBE	26.6	27.1	26.7	26.9	30.6	30.8	30.6	30.9	38.1	38.4	40.9	41.1
KJV vs. DBY	35.5	35.8	35.8	35.9	38.4	38.7	38.6	38.7	43.3	43.5	43.5	43.6
KJV vs. WEB	34.6	35.1	34.8	34.9	37.3	37.7	37.6	37.7	42.4	42.7	42.6	42.7
KJV vs. WBS	43.5	43.7	43.6	43.6	44.4	44.5	44.4	44.5	45.7	45.7	45.7	45.7
KJV vs. YLT	26.0	26.6	26.0	26.1	28.6	29.3	28.8	28.9	36.6	37.4	37.9	38.1
WEB vs. ASV	39.9	40.2	39.9	40.0	41.6	41.8	41.6	41.6	44.2	44.3	44.3	44.3
WEB vs. BBE	28.0	28.1	28.0	28.3	29.9	30.1	30.0	30.2	37.7	38.0	38.1	38.6
WEB vs. DBY	33.5	33.8	33.6	33.8	36.3	36.6	36.5	36.7	42.1	42.3	42.4	42.5
WEB vs. KJV	34.6	35.1	34.8	34.9	37.3	37.7	37.6	37.7	42.4	42.7	42.6	42.7
WEB vs. WBS	35.1	35.5	35.3	35.4	37.9	38.3	38.1	38.2	43.1	43.3	43.3	43.3
WEB vs. YLT	26.4	27.2	26.9	27.1	28.0	28.7	28.7	28.9	35.6	36.3	37.1	37.4
WBS vs. ASV	39.2	39.5	39.4	39.4	41.4	41.6	41.5	41.6	44.5	44.6	44.6	44.6
WBS vs. BBE	32.4	32.5	32.4	32.5	34.8	34.9	34.7	34.9	41.3	41.4	42.0	42.2
WBS vs. DBY	35.5	35.8	35.7	35.9	38.4	38.7	38.6	38.8	44.3	44.5	44.5	44.6
WBS vs. KJV	43.5	43.7	43.6	43.6	44.4	44.5	44.4	44.5	45.7	45.7	45.7	45.7
WBS vs. WEB	35.1	35.5	35.3	35.4	37.9	38.3	38.1	38.2	43.1	43.3	43.3	43.3
WBS vs. YLT	27.5	27.9	27.7	26.7	29.7	30.2	30.0	29.5	37.2	37.8	38.5	38.7
YLT vs. ASV	26.6	27.4	27.1	27.2	29.5	30.2	29.9	30.1	40.7	41.1	41.5	41.6
YLT vs. BBE	21.2	22.1	21.5	21.4	23.3	23.8	23.6	23.6	30.7	31.1	31.4	31.4
YLT vs. DBY	26.2	26.8	26.7	26.8	28.8	29.5	29.5	29.7	38.1	38.8	39.5	39.7
YLT vs. KJV	26.0	26.6	26.0	26.1	28.6	29.3	28.8	28.9	36.6	37.4	37.9	38.1
YLT vs. WEB	26.4	27.2	26.9	27.1	28.0	28.7	28.7	28.9	35.6	36.3	37.1	37.4
YLT vs. WBS	27.5	27.9	27.7	26.7	29.7	30.2	30.0	29.5	37.2	37.8	38.5	38.7

Tabelle 5.9: *Noisy Channel Evaluation* in *dB* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4.

Im Detail zeigte sich, dass besonders bei älteren Texten das *StringSim-Preprocessing* helfen kann, historische Schreibvarianten zu normalisieren. Weiterhin hat sich gezeigt, dass das maximale *F-Measure* F_{max} von der Größe des minimalsten Programmes im Kontext der *Conditional Kolmogorov Complexity* $\mathcal{P}_{min}$ abhängig ist. Liegt wie durch *WBS* nur eine Revision eines Originals vor, dann wird F_{max} mit einem *Shingling* maximal. Ist $\mathcal{P}_{min}$ größer, so hat sich gezeigt, dass bei einem *Word based Featuring* das *F-Measure* maximal wird. Zu guter Letzt zeigte Tabelle 5.6, dass es auch deutliche Abweichungen bei Parametern, wie dem *Scoring*-Schwellwert t, gibt. Im Kontext, dass es sich bei allen Versionen der Bibel um Varianten handelt, die einen gemeinsamen Ursprung haben, ist bereits deutlich ersichtlich, dass es niemals einen *Gold Standard* für eine *Digital Library* geben wird, welcher nicht durch einfaches, versweises Verlinken gleiche *Re-use Units* aufeinander erstellt werden kann.

Aus diesem Grund soll nachfolgend die in Abschnitt 4.4 eingeführte *Noisy Channel Evaluation* auf den gleichen Daten als Evaluierungstechnik angewendet werden. Ziel ist hierbei der Vergleich mit der *Precision*, dem *Recall* sowie des *F-Measure*. Für die *Noisy Channel Evaluation* wird eine *T2*-Randomisierung (vgl. Abschnitt 4.5 ab Seite 149 ff.) durch *Word Shuffling* zugrunde gelegt. Die Wahl dieser Randomisierungsmethode ist damit begründet, dass verschiedene Verteilungseigenschaften, wie das *Zipfsche Gesetz* oder auch der Längenverteilung der Verse, durch den *T2*-Randomisierer erhalten bleiben. Auf der anderen Seite scheinen *T3*-, *T4*- oder *T5*-Randomisierer ungeeignet, da es durch ein gemeinsames Original aller Bibelversionen bei einer *Text Re-use Analysis* um genau dieses Aufdecken der Parallelstellen geht. Vielmehr kann der bereits vielfach erwähnte Zusammenhang zwischen *KJV* und *WBS* als eine *T5*-Randomisierung sowie alle weiteren Bibelversionen als eine modifizierte *T4*-Randomisierung von *KJV* verstanden werden.

Der *T2*-Randomisier wird auf insgesamt 6097057 *Tokens* angewendet. Ganz im Sinne des *Entropie-Tests* ΔH^n aus Formel 4.35 (vgl. Abschnitt 4.6 auf Seite 153) wird während des *Word Shuffling* auf ein Höchstmaß an Zufälligkeit geachtet. Insgesamt werden $n = 181$ Iterationen durchgeführt. Die maximale *Entropie* ist $H_{max} = 22.53968$. H^n beträgt für die $n = 181$ Iterationen $H^{181} = 22.53868$. Der *Entropie-Test* ΔH^n liefert somit für 181 Iterationen einen Wert von $H^{181} = 9.9574 \cdot 10^{-4}$, so dass von einem Randomisierungsfehler ausgegangen werden kann, der kleiner als ein Promille ist.

Die Tabellen 5.8 und 5.9 reflektieren die Ergebnisse der *Noisy Channel Evaluation*. Die berechneten *Scores* in den Tabellen entsprechen der *Mining Ability* $\mathcal{L}_{quant}(\Theta)$ (vgl. Formel 4.22 auf Seite 144).

Die *Modified Noisy Channel Evaluation* aus Tabelle 5.8 ist insofern modifiziert worden, als dass in Formel 4.22 (vgl. Seite 144) die Menge E_{D_S,ϕ_Θ} nicht alle gefundenen Kanten zwischen zwei *Re-use Units* beinhaltet, sondern nur diejenigen, welche auch Teil des *Gold Standards* sind, wodurch die *Modified Noisy Channel Evaluation* einen Bezug zum *Gold Standard* herstellt und somit nicht mehr unabhängig ist. Auf der anderen Seite kann so jedoch auch die *Gold Standard*-abhängige *Modified Noisy Channel Evaluation* mit der evaluierungsbasisunabhängigen *Noisy Channel Evaluation* aus Tabelle 5.9 verglichen werden. Für die *Modified Noisy Channel Evaluation* aus Tabelle 5.8 beträgt somit die maximale *Mining Ability* (vgl. Formel 5.1) $\mathcal{L}^{max}_{modified}(\Theta) = 44.568\ dB$.

$$\mathcal{L}^{max}_{modified}(\Theta) = 10 \cdot \log_{10} 28632 dB \tag{5.1}$$

Tabelle 5.8 reflektiert deutlich das grundsätzliche Verhalten aus dem vorigen Abschnitt. Der höchste *Score* wird für den Vergleich zwischen *KJV* und *WBS* erreicht. Weiterhin werden für *BBE* und *YLT* ebenfalls die geringsten *Evaluierungsscores* bestimmt. Wie auch beim *F-Measure* zeigt sich, dass das Ergebnis für das *Word based Featuring* maximal wird.

Im Vergleich zwischen den Tabellen 5.8 und 5.9 kann festgestellt werden, dass die Er-

gebnisse nahezu identisch sind. Die *Evaluierungsscores* in Tabelle 5.9 sind im Durchschnitt zwischen 1 und 2 *dB* größer. Dies ist in erster Linie darin begründet, dass eine *Gold Standard*-unabhängige Evaluierung aller Daten zugrunde liegt. Der höhere *Score* ist auch gerechtfertigt, da die generierte Evaluierungsbasis nicht vollständig ist. So werden zwar zwischen den Bibeln paarweise gleiche Verse in den *Gold Standard* aufgenommen. Jedoch gibt es auch einen starken *Text Re-use* innerhalb der Bibel, wie der zwischen den Büchern der Apostel (vgl. Abb. 1.2 auf Seite 35), so dass ein Vers durch eine *Text Re-use Analysis* in der einen Bibelversion durch zahlreiche Selbstreferenzen innerhalb der Bibel auf zwei oder auch mehr Verse in der Zielbibel sinnbringend gelinkt werden kann. In einer ganzheitlichen Sicht wird im noch folgenden Abschnitt 5.3.4 gezeigt, dass es eine starke Korrelation zwischen dem *F-Measure* sowie der *Noisy Channel Evaluation* gibt.

5.3.3 Evaluierung durch *Text Re-use Compression*

In Abschnitt 3.10 wurde mit der *Text Re-use Compression* $\mathcal{C}_\Theta$ bereits eine weitere quantitative Evaluierungstechnik eingeführt. Ziel dieses *Scores* ist es, die durch einen *Text Re-use* erzeugte *Redundancy* zu nutzen, um bspw. zwei *Featuring*-Techniken zu vergleichen. Genau wie die *Noisy Channel Evaluation* ist auch die *Text Re-use Compression* eine reine quantitative Evaluierungstechnik, die keinerlei *Gold Standard* bzw. *Evaluierungsbasis* benötigt. Vielmehr bewertet sie das Gesamtergebnis einer *Text Re-use Analysis*.

Wie auch bei der *Noisy Channel Evaluation* werden sowohl das eigentliche Maß, die *Text Re-use Compression*, als auch eine bzgl. des *Gold Standards* modifizierte Version, die *Modified Text Re-use Compression*, miteinander verglichen.

Die *Modified Text Re-use Compression* $\mathcal{C}'_\Theta$ lehnt sich an die *Text Re-use Compression* aus Formel 3.21 (vgl. Seite 125) an. Jedoch werden genau wie bei der *Modified Noisy Channel Evaluation* nicht alle Daten in Betracht gezogen, sondern nur diejenigen, die auch Teil des *Gold Standards* waren (vgl. Formel 5.2).

$$\mathcal{C}'_\Theta = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} \theta_\Theta(s_i, s_j)}{n} \tag{5.2}$$

Die *Modified Text Re-use Compression* aus Formel 5.2 mit $n = 28632$ stellt somit ein Bindeglied beim Vergleich zwischen den klassischen Evaluierungsmaßen mit einem *Gold Standard* sowie der *Text Re-use Compression* dar, die keinerlei qualitative Evaluierungsdaten benötigt.

Die Tabellen 5.10 und 5.11 reflektieren die Ergebnisse der *Modified Text Re-use Compression* $\mathcal{C}'_\Theta$ sowie der *Text Re-use Compression* $\mathcal{C}_\Theta$. In Tabelle 5.11 sind die Werte für $\mathcal{C}_\Theta$, wie auch in Abschnitt 3.10 beschrieben, relativ klein, so dass ein Wert in Tabelle 5.11 von 6.16 für den Vergleich von *ASV* und *BBE* als $10^{-6.16}$ zu interpretieren ist. Somit werden die numerisch größten Werte von $10^{-4.14}$ wieder beim Vergleich von *KJV* mit *WBS* erreicht.

Zwei grundsätzliche Dinge können festgestellt werden. Erstens ist es auffällig, dass sich die numerischen Werte und damit die Einfärbungen zwischen beiden Tabellen deutlich gleichen. Zweitens gibt es eine starke numerische Korrelation zwischen dem *Recall* (vgl. Tabelle 5.6) und den beiden *Compression*-Varianten (vgl. Tabelle 5.10 und 5.10). Nicht nur durch die Einfärbungen der *Scores*, sondern durch das paarweise Vergleichen der Datenpunkte zwischen *Recall* und *Modified Text Re-use Compression* beider Tabellen wird dieser Zusammenhang deutlich. Vielmehr ist nicht nur die Korrelation auffällig, sondern auch, dass die *Modified Text Re-use Compression* numerisch gleich groß oder geringfügig kleiner ist als der *Recall*.

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.07	0.07	0.08	0.09
ASV vs. DBY	0.12	0.13	0.13	0.13	0.21	0.23	0.22	0.23	0.53	0.56	0.56	0.57
ASV vs. KJV	0.29	0.31	0.30	0.30	0.42	0.44	0.43	0.44	0.70	0.72	0.72	0.72
ASV vs. WEB	0.25	0.27	0.26	0.26	0.36	0.38	0.37	0.37	0.61	0.63	0.61	0.62
ASV vs. WBS	0.21	0.23	0.22	0.22	0.34	0.36	0.35	0.35	0.65	0.67	0.67	0.67
ASV vs. YLT	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.12	0.15	0.17	0.18
BBE vs. ASV	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.07	0.07	0.08	0.09
BBE vs. DBY	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.05	0.06	0.06	0.07
BBE vs. KJV	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.06	0.06	0.07	0.08
BBE vs. WEB	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.08	0.09	0.09	0.11
BBE vs. WBS	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.07	0.07	0.08	0.09
BBE vs. YLT	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03
DBY vs. ASV	0.12	0.13	0.13	0.13	0.21	0.23	0.22	0.23	0.53	0.56	0.56	0.57
DBY vs. BBE	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.05	0.06	0.06	0.07
DBY vs. KJV	0.09	0.10	0.09	0.10	0.16	0.17	0.17	0.18	0.46	0.48	0.48	0.49
DBY vs. WEB	0.05	0.06	0.06	0.06	0.10	0.11	0.10	0.11	0.33	0.36	0.36	0.37
DBY vs. WBS	0.09	0.10	0.09	0.10	0.16	0.18	0.17	0.18	0.47	0.50	0.50	0.51
DBY vs. YLT	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.12	0.15	0.18	0.19
KJV vs. ASV	0.29	0.31	0.30	0.30	0.42	0.44	0.43	0.44	0.70	0.72	0.72	0.72
KJV vs. BBE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.06	0.06	0.07	0.08
KJV vs. DBY	0.09	0.10	0.09	0.10	0.16	0.17	0.17	0.18	0.46	0.48	0.48	0.49
KJV vs. WEB	0.07	0.08	0.08	0.08	0.13	0.15	0.14	0.14	0.38	0.41	0.39	0.40
KJV vs. WBS	0.63	0.67	0.65	0.65	0.76	0.78	0.77	0.77	0.89	0.90	0.89	0.90
KJV vs. YLT	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.09	0.11	0.13	0.14
WEB vs. ASV	0.25	0.27	0.26	0.26	0.36	0.38	0.37	0.37	0.61	0.63	0.61	0.62
WEB vs. BBE	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.08	0.09	0.09	0.11
WEB vs. DBY	0.05	0.06	0.06	0.06	0.10	0.11	0.10	0.11	0.33	0.36	0.36	0.37
WEB vs. KJV	0.07	0.08	0.08	0.08	0.13	0.15	0.14	0.14	0.38	0.41	0.39	0.40
WEB vs. WBS	0.08	0.09	0.09	0.09	0.15	0.16	0.15	0.16	0.42	0.45	0.44	0.45
WEB vs. YLT	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.07	0.08	0.10	0.11
WBS vs. ASV	0.21	0.23	0.22	0.22	0.34	0.36	0.35	0.35	0.65	0.67	0.67	0.67
WBS vs. BBE	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.07	0.07	0.08	0.09
WBS vs. DBY	0.09	0.10	0.09	0.10	0.16	0.18	0.17	0.18	0.47	0.50	0.50	0.51
WBS vs. KJV	0.63	0.67	0.65	0.65	0.76	0.78	0.77	0.77	0.89	0.90	0.89	0.90
WBS vs. WEB	0.08	0.09	0.09	0.09	0.15	0.16	0.15	0.16	0.42	0.45	0.44	0.45
WBS vs. YLT	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.12	0.14	0.15
YLT vs. ASV	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.12	0.15	0.17	0.18
YLT vs. BBE	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.03
YLT vs. DBY	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.12	0.15	0.18	0.19
YLT vs. KJV	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.09	0.11	0.13	0.14
YLT vs. WEB	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.02	0.07	0.08	0.10	0.11
YLT vs. WBS	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.10	0.12	0.14	0.15

Tabelle 5.10: *Modified Text Re-use Compression* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4. Die Hintergrundfarbe wird gemäß der maximalen Kompression zwischen 0.0 (weiß) und 1.0 (schwarz) festgelegt.

	Trigram Shingling				Bigram Shingling				Word based Featuring			
	S_{11}	S_{12}	S_{13}	S_{14}	S_{21}	S_{22}	S_{23}	S_{24}	S_{31}	S_{32}	S_{33}	S_{34}
ASV vs. BBE	6.16	6.15	6.16	6.18	6.02	6.01	6.01	5.99	5.42	5.39	5.37	5.33
ASV vs. DBY	5.22	5.19	5.20	5.19	4.98	4.96	4.97	4.95	4.60	4.58	4.58	4.57
ASV vs. KJV	4.97	4.95	4.96	4.95	4.80	4.78	4.79	4.78	4.49	4.47	4.47	4.47
ASV vs. WEB	5.03	5.00	5.02	5.02	4.86	4.84	4.86	4.86	4.60	4.59	4.59	4.59
ASV vs. WBS	5.10	5.07	5.08	5.08	4.89	4.87	4.88	4.87	4.58	4.56	4.56	4.56
ASV vs. YLT	6.34	6.26	6.30	6.29	6.08	6.01	6.05	6.03	5.00	4.95	4.92	4.91
BBE vs. ASV	6.16	6.15	6.16	6.18	6.02	6.01	6.01	5.99	5.42	5.39	5.37	5.33
BBE vs. DBY	6.42	6.36	6.41	6.41	6.24	6.20	6.22	6.20	5.51	5.47	5.44	5.42
BBE vs. KJV	6.35	6.30	6.34	6.32	6.00	5.97	5.99	5.97	5.26	5.23	5.00	4.98
BBE vs. WEB	6.17	6.16	6.17	6.18	6.01	6.00	6.00	6.01	5.30	5.27	5.26	5.22
BBE vs. WBS	5.75	5.74	5.75	5.74	5.55	5.54	5.55	5.54	4.94	4.93	4.83	4.82
BBE vs. YLT	6.86	6.77	6.84	6.85	6.68	6.62	6.66	6.66	5.99	5.94	5.92	5.92
DBY vs. ASV	5.22	5.19	5.20	5.19	4.98	4.96	4.97	4.95	4.60	4.58	4.58	4.57
DBY vs. BBE	6.42	6.36	6.41	6.41	6.24	6.20	6.22	6.20	5.51	5.47	5.44	5.42
DBY vs. KJV	5.49	5.45	5.46	5.44	5.21	5.18	5.19	5.18	4.72	4.70	4.70	4.69
DBY vs. WEB	5.69	5.65	5.67	5.65	5.42	5.39	5.40	5.38	4.85	4.82	4.82	4.80
DBY vs. WBS	5.49	5.45	5.46	5.44	5.21	5.17	5.18	5.17	4.63	4.61	4.61	4.60
DBY vs. YLT	6.38	6.31	6.33	6.32	6.15	6.08	6.09	6.07	5.26	5.19	5.13	5.10
KJV vs. ASV	4.97	4.95	4.96	4.95	4.80	4.78	4.79	4.78	4.49	4.47	4.47	4.47
KJV vs. BBE	6.35	6.30	6.34	6.32	6.00	5.97	5.99	5.97	5.26	5.23	5.00	4.98
KJV vs. DBY	5.49	5.45	5.46	5.44	5.21	5.18	5.19	5.18	4.72	4.70	4.70	4.69
KJV vs. WEB	5.57	5.52	5.55	5.55	5.31	5.27	5.29	5.28	4.81	4.78	4.79	4.78
KJV vs. WBS	4.63	4.61	4.63	4.62	4.55	4.53	4.54	4.54	4.41	4.41	4.41	4.41
KJV vs. YLT	6.39	6.33	6.39	6.39	6.16	6.09	6.15	6.14	5.41	5.33	5.28	5.26
WEB vs. ASV	5.03	5.00	5.02	5.02	4.86	4.84	4.86	4.86	4.60	4.59	4.59	4.59
WEB vs. BBE	6.17	6.16	6.17	6.18	6.01	6.00	6.00	6.01	5.30	5.27	5.26	5.22
WEB vs. DBY	5.69	5.65	5.67	5.65	5.42	5.39	5.40	5.38	4.85	4.82	4.82	4.80
WEB vs. KJV	5.57	5.52	5.55	5.55	5.31	5.27	5.29	5.28	4.81	4.78	4.79	4.78
WEB vs. WBS	5.52	5.48	5.51	5.50	5.26	5.22	5.24	5.23	4.75	4.72	4.73	4.72
WEB vs. YLT	6.38	6.30	6.34	6.33	6.23	6.16	6.17	6.15	5.51	5.44	5.36	5.33
WBS vs. ASV	5.10	5.07	5.08	5.08	4.89	4.87	4.88	4.87	4.58	4.56	4.56	4.56
WBS vs. BBE	5.75	5.74	5.75	5.74	5.55	5.54	5.55	5.54	4.94	4.93	4.83	4.82
WBS vs. DBY	5.49	5.45	5.46	5.44	5.21	5.17	5.18	5.17	4.63	4.61	4.61	4.60
WBS vs. KJV	4.63	4.61	4.63	4.62	4.55	4.53	4.54	4.54	4.41	4.41	4.41	4.41
WBS vs. WEB	5.52	5.48	5.51	5.50	5.26	5.22	5.24	5.23	4.75	4.72	4.73	4.72
WBS vs. YLT	6.25	6.22	6.24	6.34	6.06	6.02	6.04	6.08	5.35	5.29	5.23	5.21
YLT vs. ASV	6.34	6.26	6.30	6.29	6.08	6.01	6.05	6.03	5.00	4.95	4.92	4.91
YLT vs. BBE	6.86	6.77	6.84	6.85	6.68	6.62	6.66	6.66	5.99	5.94	5.92	5.92
YLT vs. DBY	6.38	6.31	6.33	6.32	6.15	6.08	6.09	6.07	5.26	5.19	5.13	5.10
YLT vs. KJV	6.39	6.33	6.39	6.39	6.16	6.09	6.15	6.14	5.41	5.33	5.28	5.26
YLT vs. WEB	6.38	6.30	6.34	6.33	6.23	6.16	6.17	6.15	5.51	5.44	5.36	5.33
YLT vs. WBS	6.25	6.22	6.24	6.34	6.06	6.02	6.04	6.08	5.35	5.29	5.23	5.21

Tabelle 5.11: *Text Re-use Compression* für die *Text Re-use Analysis* zwischen sieben Bibelversionen (vgl. Tabelle 5.3) für insgesamt 12 verschiedene Experimente aus Tabelle 5.4.

Beide Beobachtungen sind nicht zufällig. Die starke Korrelation zwischen der *Modified Text Re-use Compression* und der *Text Re-use Compression* besteht darin, dass der Nenner bei der *Modified Text Re-use Compression* $\mathcal{C}'_\Theta$ n ist, während er bei der *Text Re-use Compression* $\mathcal{C}_\Theta$ (vgl. Formel 3.21 auf Seite 125) $n \cdot (n-1)$ beträgt. Der Unterschied zwischen beiden Maßen besteht somit in einer konstanten Skalierung von $\frac{1}{n-1}$. Dieser Effekt ist durch die Eigenschaft des *Gold Standards* gegeben, dass es für jeden der 28632 Verse, die auch zeitgleich die *Re-use Units* repräsentieren, einen Eintrag im *Gold Standard* gibt. Während bei der *Text Re-use Compression* n die Anzahl der *Re-use Units* beschreibt, entspricht n bei der *Modified Text Re-use Compression* der Anzahl der Datensätze, für die maximal etwas im *Gold Standard*, also 28632, gefunden werden kann. Durch die eben beschriebene Eigenschaft fallen beide Normierungsgrößen aufeinander, so dass eine entsprechende Korrelation auch mathematisch einfach nachzuweisen ist. Würde die *Text Re-use Analysis* nicht auf den 28632 Versen reduzierten Bibelversionen berechnet werden, so wäre das Mitskalieren zwar nach wie vor gegeben, jedoch deutlich schwerer zu belegen. Insofern kann die Wahl der reduzierten Bibelversionen für die Analyse als nicht zufällig gewählt angesehen werden.

Die starke Korrelation zwischen den beiden *Compression*-Techniken und dem *Recall* aus Tabelle 5.6 kann ebenfalls einfach erschlossen werden. Der *Recall* aus Formel 4.17 (vgl. Seite 138) ist im Nenner definiert als $tp + fn$. Gemäß des *Gold Standards* gilt $tp + fn = 28632$, so dass der Nenner des *Recall* sowie der *Modified Text Re-use Compression* (vgl. Formel 5.2 mit $n = 28632$) identisch sind. Im Numerator der *Modified Text Re-use Compression* werden die *Scores* von Broder's *Resemblance* aus der *Scoring*-Matrix T^R (vgl. Formel 3.20 auf Seite 119) mit $\theta_\Theta(s_i, s_j) \in [0, 1]$ so aufsummiert, dass nur die im *Gold Standard* befindlichen Daten den *Score* erhöhen und für alle anderen gefundenen Daten 0 addiert wird. Der *Recall* hingegen entspricht nicht dem Aufsummieren der *Scoring*-Matrix T^R, sondern der *Adjacency Matrix* A (vgl. Formel 2.1 auf Seite 66), wobei ebenfalls nur diejenigen Werte eingerechnet werden, die auch im *Gold Standard* enthalten sind. Da die *Adjacency Matrix* A nur die Werte 0 und 1 enthalten kann, ist auch offensichtlich, dass die Korrelation zwischen dem *Recall* und der *Modified Text Re-use Compression* umso größer wird, je höher der *Scoring*-Schwellwert t gewählt wird.

Da es sowohl eine starke Korrelation zwischen der *Modified Text Re-use Compression* und der *Text Re-use Compression* gibt, die sich nur in einer Skalierung von $\frac{1}{n-1}$ unterscheidet, als auch mathematisch nachvollziehbare Zusammenhänge zwischen der *Modified Text Re-use Compression* und dem *Recall* vorliegen, kann daraus auch ein signifikanter Zusammenhang bzw. eine starke Korrelation zwischen dem *Recall*, welcher einen *Gold Standard* voraussetzt, und der *Text Re-use Compression*, welche frei von einer Evaluierungsbasis funktioniert, hergestellt werden.

Allerdings muss auch darauf hingewiesen werden, dass dieser Zusammenhang nur für *Similarity Measures* (vgl. Abb. 3.9 auf Seite 118) gilt und dort auch vorrangig auf ungewichteten *Scores*. Für *Distance Measures*, für die $dist = 1 - sim$ gilt, ist eine starke negative Korrelation feststellbar.

5.3.4 Zusammenfassung

In den beiden letzten Abschnitten wurden Korrelationen zwischen dem *F-Measure* und der *Noisy Channel Evaluation* sowie dem *Recall* und der *Text Re-use Compression* aufgezeigt. Zusammenfassend für den Abschnitt zur *System Evaluation* sollen die Ergebnisse aus den vorangestellten Ergebnistabellen bzgl. deren statistischer Korrelation untersucht werden.

Hierzu wird Pearson's Korrelationskoeffizient $\rho(X, Y)$, wie in Formel 5.3 abgebildet, benutzt, um den Zusammenhang zwischen den Evaluierungsmaßen auf Basis der Ergebnisse aus der Tabelle 5.5 (*Precision*), Tabelle 5.6 (*Recall*) sowie Tabelle 5.7 (*F-Measure*) für X

sowie der Tabelle 5.8 (*Modified Noisy Channel Evaluation*), Tabelle 5.9 (*Noisy Channel Evaluation*), Tabelle 5.10 (*Modified Text Re-use Compression*) sowie Tabelle 5.11 (*Text Re-use Compression*) für Y zu bestimmen.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \tag{5.3}$$

Tabelle 5.12 reflektiert die Datenanalyse auf Zusammenhänge nach Pearson's Korrelationskoeffizient $\rho(X,Y)$ aus Formel 5.3. Erstens kann festgehalten werden, dass es für *Precision* P zu keinem der quantitativen *Evaluierungsmaße Modified Noisy Channel Evaluation* (NCE_M), *Noisy Channel Evaluation* (NCE), *Modified Text Re-use Compression* (TRC_M) sowie *Text Re-use Compression* (TRC) einen signifikanten statistischen Zusammenhang nach Pearson's Korrelationskoeffizient $\rho(X,Y)$ gibt.

	NCE_M	NCE	TRC_M	TRC
P	0.22095	−0.01466	−0.18785	0.03858
R	0.86054	0.83182	0.99630	0.97751
$F_{\alpha=0.5}$	0.92898	0.88806	0.96331	0.93051

Tabelle 5.12: Pearson's Korrelationskoeffizient $\rho(X;Y)$ zwischen verschiedenen Evaluierungsmetriken.

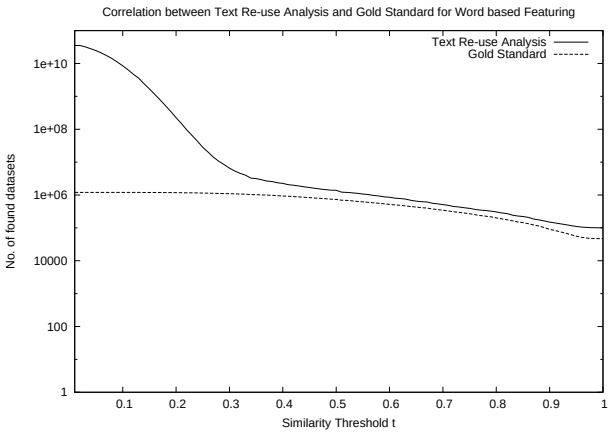
Zweitens kann die starke Korrelation zwischen sowohl dem *Recall* sowie TRC_M und TRC mit Werten von $\rho(R,TRC_M) = 0.99630$ und $\rho(R,TRC) = 0.97751$ durch Pearson's Korrelationskoeffizient auch empirisch belegt werden.

Drittens, ausgehend von NCE sowie NCE_M kann die stärkste Korrelation zu den traditionellen Evaluierungsmetriken P, R und F zum *F-Measure* festgestellt werden. Hierzu wird eine Korrelation von $\rho(F,NCE_M) = 0.92898$ sowie $\rho(F,NCE) = 0.88806$ bestimmt.

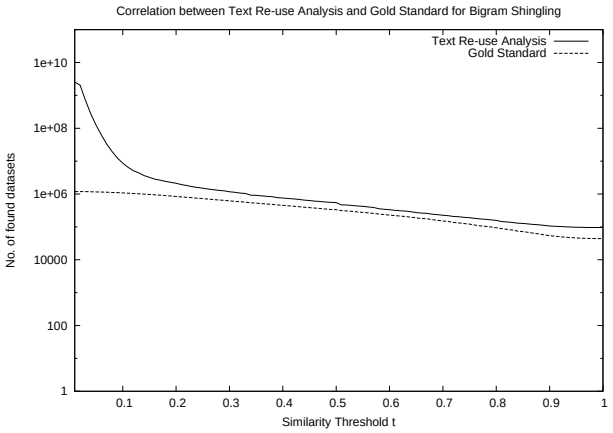
Viertens, mit Ausnahme zur *Precision* P, zu welcher keine signifikante Korrelation nach Pearson festgestellt werden kann, beträgt die Differenz $\Delta\rho(\bullet)$ zwischen der *Modified Noisy Channel Evaluation* und der *Noisy Channel Evaluation* sowie der *Modified Text Re-use Compression* und der *Text Re-use Compression* mit $\Delta\rho(\bullet) = \rho(\bullet,NCE_M) - \rho(\bullet,NCE)$ und $\Delta\rho(\bullet) = \rho(\bullet,TRC_M) - \rho(\bullet,TRC)$ lediglich zwischen 0.02 und 0.04 und ist vielmehr immer positiv, so dass zwar eine geringfügig schlechtere Korrelation erreicht wird, was jedoch ohne einen zugrunde liegenden *Gold Standard* geschieht.

Fünftens, auch wenn TRC und TRC_M stärker auf Basis der Ergebnisse der *System Evaluation* ebenfalls mit F korreliert als NCE_M und NCE, so muss auf das Verhältnis zwischen R und F in Tabelle 5.12 (vgl. 2. und 3. Zeile) hingewiesen werden. Für TRC_M und TRC wird die größte Korrelation zum *Recall* R bestimmt. Um 0.03 bzw. 0.04 kleiner ist die Korrelation zum *F-Measure*. Dem entgegengesetzt ist die größte Korrelation von NCE_M und NCE mit dem *F-Measure* zu beobachten, während die Korrelation mit dem *Recall* R etwa 0.06 in beiden Fällen kleiner ist (vgl. Tabelle 5.12).

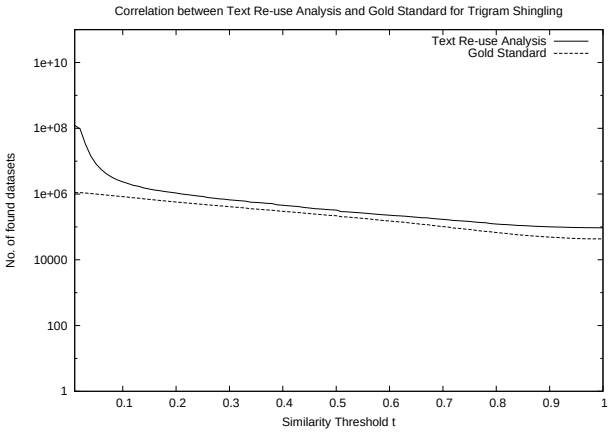
In Abschnitt 5.3.1 wurde bereits dargelegt, dass durch die vergleichsweise konstant hohe *Precision* P im Vergleich zur Streuung des *Recall* R eine signifikante Korrelation zwischen R und F induziert wird. Dies spiegelt sich auch in den Ergebnissen in Tabelle 5.12 wider. Aus diesem Grund soll der Zusammenhang zwischen *F-Measure* und *Noisy Channel Evaluation* sowie *Recall* und *Text Re-use Compression* in einer anderen Darstellung eruiert werden. Hierzu werden die Ergebnisse nicht mehr zu einem konstanten *Scoring*-Schwellwert t betrachtet, sondern t wird als Abhängige aufgefasst.



(a) Word based Featuring



(b) *Bigram Shingling*



(c) *Trigram Shingling*

Abbildung 5.7: Verlauf des gefundenen *Text Re-use* in Abhängigkeit zum *Scoring*-Schwellwert t. Sowohl für das *Trigram* als auch *Bigram Shingling* sowie das *Word based Featuring* wird der insgesamt gefundene sowie der sich mit einem *Gold Standard* überlappende *Text Re-use* abgebildet.

In Abb. 5.7 ist der Verlauf des insgesamt gefundenen (in der Abbildung *Text Re-use Analysis* genannt) sowie des zum *Gold Standard* aufgedeckten *Text Re-use* bzgl. unterschiedlicher *Scoring*-Schwellwerte t abgebildet. Der *Recall* R normalisiert den in Abb. 5.7 mit *Gold Standard* markierten Verlauf durch die Menge aller möglichen in der Evaluierungsbasis enthaltenen Datensätze. Für diese Analysen und damit auch für Abb. 5.7 geltend, wird nicht mehr jeder Bibelvergleich separat betrachtet, wodurch die maximale Menge an zu findenden Datensätzen 1202544 beträgt[13]. Die *Text Re-use Compression* wird hingegen durch $n \cdot (n-1)$ normalisiert. Da sowohl für den *Recall* R als auch die *Text Re-use Compression* der Nenner konstant bzgl. der Variablen des *Scoring*-Schwellwertes t ist, wird die Normalisierung an dieser Stelle vernachlässigt, wodurch ein direkter Vergleich der Daten bzgl. ihrer y-Achse besser gegeben ist. Abb. 5.7 reflektiert genau diese Ergebnisse.

Es wird in Abb. 5.7 die Korrelation zwischen *Text Re-use Compression* und dem *Recall* nicht nur, wie in Tabelle 5.12, zu einem konstanten *Scoring-Schwellwert* t, sondern insbesondere auch die Korrelation im Verhalten zu einem dynamischen t deutlich. Es entspricht der natürlichen Erwartung, dass der *Recall* R mit zunehmendem *Scoring*-Schwellwert t sukzessive abnimmt. Auch die *Text Re-use Compression* folgt diesem Verhalten deutlich, wobei in beiden Fällen darauf hingewiesen werden muss, dass in Abb. 5.7 die y-Achse logarithmisch skaliert ist, wodurch der negative Anstieg kleiner scheint als dieser real ist.

Im Detail zeigt sich für alle drei *Featuring*-Techniken aus Abb. 5.7 eine Abweichung bei einem kleineren *Scoring*-Schwellwert t. Aus diesem Grund wird in Tabelle 5.13 der Zusammenhang zwischen *Recall* R und *Text Re-use Compression* zu unterschiedlichen Schwellwerten t betrachtet.

	$t = 0.0$	$t = 0.1$	$t = 0.2$	$t = 0.3$	$t = 0.4$
Word based Featuring	0.41801	0.35972	0.415979	0.86754	0.98144
Bigram Shingling	0.39993	0.83656	0.98122	0.99274	0.99415
Trigram Shingling	0.53019	0.96880	0.99220	0.99536	0.99750

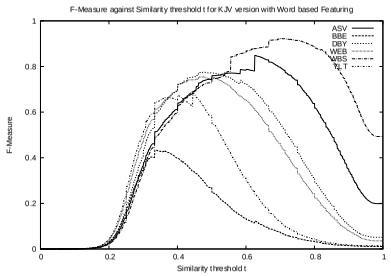
Tabelle 5.13: Pearson's Korrelationskoeffizient $\rho(X;Y)$ zwischen *Recall* R und *Text Re-use Compression* bzgl. unterschiedlicher Schwellwerte t sowie für drei *Featuring*-Techniken.

Tabelle 5.13 reflektiert zwei Aspekte. Erstens, es kann gezeigt werden, dass bereits bei kleinen *Scoring*-Schwellwerten von $t = 0.4$ eine Korrelation von $\rho_{t=0.4}(R, TRC) \geq 0.98$ erreicht werden kann. Zweitens, es zeigt sich auch hier wieder, unabhängig von verschiedenen *Bibeln*, dass sich insbesondere bei niedrigen Schwellwerten t, bspw. $t = 0.1$, das Ergebnis teilweise deutlich bei der Wahl einer anderen *Featuring*-Methode unterscheidet.

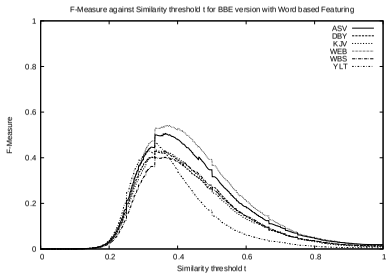
In Anlehnung an die berechneten Korrelationen nach Pearson (vgl. Tabelle 5.12) soll nachfolgend auch der Zusammenhang zwischen dem *F-Measure* F und der *Mining Ability* $\mathcal{L}(\Theta)$ der *Noisy Channel Evaluation* verifiziert werden. Genau wie beim Vergleich zuvor zwischen dem *Recall* R und der *Text Re-use Compression* wird das Verhalten des *F-Measures* sowie der *Noisy Channel Evaluation* in Abhängigkeit vom *Scoring*-Schwellwert t betrachtet. Die Abb. 5.8 und 5.9 reflektieren die entsprechenden Analysen mit den gleichen Einstellungen wie bei der vorigen Analyse zwischen *Recall* und *Text Re-use Compression*.

Erstens, im direkten Vergleich der entsprechenden Verläufe wird ein ähnliches Verhalten ersichtlich. Sowohl das *F-Measure* als auch die *Noisy Channel Evaluation* haben genau ein globales Maximum, welches sich deutlich von $t = 0.0$ bzw. $t = 1.0$ unterscheidet, die dem globalen Maximum des *Recall* R (vgl. Tabelle 5.7) sowie der *Precision* P entsprechen.

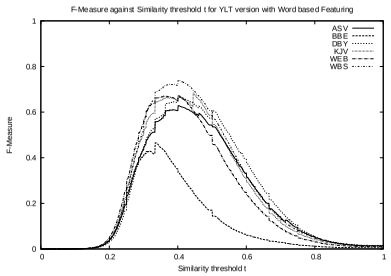
[13] Es gilt $1202544 = 42 \cdot 28632$. 42 repräsentiert die insgesamt 42 paarweisen Vergleiche der sieben Bibelversionen mit den sechs anderen Editionen.



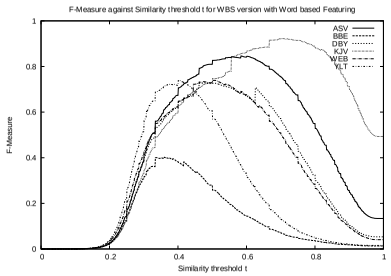
(a) KJV mit *Word based Featuring*



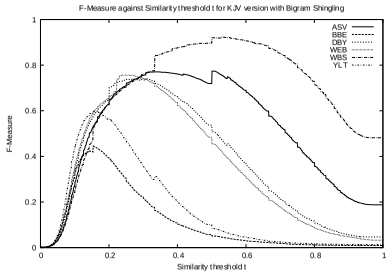
(b) BBE mit *Word based Featuring*



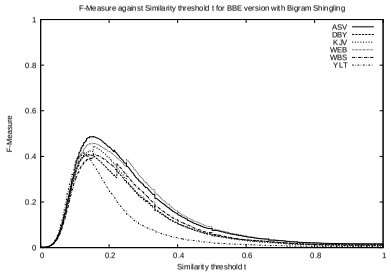
(c) YLT mit *Word based Featuring*



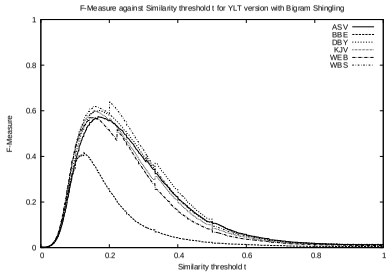
(d) WBS mit *Word based Featuring*



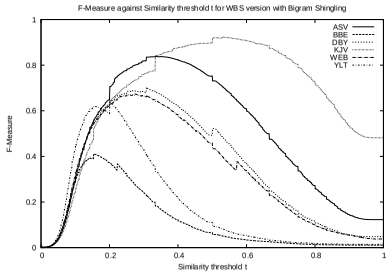
(e) KJV mit *Bigram Shingling*



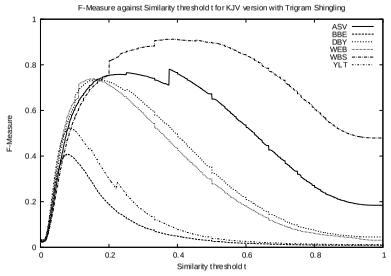
(f) BBE mit *Bigram Shingling*



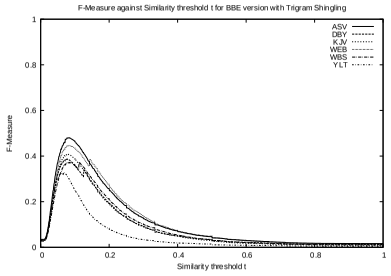
(g) YLT mit *Bigram Shingling*



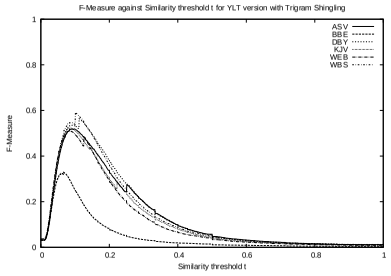
(h) WBS mit *Bigram Shingling*



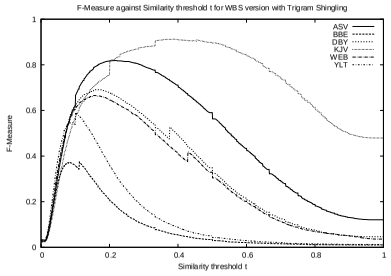
(i) KJV mit *Trigram Shingling*



(j) BBE mit *Trigram Shingling*

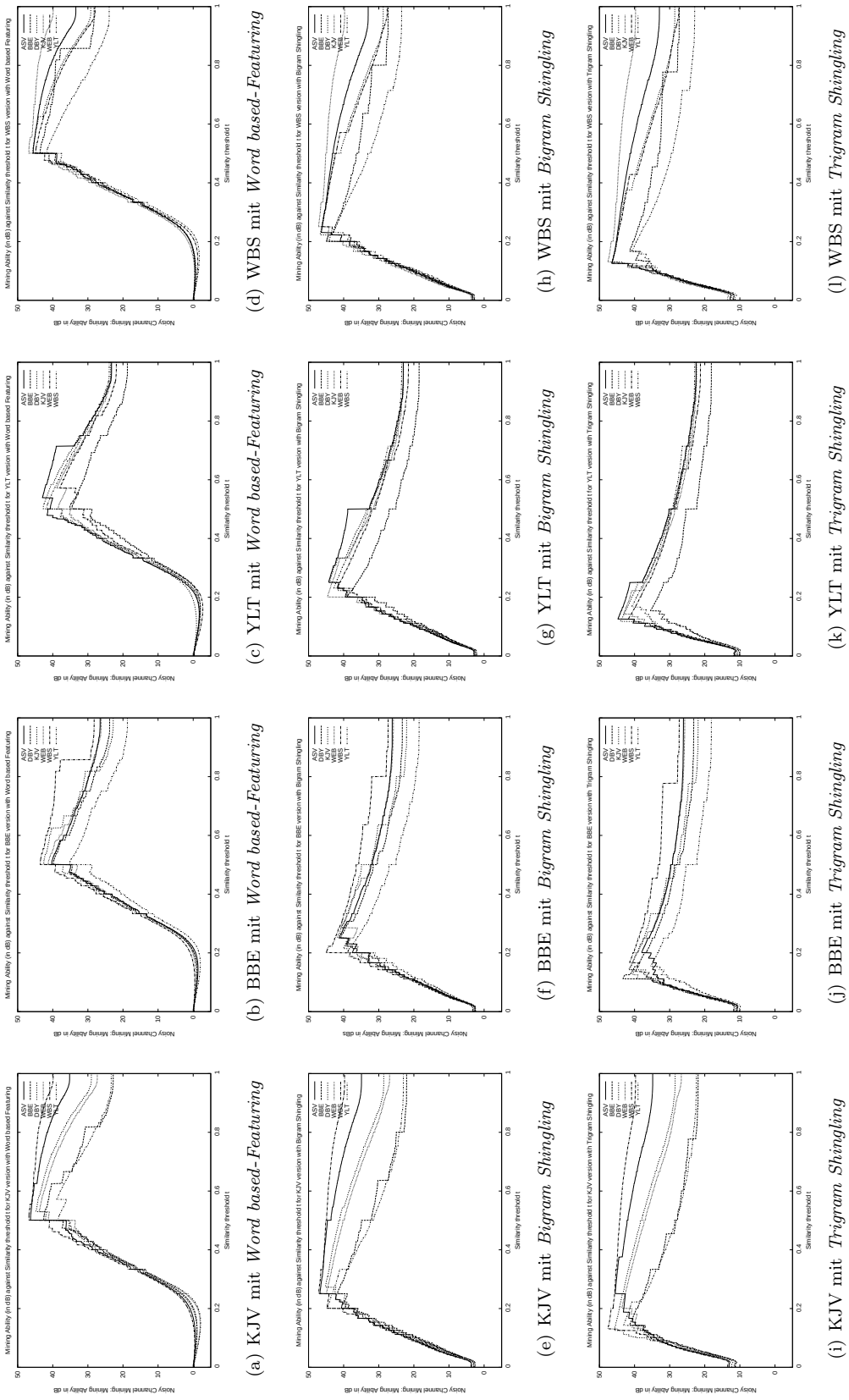


(k) YLT mit *Trigram Shingling*



(l) WBS mit *Trigram Shingling*

Abbildung 5.8: *F-Measure* F in Abhängigkeit vom *Scoring*-Schwellwert t. In den Abbildungen wird das Verhalten des *F-Measure* gegen t für das *Trigram Shingling*, *Bigram Shingling* sowie für vier ausgewählte Bibelversionen dargestellt. In jedem Plot wird das Verhalten des *F-Measure* bzgl. der anderen sechs Bibelversionen aufgezeigt.



(a) KJV mit *Word based-Featuring* (b) BBE mit *Word based-Featuring* (c) YLT mit *Word based-Featuring* (d) WBS mit *Word based-Featuring*

(e) KJV mit *Bigram Shingling* (f) BBE mit *Bigram Shingling* (g) YLT mit *Bigram Shingling* (h) WBS mit *Bigram Shingling*

(i) KJV mit *Trigram Shingling* (j) BBE mit *Trigram Shingling* (k) YLT mit *Trigram Shingling* (l) WBS mit *Trigram Shingling*

Abbildung 5.9: *Noisy Channel Evaluation* in Abhängigkeit vom *Scoring*-Schwellwert t. In den Abbildungen wird das Verhalten der *Mining Ability* $\mathcal{L}(\Theta))$ in dB gegen t für das *Trigram Shingling*, *Bigram Shingling* sowie für vier ausgewählte Bibelversionen dargestellt. In jedem Plot wird das Verhalten der *Mining Ability* bzgl. der anderen sechs Bibelversionen aufgezeigt.

Zweitens, das *F-Measure* in Abhängigkeit vom *Scoring*-Schwellwert t kann in zwei Domänen eingeteilt werden, die durch das globale Maximum von einander getrennt sind (vgl. Abb. 5.8). Während der *Recall* R in der Domäne der niedrigen Schwellwerte t dominant ist, wird die Domäne der höheren Schwellwerte t von der *Precision* P vorrangig beherrscht. Auch für die *Noisy Channel Evaluation* existieren diese beiden Domänen, wobei die Domäne der niedrigen Schwellwerte vom statistischen Grundrauschen eines *Mining*-Verfahrens bestimmt ist und die Domäne der höheren Schwellwerte von der quantitativen Genauigkeit.

Drittens, sowohl für das *F-Measure* als auch die *Noisy Channel Evaluation* kann in der Domäne der niedrigen *Scoring*-Schwellwerte t ein gebündelter Anstieg zum globalen Maximum beobachtet werden. Dies ist sowohl von der Wahl der *Featuring*-Technik als auch der unterschiedlichen Bibelversionen unabhängig. Nach dem Maximum streuen die Plots innerhalb eines Verlaufes in Abb. 5.9 unterschiedlich stark.

Viertens, die Streuung von *BBE* und *YLT* (vgl. 2. und 3. Spalte in den Abbildungen 5.8 und 5.9) ist sowohl für das *F-Measure* als auch die *Mining Ability* deutlich geringer als bei *KJV* und *WBS* (vgl. 1. und 4. Spalte). Insgesamt fällt die Streuung bei der *Noisy Channel Evaluation* in Abb. 5.9 geringer aus als beim *F-Measure*. Das ist der logarithmischen Skalierung der *Mining Ability* aus Formel 4.22 (vgl. Seite 144) geschuldet.

Fünftens, eine der grundlegenden Fragen dieser Arbeit ist, ob im Kontext der *Data Diversity* ein *Gold Standard* für die Evaluierung benötigt wird, da ebendieser oftmals sehr schwer vor allem unter Berücksichtigung der Repräsentativität zu erstellen ist. Neben dem bereits aufgezeigten Zusammenhang zwischen dem *Recall* R und der *Text Re-use Compression* zeigen die Abb. 5.8 und 5.9 ebenfalls auf, dass auch in Abwesenheit eines *Gold Standards* die *Mining Ability* der *Noisy Channel Evaluation* eingesetzt werden kann. Bei der Evaluierung von Sprachmodellen kommt es oftmals gar nicht darauf an, genau sagen zu müssen, dass die *Precision* und der *Recall* einen bestimmten Wert einnehmen, sondern diese *Scores* werden meistens nur dazu verwendet, unterschiedliche Verfahren und Parameter zu vergleichen und entsprechende Verbesserungen festzustellen. Aus diesem Grund soll das *F-Measure* und die *Noisy Channel Evaluation* auf deren Aussagekraft nicht nur im Sinne des *Korrelationskoeffizienten* untersucht werden, sondern auch auf deren Gleichheit in der Reihenfolge der Ergebnisse. Aus den Abbildungen 5.8 und 5.9 können so für *KJV* (vgl. 1. Spalte in den Abbildungen) die folgenden Rangfolgen

F-Measure: *WBS*, *ASV*, *DBY*, *WEB*, *YLT*, *BBE*
NCE: *WBS*, *ASV*, *DBY*, *WEB*, *BBE*, *YLT*

in der zweiten Domäne der höheren *Scoring*-Schwellwerte t, bspw. $t = 0.5$ oder $t = 0.6$, erstellt werden. Neben einer starken Überlappung in der Reihenfolge sind zwei weitere Dinge grundsätzlich auffällig. Auf der einen Seite sind die Abstände zwischen den Plots vergleichbar. Während sich *WBS* deutlich von den anderen Plots abhebt und von *ASV* gefolgt wird, liegen die Verläufe von *DBY* und *WEB* sowohl beim *F-Measure* als auch bei der *Noisy Channel Evaluation* dicht beieinander. *YLT* und *BBE* fallen ebenfalls in beiden Fällen deutlich schneller ab, als bei den anderen Bibelversionen. Auf der anderen Seite muss auch festgestellt werden, dass *YLT* und *BBE* bei einer Evaluierung nach *F-Measure* ihre Positionen tauschen. Dies kann einfach damit begründet werden, dass die Evaluierung mit einem *Gold Standard BBE* insofern benachteiligt ist, dass immer nur die Datensätze evaluiert werden, die dem gleichen Vers entsprechen. Durch den hohen *Text Re-use* innerhalb der Bibel, bspw. durch die Bücher der Apostel Markus und Lukas (vgl. Abb. 1.2 auf Seite 35), kann das Ergebnis der *Noisy Channel Evaluation* abweichen, da nicht nur die Daten des *Gold Standards*, sondern alle Daten betrachtet werden. Aufgrund der einfachen Sprache in

BBE fällt das Ergebnis von *BBE* gegenüber *YLT* bei der *Noisy Channel Evaluation* besser aus, so dass der Rang getauscht wird. Weiterhin zeigt sich auch, dass die Reihenfolgen in diesem Fall unabhängig von der *Featuring*-Technik sind. Auch für *WBS* (vgl. 4. Spalte in Abbildungen 5.8 und 5.9) kann Ähnliches festgestellt werden.

Für *YLT* und *BBE* ist nicht nur beobachtbar, dass die Streuung nach dem globalen Maximum deutlich geringer ist, sondern auch, dass *YLT* im Plot der *BBE* (vgl. 2. Spalte) als Ausreißer dieser geringen Streuung anzusehen ist. Gleiches gilt für *BBE* im Plot von *YLT* (vgl. 3. Spalte).

Sechstens kann festgestellt werden, dass sich das globale Maximum des *F-Measure* F_{max} (vgl. Abb. 5.8) sowie der *Noisy Channel Evaluation* (vgl. Abb. 5.9) bzgl. der *Featuring*-Techniken *Trigram Shingling*, *Bigram Shingling* sowie *Word based Featuring* bei kleiner werdender Größe der Atome sukzessive zu einem höheren *Scoring*-Schwellwert t bewegt.

Dieser Effekt soll nachfolgend in den Tabellen 5.14 bis 5.16 auch für weitere *Preprocessing*-Techniken aufgezeigt werden. Hierbei wird ξ_i, wie auf Seite 177 bereits eingeführt, benutzt.

	ASV	BBE	DBY	WEB	WBS	YLT
ξ_1	(0.38, 0.77)	(0.08, 0.40)	(0.15, 0.74)	(0.15, 0.74)	(0.37, 0.91)	(0.08, 0.52)
ξ_2	(0.38, 0.79)	(0.08, 0.40)	(0.16, 0.74)	(0.15, 0.74)	(0.40, 0.92)	(0.08, 0.53)
ξ_3	(0.38, 0.78)	(0.08, 0.41)	(0.15, 0.74)	(0.15, 0.74)	(0.39, 0.91)	(0.08, 0.52)
ξ_4	(0.38, 0.78)	(0.08, 0.42)	(0.16, 0.74)	(0.15, 0.74)	(0.39, 0.91)	(0.09, 0.53)

Tabelle 5.14: (t, F_{max})-Tupel für *Trigram Shingling* bei vier *Preprocessing*-Techniken im Vergleich von *KJV* zu den anderen Bibelversionen.

	ASV	BBE	DBY	WEB	WBS	YLT
ξ_1	(0.32, 0.76)	(0.15, 0.43)	(0.26, 0.73)	(0.24, 0.75)	(0.53, 0.92)	(0.15, 0.57)
ξ_2	(0.51, 0.77)	(0.15, 0.43)	(0.27, 0.74)	(0.24, 0.76)	(0.53, 0.92)	(0.15, 0.59)
ξ_3	(0.51, 0.77)	(0.15, 0.45)	(0.31, 0.74)	(0.24, 0.76)	(0.53, 0.92)	(0.16, 0.60)
ξ_4	(0.51, 0.78)	(0.15, 0.45)	(0.31, 0.74)	(0.24, 0.76)	(0.53, 0.91)	(0.16, 0.60)

Tabelle 5.15: (t, F_{max})-Tupel für *Bigram Shingling* bei vier *Preprocessing*-Techniken im Vergleich von *KJV* zu den anderen Bibelversionen.

	ASV	BBE	DBY	WEB	WBS	YLT
ξ_1	(0.63, 0.84)	(0.34, 0.44)	(0.46, 0.77)	(0.46, 0.75)	(0.70, 0.92)	(0.34, 0.64)
ξ_2	(0.63, 0.85)	(0.34, 0.45)	(0.48, 0.78)	(0.46, 0.76)	(0.70, 0.92)	(0.36, 0.65)
ξ_3	(0.63, 0.85)	(0.34, 0.43)	(0.48, 0.78)	(0.46, 0.76)	(0.70, 0.92)	(0.44, 0.68)
ξ_4	(0.63, 0.85)	(0.36, 0.44)	(0.48, 0.78)	(0.47, 0.76)	(0.70, 0.92)	(0.44, 0.70)

Tabelle 5.16: (t, F_{max})-Tupel für *Word based Featuring* bei vier *Preprocessing*-Techniken im Vergleich von *KJV* zu den anderen Bibelversionen.

Die Tabellen 5.14 bis 5.16 reflektieren die Ergebnisse für die drei eingesetzten *Featuring*-Techniken sowie für die vier eingesetzten *Preprocessing*-Methoden. Die enthaltenen Tupel (x, y) repräsentieren mit $x = t_{max}$ sowie $y = F_{max}$ die Ergebnisse der Analyse des maximalen *F-Measure* F_{max}.

Die Tabellen 5.14 bis 5.16 zeigen drei Sachverhalte auf. Erstens wird deutlich, dass in einer vertikalen Analyse über alle drei Tabellen mit kleiner werdender Größe der *Atome* F_{max} sukzessive steigt. Während für bspw. *ASV* beim *Trigram Shingling* das *F-Measure* mit etwa $F_{max} = 0.78$ bei einem *Scoring*-Schwellwert t von $t_{max} = 0.38$ maximal wird, so ist $t_{max} = 0.51$ bzw. $t_{max} = 0.63$ für das *Bigram Shingling* bzw. das *Word based Featuring*. Ein ähnliches Verhalten ist bei allen anderen Vergleichen von Bibelversionen mit *KJV* aus den Tabellen 5.14 bis 5.16 ablesbar. Zweitens kann in einer zeilenweisen Analyse wiederum beobachtet werden, dass es eine große Streuung von t_{max} für Vergleiche zwischen unterschiedlichen Bibelversionen gibt. Während t_{max} für den Vergleich von *KJV* mit *WBS* immer am größten ist, liegt t_{max} für den Vergleich *KJV* mit *BBE* im Schnitt etwa um 0.3 niedriger. Drittens, es muss aber auch festgestellt werden, dass das *Preprocessing* ξ_i nahezu keinen Einfluss auf zumindest das Maximum des *F-Measure* hat. Die Tabellen 5.14 bis 5.16 reflektieren das bis auf zwei Ausnahmen deutlich. Auf der einen Seite kann in Tabelle 5.15 für das *Bigram Shingling* ein signifikanter Unterschied zwischen ξ_1 mit $t_{max} = 0.32$ sowie ξ_2, ξ_3 und ξ_4 mit jeweils $t_{max} = 0.51$ beim Vergleichen der Bibelversionen *KJV* mit *ASV* ausgemacht werden (vgl. 1. Spalte). Auf der anderen Seite zeigt sich beim *Word based Featuring* (vgl. Tabelle 5.16), dass es beim Vergleich zwischen *KJV* und *YLT* einen deutlichen Unterschied gibt, ob eine *Lemmatisation* wie in ξ_3 und ξ_4 durchgeführt wird oder wie in ξ_1 und ξ_2 nicht. Tabelle 5.16 zeigt hierfür eine Differenz von F_{max} von 0.1 bzw. 0.08 auf (vgl. letzte Spalte).

Zusammenfassend zeigt die *System Evaluation* die *Diversity* auch bei den Ergebnissen auf. Ausgehend von sieben verschiedenen englischsprachigen Bibelversionen, zu welchen sicher angenommen werden kann, dass sie alle einen gemeinsamen Ursprung haben, hat sich gezeigt, dass sich durch editorspezifische Motivationen für die Erstellung einer neuen Bibelversion bereits hinreichende Veränderungen im Sinne des *Noisy Channel* ergeben, so dass sich eine große Varianz an Möglichkeiten ergibt. Weiterhin ist es ein Ergebnis, dass die Auswirkungen des *Preprocessing* vom *Scoring*-Schwellwert t abhängig ist. Je höher t ist, desto größer ist der Einfluss des *Preprocessing*. Weiterhin hat sich gezeigt, dass keine Annahme über einen guten *Scoring*-Schwellwert t gemacht werden kann. Je nach Größe des minimalsten Programmes $\mathcal{P}_{min}$ im Sinne der *Conditional Kolmogorov Complexity*, welches den Regeln entspricht, einen Vers in einen anderen Vers einer anderen Bibelversion zu transformieren, gibt es teilweise deutliche Unterschiede bei dem zu wählenden Schwellwert t. Zusätzlich ist es ebenfalls ein Ergebnis, dass nicht per se eine Annahme über die zu wählende *Featuring*-Technik getroffen werden kann.

Aus genau dieser *Diversity* heraus stellt sich die Frage nach der Evaluierungsmöglichkeit. Das in diesem Abschnitt dargelegte Szenario mit einem gemeinsamen Ursprung zeigt bereits vielschichtige Probleme mit der Erstellung eines *Gold Standards* auf. Aus diesem Grund wurden die rein quantitativen Maße der *Text Re-use Compression* sowie der *Noisy Channel Evaluation* vorgestellt und getestet, welche ein *Mining*-Ergebnis als Ganzes und nicht nur eine überlappende Teilmenge aus Ergebnis und Testgrundlage evaluiert. Im Ergebnis kann festgehalten werden, dass die *Text Re-use Compression* den *Recall* R nahezu perfekt ersetzen kann. Des Weiteren konnte dargestellt werden, dass das *F-Measure* in seinen Grundeigenschaften durch die *Noisy Channel Evaluation* mit dem *Score* der *Mining Ability* so ausgetauscht werden kann, dass es keinen *Gold Standard* bzw. keine qualifizierte *Evaluierungsbasis* benötigt. Einzig für die *Precision* konnte kein adäquates Pendant gefunden werden. Aus diesem Grund kann an dieser Stelle auf Basis der Erfahrungen aus Abschnitt 5.2 zum *Text Re-use* auf der *Perseus Digital Library* darauf verwiesen werden, dass eine manuelle Evaluierung einer zufälligen *Stichprobe* der Ergebnisse einer *Text Re-use Analysis* bereits sehr gut einen fehlenden *Gold Standard* ersetzt. Somit kann durch die *Text Re-use Compression*, die *Noisy Channel Evaluation* sowie einer manuellen Überprüfung einer

Stichprobe eine Evaluierungsmöglichkeit geschaffen werden, welcher keine Evaluierungsbasis zugrunde liegt, was, wie eingangs bereits beschrieben wurde, durch die *Diversity* faktisch nicht in einem guten Maße umsetzbar wäre.

5.4 Component & Aggregated Evaluation

In der *Text Re-use Analysis* zur *System Evaluation* aus Abschnitt 5.3 wurden einige Level, wie das *Selection*, und Parameter der *7-Level-Architektur* des *Historical Text Re-use* als konstant angenommen. Für die Level und Parameter, wie dem *Scoring*-Schwellwert t, die hingegen als nicht konstant angenommen wurden, haben sich vielschichtige Unterschiede ergeben.

Ziel des folgenden Abschnittes ist es, die als fix angenommenen Level, wie das *Selection* und das *Linking*, ebenfalls genauer zu betrachten. Das soll erreicht werden, indem die klassische *System Evaluation* aufgegeben wird und ein Level allein bzw. in einer Aggregation von zwei oder maximal drei Level getestet wird. Daher wird diese Form der Evaluierung auch *Component* bzw. *Aggregated Evaluation* genannt. Die grundlegende Frage hierbei ist, wie die *Text Re-use Analysis* auf den sieben Versionen der Bibel aus Abschnitt 5.3 verbessert werden kann. Aus der *System Evaluation* wird zwar bspw. der Trend zu einem *Word Based Featuring* deutlich. Jedoch ergibt sich aus dieser Form der Evaluierung nicht, wie die Analyse verbessert werden kann. Vielmehr endet der Versuch des Optimierens einer *Text Re-use Analysis* durch eine *System Evaluation* mit *Precision*, *Recall* und *F-Measure* oftmals im *Trial & Error*.

Die in diesem Abschnitt vorgestellte *Component* bzw. *Aggregated Evaluation* basiert auf den Erkenntnissen aus der *Biometrie*. Ein biometrisches System, wie das Erkennen und Vergleichen eines menschlichen Fingerabdruckes (vgl. [Maltoni 2009]), wird in der forensischen Analyse eingesetzt, so dass ein Höchstmaß an Genauigkeit erreicht werden muss. Hierzu wird das *System* während der *Test- und Entwicklungsphase* nicht ganzheitlich evaluiert, sondern es gibt etwa 20 - 30 verschiedene *Error Rates* (vgl. die *Biometry Glossaries* in [BIMA 2012, NSTC 2006]), die einen spezifischen Teil des Gesamtsystems evaluieren. So wird einerseits der Scanner separat betrachtet und insbesondere auch auf seine Fehleranfälligkeit bei Rauschen, wie es durch Schmutz und Schweiß gegeben ist, untersucht (vgl. *T5*-Randomisierer der *Noisy Channel Evaluation* aus Abschnitt 4.5). Auf der anderen Seite wird das *Feature Extraction* ebenfalls separat betrachtet, wobei der Fokus auf der Erkennungsqualität der *Minutiae*, also den Haupterkennungsmerkmalen eines menschlichen Fingerabdruckes, liegt (vgl. Abbildung 3.4 auf Seite 99 in [Maltoni 2009]).

Die nachfolgende *Component Evaluation* hat zwei Ziele. Erstens, es soll aufgezeigt werden, wie die *Text Re-use Analysis* verbessert werden kann. Zweitens, anhand der Qualitätskriterien für *Text Mining*, wie der *Performance* (vgl. Abschnitt 2.4 auf Seite 62), wird evaluiert, wie eine *Text Re-use Analysis* möglichst ohne Verlust an Ergebnissen bzgl. der Geschwindigkeit optimiert werden kann. Dies ist u. a. deswegen nicht zu vernachlässigen, da eine Analyse durch ein *Intra Digital Library Linking* immer von quadratischer Komplexität $O(n^2)$ ist (vgl. Abschnitt 3.6 auf Seite 113).

In diesem Abschnitt werden zunächst die *Lemmatisation* (vgl. Abschnitt 5.4.1), der Umgang mit paradigmatischen Relation (vgl. Abschnitt 5.4.2) sowie die Fähigkeit im Umgang mit historischen Schreibweisen (vgl. Abschnitt 5.4.3) evaluiert. Dem schließt sich eine *Evaluierung* der *Digital Signature* an. Abgeschlossen wird der Abschnitt von einer *Linking*-Analyse, bei der die *Performance* einer *Text Re-use Analysis* im Vordergrund steht.

5.4.1 Qualität der *Lemmatisation*

Lemmatisierung von historischen Schreibweisen ist ein in den Geisteswissenschaften seit Jahrhunderten bzw. Jahrzehnten mit unterschiedlichen Methoden andauernder Prozess. Doch warum ist die Lemmatisierung auf historischen Text schwierig?

Bis auf wenige Ausnahmen, wie die Lemmatisierung von irregulären Verben, besteht die *Lemmatisation* in der Behandlung des Wortsuffixes, einem *Suffix Stripping*, wie es Porter für den bekanntesten Lemmatisierungsansatzes nennt (vgl. [Porter 1980]). Bei der Lemmatisierung wird zweigeteilt vorgegangen. Erstens, es wird die Wortart, wie Substantiv oder Adjektiv, bestimmt. Basierend auf der Wortart wird anschließend eine Suffixregel angewendet, die bspw. aus einem Plural einen Singular transformiert.

Für die *Lemmatisation* auf historischen Texten können entgegen der Lemmatisierung von modernen Sprachen beide Schritte nicht als statisch angenommen werden. Einerseits ändern sich Wortartklassen einer Wortform. Andererseits ändern sich auch die Regeln. Bei den zugrunde liegenden Bibelversionen können insbesondere in *KJV* alte Wortformen gefunden werden, die bspw. auf *eth* enden. So entspricht *could* der archaischen Form *couldeth*.

Jede Epoche brachte ihre teilweise sehr eigenen *Lemmatisierungsregeln* mit sich, so dass zwar dadurch Wortformen und damit auch Texte datiert werden können, jedoch auch eine große *Lemmatisation Diversity* generiert wird, die bisher immer nur für bestimmte Epochen als gelöst gilt. Ferner verspricht auch das reine Sammeln solcher Regeln keine allzu großen Erfolge, da durch mehr Lemmatisierungsregeln auch die Chance steigt, eine falsche Regel anzuwenden und somit die *Precision* zu reduzieren. Für vielversprechende Ansätze sei an dieser Stelle auf die Arbeiten von Philip Burns (vgl. [Burns 2012]) für die *Lemmatisation* von historischem Englisch, Brian Jurish (vgl. [Jurish 2012]) für historisches Deutsch sowie von Michael Piotrowski für einen allgemeinen Überblick existierender Techniken (vgl. vor allem Kapitel 6 - *Handling Spelling Variations*, Seite 69 ff. in [Piotrowski 2012]) verwiesen.

Im Sinne einer *Component Evaluation* muss vielmehr die Frage gestellt werden, wie eine *Lemmatisation* für eine *Digital Library* evaluiert werden kann. Aus der Einführung zu diesem Unterabschnitt sollte bereits deutlich geworden sein, dass es keine *Lemmatisation* gibt, die auf jedem Text gleich gut funktioniert. Da es für eine *Text Re-use Analysis* nicht zwangsläufig von Interesse ist, dass eine Wortform auf ihre Grundform reduziert wird, sondern es oftmals bereits ausreicht, wenn alle Wortformen einer Grundform auf die gleiche Form transformiert werden, sei der Aspekt der Qualität der *Lemmatisation* auf das *Lemmatisation Coverage* C^L reduziert. Hierbei ist dementsprechend die Richtigkeit einer *Lemmatisation* eines einzelnen *Tokens* nebensächlich. Deshalb gilt es erst einmal festzustellen, für wie viele der Wortformen (im Text *Tokens*) überhaupt auf eine andere Form transformiert werden kann. Das *Lemmatisation Coverage* C_i^L ist in Formel 5.4 definiert, wobei i dem Rang eines Wortes entspricht. Ein $i = 200$ bedeutet, dass die 200 häufigsten Wörter für die Bestimmung des *Lemmatisation Coverage* nicht berücksichtigt werden. Das ist einfach damit begründet, dass diese wenigen aber hochfrequenten Wörter bereits allein einen *Coverage* von 0.3 bis 0.4 erreichen, so dass die Interpretation des *Lemmatisation Coverage* bei einer perfekten *Lemmatisation* dieser häufigsten Wörter verfälscht sein würde.

$$C_{200}^L = \frac{\sum_{i=201}^{n} t_i^L}{\sum_{i=201}^{n} t_i} \tag{5.4}$$

Während t_i in Formel 5.4 der Wortfrequenz eines *Word Types* entspricht, so ist t_i^L die Häufigkeit eines *Word Types*, für welches auch eine Lemmatisierungsregel vorliegt. Für die sieben Bibelversion können zwischen $n = 7350$ und knapp $= 15000$ unterschiedliche Wortformen festgestellt werden (vgl. Tabelle 5.3 auf Seite 177). Für alle sieben Bibelversionen ergeben sich insgesamt $n = 17417$ verschiedene Wörter.

Für das *Lemmatisation Coverage* kann für alle sieben Bibelversionen mit $n = 17417$ ein Wert von $C^L_{200} = 0.4847$ berechnet werden. Dieser Wert kann als Durchschnitt für alle sieben Bibelversionen verstanden werden. Das *Lemmatisation Coverage* für die einzelnen Bibelversionen selbst streut jedoch deutlich. Während für *BBE* mit einem sehr einfachen und deutlich kürzerem Vokabular (vgl. Tabelle 5.3 auf Seite 177) ein *Lemmatisation Coverage* von $C^L_{200} = 0.9462$ bestimmt werden kann, ist mit $C^L_{200} = 0.2271$ dieser Score für *KJV* mit seinen archaischen Formen deutlich schlechter.

In der Summe zeigt das *Lemmatisation Coverage* sowohl insgesamt als auch speziell für die älteren Versionen der Bibel, dass mit knapp der Hälfte der *Tokens* im Durchschnitt zu wenige Wörter lemmatisiert werden können. Insbesondere für das Erkennen von *Paraphrase* oder *Allusion* ist das jedoch nötig. Insofern bleibt an dieser Stelle nur festzustellen, da nicht Gegenstand dieser Arbeit, dass mit einer Verbesserung in der Lemmatisierung von historischen Varianten noch enorme Potenziale möglich sind, auch wenn die Ergebnisse einer *Text Re-use Analysis* in Abschnitt 5.3 für ein *Word based Featuring* bereits sehr gute Evaluierungswerte liefern.

5.4.2 Qualität im Umgang mit paradigmatischen Relationen

Über die Jahrhunderte haben Objekte unterschiedliche Namen erhalten, so dass die Varianten als synonym angesehen werden können. Insbesondere bei einer *Text Re-use Analysis* auf historischen Daten sind mit zunehmender zeitlicher Distanz zweier Werke Ersetzungen aus dem Original feststellbar (vgl. hierzu auch die *Microview*[14] des *Historical Text Re-use* aus Abb. 1.1 auf Seite 35).

Für die in Abschnitt 5.3 eingesetzten sieben Bibelversionen sind nur drei Beispiele für derartige paradigmatische Relationen die Beziehung zwischen *punishment* und *torment* sowie *not wait* und *not delay* aber auch, wie in Tabelle 5.2 dargestellt, die synonyme Beziehung zwischen *create*, *make* und *prepare*.

Genau wie bei dem *Lemmatisation Coverage* sei auch die Fähigkeit der zugrunde liegenden Daten für eine Behandlung von möglichen Kandidaten einer paradigmatischen Relation durch ein *Coverage* bestimmt. Als Datenbasis für die Synonymbehandlung wird *WordNet* (vgl. [Miller 1995, Fellbaum 1998]) eingesetzt. Für die sieben Bibelversionen kann die Datenmenge in *WordNet* nach der *Lemmatisation* auf 33074 Synonympaare reduziert werden. Die Behandlung von paradigmatischen Relationen kann im Sinne des *Information Retrieval* als *Query Expansion* aufgefasst werden. Das *Synonym Coverage* C^S_{200} ist in Formel 5.5 definiert. Hierbei wird davon ausgegangen, dass für jedes Wort auch mindestens ein Synonym existiert, wie es im Sprachgebrauch üblich ist, so dass zumindest theoretisch ein Wert von $C^S_{200} = 1.0$ möglich wäre.

$$C^S_{200} = \frac{\sum_{i=201}^{n} t_i^S}{\sum_{i=201}^{n} t_i} \tag{5.5}$$

Für das *Synonym Coverage* kann über allen sieben Bibelversionen mit $n = 17417$ ein Wert von $C^S_{200} = 0.2955$ berechnet werden. Dieser Wert kann als Durchschnitt für alle sieben Bibelversionen verstanden werden. Da ein *Query Expansion* durch Synonyme sehr stark von der *Lemmatisation* abhängig ist, zeichnet sich für das *Synonym Coverage* C^S_{200} ein ähnliches Bild wie für das Lemmatisation Coverage C^L_{200} ab. Während sich C^S_{200} für *BBE* deutlich vom Durchschnitt abhebt, bleibt insbesondere *KJV* weit hinter dem Durchschnitt aller sieben Bibelversionen von $C^S_{200} = 0.2955$ mit nur $C^S_{200} = 0.0734$ zurück.

[14] Die *Microview* eignet sich ideal dafür, Varianten eines Textes bzw. eines *Chunks* auf Ersetzungen zu visualisieren. Hierbei wird eine Ersetzung als ein eigener Zweig dargestellt, so dass entsprechende Veränderungen für den Fachwissenschaftler leichter zu explorieren sind.

Auch wenn *WordNet* (vgl. [Miller 1995, Fellbaum 1998]) für Englisch als die größte Sammlung von *Synsets* angesehen werden kann, so gibt es, genau wie in den Experimenten aus Abschnitt 5.3, eine natürliche Selektion, wodurch keine *Digital Library* perfekt auf die Daten in *WordNet* passt, weshalb der quantitative und semiautomatische Aufbau eines domänenspezifischen *WordNet* (vgl. [Piasecki 2009]) zukünftig nahe liegt, um besser mit der Varianz in der Benutzung von semantischen Relationen umgehen zu können. Weiterhin können Wörterbücher eingesetzt werden, so dass durch ein *Word Sense Alignment* (vgl. u. a. [Ackermann 2013])[15] ähnliche semantische Wörter auf Basis deren Wörterbuchbeschreibungen bestimmt werden können.

5.4.3 Qualität im Umgang mit historischen Varianten

Genau wie unterschiedliche Lemmatisierungsregeln haben Konzepte über die Jahrhunderte verschiedene Schreibweisen erfahren. Im Lateinischen gibt es für *Amsterdam* Schreibweisen, wie *Amsteldam*, *Amsterodamum* oder *Amstelodamum*[16]. Die Gründe für unterschiedliche Varianten sind vielschichtig und reichen vor allem in der vorchristlichen Zeit von der fehlenden Existenz einer Rechtschreibung im heutigen Sinne bis hin zu regionalen Varianten. Historische Varianten haben gemeinsam, dass sie sich durch eine ähnliche Schreibweise auszeichnen. Daher können einfache Stringähnlichkeitsalgorithmen, wie die *Levenshtein Distance* (vgl. [Levenshtein 1966]), eingesetzt werden, um die historischen Schreibweisen eines Wortes zu erkennen (vgl. für die Anwendung in den *eHumanities* auch [Piotrowski 2012] bzw. für die Informatik [Crochemore 2003]). Weiterhin sind in den Geisteswissenschaften über die letzten Jahrzehnte Wörterbücher mit Varianten entstanden, welche perspektivisch, sofern sie frei verfügbar sind, ebenfalls für diesen Zweck eingesetzt werden können[17].

Der *StringSim*-Ansatz (vgl. Abschnitt 5.3) basiert auf Buchstabenbigrammen, durch welche mittels des *Dice Coefficient* die Ähnlichkeit der Überlappung zweier Mengen von *Buchstabenbigrammen*[18] bestimmt wird. Auf diese Weise können für alle sieben Bibelversionen mit in der Summe 17417 Wörtern 83866 ungerichtete Kanten zwischen Wörtern bestimmt werden. Der Mindestähnlichkeitsschwellwert wurde mit 0.7 gewählt, so dass die berechnete Menge weitestgehend aus historischen Schreibweisen sowie morphologischen Varianten besteht.

In Anlehnung an das *Lemmatisation* und *Synonym Coverage* wird in Formel 5.6 das *Variant Coverage* C^V_{200} definiert. Theoretisch ist es zwar möglich, dass $C^V_{200} = 1$ werden kann, realistisch sind jedoch kleinere obere Grenzen in Abhängigkeit von der *Digital Library* von $0.6 \leq C^V_{200} \leq 0.8$ [19].

[15] Die Bachelorarbeit von Markus Ackermann wird im Rahmen vom *eTRACES*-Projekt betreut. Die grundlegende Idee hierzu basiert auf einem persönlichen Gespräch mit Thomas Meyer aus Darmstadt vom *Ubiquitous Knowledge Processing Lab* unter Leitung von Prof. Dr. Iryna Gurevych, der auf diesem Gebiet bereits seit geraumer Zeit arbeitet.

[16] Dieses Beispiel wurde von Dr. Reinhard Gruhl, CAMENA-TERMINI Heidelberg, Deutschland, im Rahmen seines Vortrages *Das Wissensnetz der Frühen Neuzeit - Von der virtuellen Bibliothek zur virtuellen Enzyklopädie* während der Eröffnungsveranstaltung zum *eAQUA*-Projekt am 15. April 2008 vorgestellt.

[17] Einen guten Überblick hierzu gab der von Dr. Ingelore Hafemann am 12./13. Dezember 2011 organisierte Workshop *Perspektiven einer corpusbasierten Linguistik und Philologie* an der Berlin-Brandenburgischen Akademie der Wissenschaften, Berlin, Deutschland. Im Rahmen dieses Workshops wurden u. a. solche Aspekte diskutiert. Insbesondere wurde auch deutlich, dass es für unterschiedliche Epochen auch unterschiedliche Problemstellungen, wie die Sicherheit der Bedeutung während der Überlieferung über die Jahrhunderte, mit solchen Daten gibt.

[18] Eine Menge von Buchstabenbigrammen entspricht einem Wort.

[19] Dieses Intervall basiert auf Erfahrungen sowohl auf den englischsprachigen Bibeln sowie Analysen auf der *Perseus Digital Library* als auch deutschsprachigen Texten.

$$C_{200}^{V} = \frac{\sum_{i=201}^{n} t_i^V}{\sum_{i=201}^{n} t_i} \tag{5.6}$$

Das *Variant Coverage* C^V erreicht einen Wert von $C^V = 0.4333$ über alle sieben Bibelversionen. Entgegen den beiden anderen *Coverage*-Maßen ist das *Variant Coverage* mit $C^V = 0.5283$ für *KJV* am größten. Das kann neben verschiedenen archaischen Formen auf unterschiedliche historische Varianten zurückgeführt werden. *KJV* ist letztlich mindestens 300 Jahre älter als die anderen in Abschnitt 5.3 eingesetzten Bibelversionen.

Da die Massendigitalisierung durch *OCR* stetig steigt, kann das *Variant Coverage* C^V zukünftig auch eingesetzt werden, um die Qualität des Säuberns von *OCR-Fehlern* während einer *Text Re-use Analysis* zu messen. Letztlich ist der Unterschied zwischen einer historischen Variante und einem *OCR*-Fehler nur durch die menschliche Kognition gegeben. Für den Computer hingegen sind Varianten eines Wortes oftmals nur selten auftretende Formen mit einer unwahrscheinlichen Buchstaben- bzw. Zeichenfolge.

5.4.4 Qualität der *Digital Signature*

Die *Digital Signature* wurde in Abschnitt 3.5 als diejenige reduzierte *Feature*-Menge einer *Re-use Unit* definiert, welche für das *Linking* mit anderen *Re-use Units* als Grundlage dient. Neben den zuvor genannten Aspekten des *Preprocessing*, wie der Qualität der *Lemmatisation*, ist die *Digital Signature* im Wesentlichen vom *Featuring* sowie der *Selection* abhängig. Während der *System Evaluation* ist das *Selection* mit dem *max pruning* als konstant angenommen gewesen. Für das *Featuring* wurde in verschiedenen Analysen jeweils *Trigram Shingling*, *Bigram Shingling* sowie das *Word based Featuring* ausgewählt.

Die grundlegende Frage ist schließlich, wie gut die auf diese Weise generierte *Digital Signature* ist. Was ist ein gutes *Featuring* bzw. was ist eine gute *Selection*? Letztlich hat das Zusammenspiel dieser beiden Größen einen maßgeblichen Einfluss auf das Qualitätskriterium *Performance* (vgl. Abschnitt 2.4 ab Seite 62 ff.). Sowohl die damit implizierte Qualität als auch die Geschwindigkeit hängen stark davon ab. Vielmehr treten sie gegensätzlich auf. So sichert eine hohe *Feature Density* (vgl. Abschnitt 3.5 ab Seite 104 ff.), wenn auch nicht immer eine hohe *Precision*, zumindest einen hohen *Recall*. Auf der anderen Seite bedeutet eine hohe *Feature Density*, dass die Geschwindigkeit aufgrund des quadratischen Verhaltens einer *Text Re-use Analysis* progressiv steigt.

Um die Qualität einer *Digital Signature* einschätzen zu können, wurde in Abschnitt 3.5 der *Re-use Nucleus* definiert, welcher der theoretisch nötigen Information einer *Re-use Unit* entspricht, um diese zu repräsentieren. Der *Re-use Nucleus* ist insofern ein theoretisches Konstrukt, da es sehr unwahrscheinlich ist, dass aus einer Menge von 10 Personen eine exakte Übereinstimmung bei der Deutung bzw. Ermittlung des *Re-use Nucleus* vorliegt. Vielmehr ist er ein theoretisches Ideal, dem sich nur quantitativ genähert werden kann.

Um die in Abschnitt 5.3 zugrunde liegende *Digital Signature* zu evaluieren, sei zunächst von den Bibeltexten zurückgetreten und der Fokus auf allgemeinsprachliche Redewendungen gelegt. Dies ist einfach damit begründet, dass der *Re-use Nucleus* als theoretisches Ideal unabhängig vom Text sein soll. Deshalb ist eine erste Analyse auf allgemeinsprachlichen *Idioms* insofern hilfreich, als dass sie weit verbreitet in Texten beobachtet werden können, während die Bibelverse zu spezifisch in religiösen Kontexten gebraucht werden. Die Untersuchungsbasis wird auf zwei Bücher eingeschränkt (vgl. [Wagner 2011b] und [Wagner 2011a]). Während Ersteres mit dem Titel *Wer's glaubt wird selig!* den Fokus auf *Idiome* mit biblischem Bezug legt, sind in letzterem Werk mit dem Titel *Das geht auf keine Kuhhaut Idiome* des Mittelalters gesammelt. Beide Werk enthalten 200 sowie 202 gesammelte Rede-

wendungen, welche für diese Zwecke abgeschrieben worden sind[20]. Anschließend wurden die Texte noch mit dem *Perseus Tag System*[21] manuell getaggt[22], welches aus vierzehn *Part of Speech*-Tags besteht (vgl. Tabelle 5.17). Die *PoS*-Tags wurden für die Probanden entfernt und erst nachträglich wieder den ausgewählten Wörtern zugeordnet.

Part of Speech-Tag	Wortartklasse
n	noun
v	verb
t	participle
a	adjective
d	adverb
l	article
g	particle
c	conjunction
r	preposition
p	pronoun
m	numeral
i	interjection
e	exclamation
u	punctuation

Tabelle 5.17: *Perseus Tag System*. Die Tabelle repräsentiert die 14 verschiedenen *Part of Speech*-Tags des *Perseus Tag System*.

Die insgesamt 402 Redewendungen wurden nach Bibel- und Mittelalterbezug separat 24 Testpersonen[23] vorgelegt, welche die Redewendungen um diejenigen Wörter verkürzen sollten, die für das Erkennen nötig sind. Da die Prämissen der Probanden in Alter, Geschlecht und Bildungsgrad unterschiedlich waren, wurde als initiale Hilfestellung *Google* gewählt. Die Testpersonen sollten die Redewendungen verkürzen, so dass sich trotz Reduzierung der *Proband* eine hohe Trefferwahrscheinlichkeit verspricht. Aufgrund des Informationszeitalters schien diese Hilfestellung der passendste gemeinsame Nenner zu sein. Vielmehr wurden die Probanden angehalten, *Google* nicht zu benutzen, sondern nur ihre Anfrage an *Google* zu formulieren. Alle Testpersonen sind Muttersprachler für Deutsch, so dass mangelndes Verständnis einer Redewendung ausgeschlossen werden kann. Die durchschnittliche Länge eines *Idioms* mit biblischen bzw. einem Mittelalterbezug beträgt 3.99 und 3.77 Wörter.

[20] Die Redewendungen wurden dankenswerterweise von Petra Gamrath im Rahmen von Arbeiten des *eTRACES*-Projektes abgeschrieben und stehen somit digital, wenn auch nicht mit *Open Access*, zur Verfügung.

[21] vgl. `http://nlp.perseus.tufts.edu/syntax/treebank/agdt/1.7/docs/README.txt`

[22] Das manuelle Taggen der 402 Redewendungen wurde dankenswerterweise von Markus Ackermann übernommen, der zum Zeitpunkt dieser Tätigkeit als studentische Hilfskraft im *eTRACES*-Projekt eingestellt ist.

[23] Die 24 Testpersonen sind namentlich (in alphabetischer Reihenfolge): Markus Ackermann (*eTRACES*, Leipzig), Frederik Baumgardt (*eTRACES*, Halle/S.), Volker Boehlke (*CLARIN*, Leipzig), Anett Büchler (Leipzig) Gabriele Büchler (Delitzsch), Thomas Eckart (*CLARIN*, Leipzig), Thomas Efer (*eXCHANGE*, Leipzig), Lars Gadegast (Leipzig), Annette Geßner (*eTRACES*, Leipzig), Katarina Jacob (Rackwitz), Ronny Jacob (Rackwitz), Stefan Jänicke (*eTRACES*, Leipzig), Christian Kötteritzsch (*eTRACES*, Leipzig), Matthias Leopold (Deutsche Zentralbibliothek für Blinde, Leipzig), Sebastian Lissmann (*eAQUA*, Leipzig), Maria Moritz (*eTRACES*, Leipzig), Clemens Neudecker (Niederländische Nationalbibliothek, Den Haag, Niederlande), Ute Pietruschka (*CASG*, Halle/S.), Elke Rosenkranz (Markkleeberg), Martin Schierle (Ulm), Thomas Stäcker (Herzog-August-Bibliothek, Wolfenbüttel), Lydia Steiner (Institut für Medizinische Informatik, Statistik und Epidemiologie, Leipzig), Sabine Thänert (Deutsches Archäologisches Institut, Berlin) sowie Diana Winger (Leipzig).

In einem ersten Test ist es das Ziel, herauszufinden, welche Wörter besonders häufig entfernt worden sind. Hierzu können durch eine Differenzanalyse aus der originalen mit der reduzierten Redewendung, diejenigen Wörter bestimmt werden, die gelöscht wurden. Für die Redewendungen mit biblischem Kontext (Bibel) sowie für die Redewendungen, die einem Ursprung im Mittelalter zugewiesen sind (Mittelalter), können so die folgenden Listen generiert werden, wobei die Zahl in Klammer der Häufigkeit entspricht, mit welcher das entsprechende *Token* entfernt worden ist:

- **Bibel**: *ein* (563), *die* (276), *das* (193), *sein* (176), *den* (170), *der* (169), *wie* (131), *und* (127), *im* (107), *ist* (105), *etwas* (94), *einen* (93), *in* (92), *eine* (88), *auf* (78), *sich* (76), *sein* (73), *jemanden* (71), *haben* (58), *!* (55), *,* (50), *von* (46), *vom* (43), *jemandem* (42), *gehen* (41), *das* (38), *machen* (38), *werden* (38), *dem* (37), *mit* (37)
- **Mittelalter**: *ein* (563), *die* (276), *das* (193), *sein* (186), *einen* (172), *ein* (140), *und* (117), *sich* (111), *haben* (107), *auf* (98), *dem* (93), *!* (85), *der* (77), *,* (75), *eine* (64), *mit* (64), *jemandem* (59), *jemanden* (46), *in* (40), *ins* (40), *am* (38), *kommen* (37), *einer* (35), *machen* (35), *wie* (34), *aus* (33), *es* (31), *das* (30), *legen* (29)

Im Wesentlichen bestehen diese Liste übereinstimmend aus *Stoppwörter*, *Satzzeichen*, die explizit als eigenständige *Tokens* behandelt werden sollten, sowie einigen Hilfsverben, wie *sein* und *haben*. Da genau diese *Tokens* insbesondere durch das eingesetzte *max pruning* ebenfalls aus der *Digital Signature* entfernt werden, kann die bei der *System Evaluation* auf den sieben Bibelversionen gewählte *Selection*-Strategie nachträglich hinreichend motiviert werden.

In einem zweiten Test stellt sich die Frage, wie gut die ausgewählte *Feature Density* von $\mathcal{F} = 0.8$ (vgl. *System Evaluation* in Abschnitt 5.3) im Vergleich zu einer manuellen Selektion ist. Für die Redewendungen mit biblischem Kontext kann eine *Feature Density* von $\mathcal{F}^{\mathcal{B}} = 0.7585$ bestimmt werden. Die *Feature Density* für die mittelalterlichen Redewendungen liegt bei $\mathcal{F}^{\mathcal{M}} = 0.7699$. Der Bereich der manuellen *Feature Selection* liegt zwischen $0.5310 \leq \mathcal{F}^{\mathcal{B}} \leq 1.0$ sowie zwischen $0.5099 \leq \mathcal{F}^{\mathcal{M}} \leq 1.0$. Die jeweiligen Standardabweichungen von $\sigma_B = 0.1367$ und $\sigma_M = 0.1435$ zeigen auf, dass die *Feature Density* der Probanden vergleichsweise stark streut. Dies wird insbesondere im Vergleich zum Wertebereich von jeweils knapp 0.5 deutlich.

In einem dritten Test kann untersucht werden, welche *PoS*-Tags besonders häufig selektiert werden. Hierzu wird die Wahrscheinlichkeit eines *PoS*-Tags aus Tabelle 5.17 sowohl für die vollständigen $P_v(tag)$ als auch auf den reduzierten Redewendungen $P_r(tag)$ berechnet. Anschließend wird für jedes *PoS*-Tag das Verhältnis $P_r(tag)/P_v(tag)$ bestimmt, welches als die *Feature Density* jedes *PoS*-Tags verstanden werden kann und zu welchem bspw. Substantive (n) von den 24 Probanden im Durchschnitt ausgewählt worden sind.

	n	v	t	a	d	l	g	c	r	p	m	u
Bibel	0.98	0.86	0.81	0.95	0.69	0.39	0.71	0.70	0.72	0.56	0.80	0.58
Mittelalter	0.98	0.88	0.93	0.95	0.79	0.42	0.81	0.71	0.79	0.49	0.84	0.52

Tabelle 5.18: *Feature Density* pro *PoS*-Tag aus Tabelle 5.17.

Hierzu stellen $\mathcal{F}^{\mathcal{B}} = 0.7585$ und $\mathcal{F}^{\mathcal{M}} = 0.7699$ aus dem vorigen Test eine gute *Baseline* dar, um diejenigen *PoS*-Tags zu identifizieren, welche auffällig häufiger in die reduzierten Redewendungen übernommen und welche besonders häufig entfernt worden sind. Tabelle 5.18 beinhaltet die Ergebnisse dieses Tests, wobei Ergebnisse mit einem schwarzen Hintergrund für eine starke Übernahme eines Wortes des entsprechenden *PoS*-Tags in die

reduzierte Version einer Redewendung stehen. Ergebnisse mit einem weißen Hintergrund hingegen können als stark selektiert angesehen werden. Ein grauer Hintergrund repräsentiert eine gegenüber dem jeweiligen Durchschnitt von $\mathcal{F}^{\mathcal{B}} = 0.7585$ und $\mathcal{F}^{\mathcal{M}} = 0.7699$ nur schwach signifikante Übernahme in die reduzierte Version einer Redewendung.

Sowohl für die biblischen als auch den Redewendungen des Mittelalters können Substantive (n), Adjektive (a) sowie Verben (v) übereinstimmend als überdurchschnittlich relevant für die 24 Probanden angesehen werden. Speziell für Verben (v) lässt sich $\mathcal{F}^{\mathcal{B}} = 0.81$ bzw. $\mathcal{F}^{\mathcal{B}} = 0.93$ (vgl. Tabelle 5.18) insofern verbessern, wenn Hilfsverben, wie *haben* oder *sein*, separat betrachtet werden. Weiterhin können auch die Klassen der Partizipien (t) sowie Numerale (m) als relevant angesehen werden, wobei darauf hingewiesen sein muss, dass für die Analysen auf der Bibel beide Wortartklassen lediglich schwach relevant sind.

Bei einer Adaption dieser Ergebnisse auf ein wortartspezifisches *Selection* für die sieben Bibelversionen aus Abschnitt 5.3 wird eine *Feature Density* von $\mathcal{F} = 0.354$ bestimmt, wenn lediglich Substantive, Adjektive sowie Verben selektiert und alle anderen Wortarten gefiltert werden. Bleiben zusätzlich noch Partizipien und Numerale während des *Selection* berücksichtigt, so steigt die *Feature Density* auf $\mathcal{F} = 0.438$. Während sich der *Recall* nicht entscheidend verändert, steigt die *Precision* bei einer Wiederholung der Experimente zur *System Evaluation* aus Tabelle 5.5 (vgl. Seite 180) um durchschnittlich etwa 4.5%, wobei die *Feature Density* mit $\mathcal{F} = 0.354$ bzw. $\mathcal{F} = 0.438$ in etwa um die Hälfte kleiner ist als in den initialen Experimenten. Insgesamt kann festgestellt werden, dass ein *max pruning* als *Feature Selection* zwar einfach und effizient ist, jedoch durch eine deutlich geringere *Feature Density* eine wortartbasierte *Selection* die *Performance* (vgl. Abschnitt 2.4 auf Seite 62) deutlich verbessert. Das umfasst die Qualität in Form der *Precision*, die durchschnittlich etwa 4.5% verbessert wird sowie insbesondere auch der Geschwindigkeit (vgl. auch Abschnitt 5.4.5).

In einem letzten Test können nicht nur die einzelnen Wortartklassen betrachtet werden, sondern auch signifikante Verbindungen zwischen verschiedenen Wortarten. Hierzu werden Beziehungen zwischen zwei, drei und vier Wortartklassen untersucht, die nicht notwendigerweise *NGrams* sein müssen. Im Ergebnis kann festgehalten werden, dass es keine signifikante Beziehung zwischen vier Wortartklassen gibt, was wiederum auch auf die geringe Menge an Daten zurückgeführt werden kann. Dennoch konnte zwischen zwei Wortartklassen die fünf **n v** sowie **[a|l|r|m] n** Patterns als auffällig relevant bestimmt werden. Die vier Muster **[l|r] n v** und **n [r|c] n** sind hinreichend auffällig für Beziehungen zwischen drei Wortartklassen. Auch wenn insgesamt neun Patterns zwischen verschiedenen Wortarten durch die 24 Probanden als relevant gemessen werden konnten, so muss auf die Dominanz von **n v** hingewiesen werden. Derartige Substantiv-Verb-Verbindungen entsprechen knapp 40% aller signifikanten Verbindungen zwischen Wortartklassen. Interessanterweise, wenn auch an dieser Stelle nicht weiter verfolgt, ist das Muster **n v** eines der ersten Muster, die Kinder beim Spracherwerb erlernen. So können Instanzen dieses Musters, wie *Oma fahren*, als eine der ersten sprachlichen Kommunikationen zwischen Eltern und Kind beobachtet werden[24]. Auch hier werden zuerst keine vollständigen Sätze formuliert, sondern reduzierte Sätze, die bis auf Patterns, wie **n v**, verkürzt werden.

Zusammenfassend kann für die Evaluierung der *Digital Signature* festgehalten werden, dass mit einfachen und intuitiven *Selection*-Strategien, wie dem *max pruning*, bereits hinreichend gute Ergebnisse erzielt werden können. Unter dem Aspekt der *Performance*, insbesondere von *Big Scale Data*, zeigt sich jedoch auch, dass zukünftig weitere Forschungen im Bereich der *Minutiae* und dem besseren Verständnis des *Re-use Nucleus* nötig sind. Das wird bspw. auch beim *Featuring* deutlich. Während im aktuellen Forschungsstand immer

[24] Der Autor konnte das explizit parallel zu dieser Arbeit an seinem zum Zeitpunkt dieser Arbeit vierjährigen Neffen Marlon Jacob beobachten.

ein Atom fester Größe gewählt wird (vgl. auch *Trigram Shingling*, *Bigram Shingling* sowie *Word based Featuring* in Abschnitt 5.3 auf Seite 176 ff.), zeigen die eher einfachen Analysen dieses Abschnittes, dass bestimmte *Features*, wie Substantive und Verben, im Verbund als Einheit bzw. als ein zusammengesetztes *Feature* zu behandeln sind. Somit wäre eine Tendenz gegeben, zu welcher es zukünftig keine konstante *Atom*-Länge eines *Minutiae* in der *Digital Signature* geben wird. Vielmehr kann dies als der Beginn der *Minutiae*-Forschung für den *Historical Text Re-use* verstanden werden.

5.4.5 Qualität des *Linking*

Das *Linking* ist, wie bereits mehrfach erwähnt, von quadratischer Komplexität $O(n^2)$. Die dafür aufgebrachte Zeit ist hauptsächlich von drei Parametern abhängig. Erstens, die Anzahl der *Re-use Units* hat einen direkten Einfluss auf das Laufzeitverhalten. Zweitens, da nicht einfach jede *Re-use Unit* mit jeder anderen verglichen wird, so wie es in einem *Brute Force*-Ansatz nötig wäre, sondern nur diejenigen *Re-use Units* miteinander verglichen werden, die auch mindestens ein *Feature* in der *Digital Signature* gemeinsam haben, ist die Wahl einer geeigneten *Feature Density* $\mathcal{F}$ von besonderer Bedeutung. Drittens, die Frequenz f eines *Features* in der *Digital Signature* hat einen direkten Einfluss auf die *Linking Complexity*, da jedes *Token* eines *Features* mit jedem anderen verlinkt wird, so dass die *Linking*-Kosten für ein *Feature* $f \cdot (f-1)$ betragen. Wird nun ein *Feature* einem anderen bevorzugt, welches doppelt so häufig in einer *Digital Library* vorkommt, dann erhöht das die *Linking*-Kosten für dieses *Feature* um etwa den Faktor 4.

Die zuvor genannten Aspekte haben alle einen direkten Einfluss auf die *Performance*. Im folgenden Abschnitt liegt der Fokus nicht auf der Genauigkeit bzw. der Qualität, sondern der Geschwindigkeit. Hierzu werden insgesamt drei Analysen vorgestellt, die das Laufzeitverhalten einer *Text Re-use Analysis* aus Abschnitt 5.3 beschreiben.

In einem ersten *Linking*-Test werden das *Trigram Shingling*, *Bigram Shingling* sowie das *Word based Featuring* verglichen (vgl. Tabelle 5.19). Allen drei *Featuring*-Methoden liegt das *Preprocessing* ξ_4 zugrunde (vgl. Seite 177). Die Wahl dieses *Preprocessing* ist für die *Linking*-Analyse damit begründet, dass es von den auf Seite 177 vorgestellten *Preprocessing*-Techniken die aktivste Form ist. Im Detail bedeutet dies, dass durch die *Lemmatisation* sowie durch die Behandlung von Synonymen die Wortfrequenzen im Vergleich zu den drei anderen Techniken im Durchschnitt am höchsten sind. Somit kann die nachfolgende *Linking*-Analyse für alle *Preprocessing*-Techniken aus Abschnitt 5.4.5 als obere Grenze angesehen werden. Weiterhin ist die *Feature Density* $\mathcal{F} = 0.8$. Vielmehr entspricht der erste Test der *Text Re-use Analysis* aus Abschnitt 5.3, mit der Einschränkung, dass nur ein *Preprocessing* als obere Grenze betrachtet wird und die sieben Bibelversionen nicht separat betrachtet werden.

Tabelle 5.19 reflektiert die Ergebnisse der *Linking*-Analyse. Entgegen der Analyse in der *System Evaluation*, wo alle 42 paarweisen Bibelvergleiche separat behandelt worden sind, werden die sieben Bibelversionen in Tabelle 5.19 als eine *Digital Library* betrachtet. Somit ergeben sich insgesamt 200424 Verse[25]. Die Spalte *Unique Links* repräsentiert die Anzahl der Kanten im *Text Re-use Graph*. Es sei an dieser Stelle darauf hingewiesen, dass für die *Linking*-Analyse kein *Scoring*-Schwellwert t eingesetzt wird, so dass von einem *Long Tail* von vielen *Links* mit nur einem oder zwei *Features* im *Re-use Overlap* ausgegangen werden kann. Die Spalte *Linked Links* entspricht der Anzahl an *Links*, die während des *Linking* insgesamt aufgebaut werden. Haben zwei *Re-use Units* bspw. vier *Features* im *Re-use Overlap* gemeinsam, dann werden vier *Linked Links* generiert, während nur ein *Unique Link* gezählt wird. Die Spalte *Average Links* entspricht dem Verhältnis von *Linked Links*

[25] Jede reduzierte Bibel besitzt 28632 Verse, so dass $28632 \cdot 7 = 200424$ gilt.

zu *Unique Links*. Der *Brute Force Score* stellt eine *Linking*-Analyse in Relation zum *Brute Force Linking* (BFS), welches jede *Re-use Unit* s_i mit jeder anderen s_j vergleicht, wodurch sich, wie in Formel 5.7 dargestellt, ein quadratischer *Linking*-Aufwand ergibt.

$$BFS = |\bigcup s_i| \cdot (|\bigcup s_i| - 1) \tag{5.7}$$

Für die insgesamt 200424 zugrunde liegenden Verse, den Re-use Units s_i, ergeben sich somit 40,169 Milliarden paarweise Vergleiche aller Verse. Ein *Brute Force Score* von $BFS = 1$ bedeutet demnach, dass exakt genauso viele *Linked Links* generiert werden, wie durch ein *Brute Force Linking* nötig wären.

Tabelle 5.19 zeigt auf, dass sich die Länge eines Atoms umgekehrt logarithmisch proportional zu den *Unique Links* verhält. Vielmehr kann durch das Verkürzen eines *Atoms* um ein Wort, die Anzahl der berechneten *Links* verzehnfacht werden. In Anlehnung an den *Recall* aus Tabelle 5.6 (vgl. Seite 181) sind die *Average Links* für das *Trigram Shingling* und *Bigram Shingling* mit Werten von 1.303 und 1.197 vergleichsweise niedrig, während sich für das *Word based Featuring* der *Score* fast verdreifacht zumindest aber mehr als verdoppelt.

	Unique Links	Linked Links	Average Links	Brute Force Score
Trigram	123 M	160 M	1.303	0.00399
Bigram	2531 M	3030 M	1.197	0.07543
Word	35294 M	106211 M	3.009	2.64408

Tabelle 5.19: *Linking*-Analyse mit unterschiedlichen *Featuring*-Techniken. Die Tabelle reflektiert die *Linking*-Analyse für das *Trigram Shingling*, *Bigram Shingling* sowie das *Word based Featuring*. Die Ergebnisse basieren auf dem *Preprocessing* ξ_4 der *System Evaluation* von Seite 176 ff. Sowohl *Unique Links* als auch *Linked Links* sind in Millionen (M) angegeben.

Tabelle 5.19 zeigt ebenfalls auf, dass der *Brute Force Score* (vgl. Formel 5.7) für das *Trigram Shingling* und *Bigram Shingling* relativ klein ist, so dass das Laufzeitverhalten gegenüber dem *Brute Force Linking* signifikant verbessert wurde. Auf der anderen Seite muss auch konstatiert werden, dass für das *Word based Featuring* ein *Brute Force Score* von $BFS = 2.64408$ berechnet wird, was zumindest für das *Linking* die 2.6-fache Laufzeit gegenüber der *Brute Force*-Methode bedeutet.

Aus diesem Grund soll in einem zweiten Test das initiale Experiment der *System Evaluation* aus Abschnitt 5.3 zu unterschiedlichen *Feature Density* wiederholt werden. Tabelle 5.20 reflektiert diese Ergebnisse. Der Tabelle kann entnommen werden, dass für das zugrunde liegende *Selection* des *max pruning* ab einer *Feature Density* von $\mathcal{F} \in [0.6, 0.7]$ ein *Brute Force Score* von etwa $BFS = 1.0$ erreicht wird. Weiterhin ist ersichtlich, dass bereits bei einer *Feature Density* von $\mathcal{F} = 0.5$ der *BFS* nur noch 0.53669 beträgt und mit weiterem Herabsenken von $\mathcal{F}$ deutlich fällt.

Abschließend zu Tabelle 5.20 sei noch auf die Konvergenz der *Average Links* und des *Brute Force Score* mit zunehmender *Feature Density* $\mathcal{F}$ hingewiesen. Beide *Scores* haben im Zähler die *Linked Links* gemeinsam. Während der Nenner der *Average Links* der Anzahl der *Unique Links* entspricht, wird der *Brute Force Score* durch die Anzahl jedes paarweisen Vergleiches einer *Re-use Unit* mit jeder anderen normalisiert. Da mit zunehmender *Feature Density* die Anzahl der *Unique Links* gegen die Anzahl jedes paarweisen Vergleiches konvergiert, laufen auch der *Score* der *Average Links* sowie der *Brute Force Score* mit steigender *Feature Density* zusammen. Dieses Verhalten kann dazu eingesetzt werden, die *Average*

Links als obere Schranke für den *Brute Force Score* anzusehen, um bereits während einer *Text Re-use Analysis* eine Laufzeitabschätzung des *Linking*-Schrittes abzugeben.

	Unique Links	Linked Links	Average Links	Brute Force Score
$\mathcal{F} = 0.1$	161 M	165 M	1.02350	0.00411
$\mathcal{F} = 0.2$	1099 M	1152 M	1.04786	0.02868
$\mathcal{F} = 0.3$	3778 M	4198 M	1.11109	0.10452
$\mathcal{F} = 0.4$	8347 M	10246 M	1.22754	0.25508
$\mathcal{F} = 0.5$	15130 M	21558 M	1.42489	0.53669
$\mathcal{F} = 0.6$	21687 M	37003 M	1.70621	0.92117
$\mathcal{F} = 0.7$	29224 M	64521 M	2.20779	1.60623
$\mathcal{F} = 0.8$	35294 M	106211 M	3.0093	2.64408
$\mathcal{F} = 0.9$	38685 M	157160 M	4.06248	3.91241
$\mathcal{F} = 1.0$	39914 M	204354 M	5.11974	5.08729

Tabelle 5.20: *Linking*-Analyse mit unterschiedlichen *Feature Density* $\mathcal{F}$. Die Tabelle reflektiert die *Linking*-Analyse für unterschiedliche *Feature Density* $\mathcal{F}$. Die Ergebnisse basieren auf dem *Preprocessing* ξ_4 der *System Evaluation* von Seite 176 ff.

In Abschnitt 5.4.4 zur *Qualität der Digital Signature* einer *Re-use Unit* konnte analysiert werden, dass bestimmte Wortartklassen (*Part of Speech*) überdurchschnittlich von insgesamt 24 Probanden ausgewählt worden sind. Hierzu wurden zwei Mengen von *PoS*-Tags definiert. Einerseits konnte in zwei Experimenten gezeigt werden, dass Substantive (n), Adjektive (a) sowie Verben (v) bevorzugt selektiert werden. Weiterhin hat sich gezeigt, dass in einer zweiten Menge zusätzlich auch die Partizipien (t) sowie Numerale (m) berücksichtigt werden können. Für diese beiden Mengen einer wortartspezifischen *Selection* konnte in Abschnitt 5.4.4 eine *Feature Density* von $\mathcal{F} = 0.354$ bzw. $\mathcal{F} = 0.438$ berechnet werden.

Tabelle 5.21 spiegelt das Ergebnis einer *Linking*-Analyse mit einer wortartspezifischen *Selection* wieder. Hierbei bleiben die bisherigen Parameter der *System Evaluation* (vgl. Abschnitt 5.3) bis auf die veränderte *Selection*-Strategie unverändert.

	Unique Links	Linked Links	Average Links	Brute Force Score
n, a, v	26732 M	45260 M	1.69305	1.12673
n, a, v, t, m	27606 M	51091 M	1.85072	1.27189

Tabelle 5.21: *Linking*-Analyse mit einer wortartbasierten *Selection*. Die Wahl der zwei Mengen von berücksichtigten *PoS*-Tags (vgl. Tabelle 5.17 auf Seite 206) basiert auf Ergebnissen einer manuellen *Selection* von insgesamt 24 Probanden (vgl. Abschnitt 5.4.4).

Es können zwei Dinge festgestellt werden. Erstens, in Abschnitt 5.4.4 konnte aufgezeigt werden, dass der *Recall* zwischen einem *Selection* durch *max pruning* bei einer *Feature Density* von $\mathcal{F} = 0.8$ und einer wortartbasierten *Selection* mit genannten *PoS*-Tags bei einer *Feature Density* von $\mathcal{F} = 0.354$ bzw. $\mathcal{F} = 0.438$ nahezu unverändert bleibt. Die *Precision* konnte in beiden Fällen durch eine wortartbasierte *Selection* um etwa 4.5% verbessert werden. Im Kontext dieser Verbesserung kann weiterhin anhand des *Brute Force Score* aus Tabelle 5.21 konstatiert werden, dass dieser sich in beiden Fällen gegenüber dem initial eingesetzten *max pruning* bei einer *Feature Density* von $\mathcal{F} = 0.8$ (vgl. Tabelle 5.20) in etwa halbiert und somit die *Performance* sowohl in der Qualität als auch in der Geschwindigkeit verbessert werden konnte.

Zweitens, es muss dennoch festgestellt werden, dass trotz verbesserter Ergebnisse in der *Performance* ein *Brute Force Score* von deutlich über 1.0 berechnet werden muss. Somit verbessert die wortartbasierte *Selection* zwar das Laufzeitverhalten gegenüber einem einfachen *max pruning*, bleibt jedoch im Laufzeitverhalten noch hinter einem paarweisen Vergleich aller möglichen *Re-use Units* untereinander zurück. Es bleibt an dieser Stelle Teil der offenen und zukünftigen Forschung im Bereich des *Text Re-use Mining*, dieses Laufzeitverhalten zu verbessern, wobei sich etwaige Antworten bereits aufzeigen. In den *Linking*-Analysen dieses Abschnittes wurde immer nur ein *Atom* fester Länge, wie das *Trigram Shingling* oder das *Word based Featuring*, behandelt. In Abschnitt 5.4.4 konnte auf Basis der Ergebnisse von 24 Probanden zur Näherung des *Re-use Nucleus* festgestellt werden, dass bestimmte Cluster von Wortarten als zusammengesetzte *Features*, wie die Substantiv-Verb- oder die Adjektiv-Substantiv-Verbindung, betrachtet werden können. Einerseits muss die weitere Forschung von einem *Featuring* mit konstanter Länge, welches wiederum sehr einfach zu bestimmen ist, hin zu *Features* mit unterschiedlicher Länge verlagert werden. Andererseits zeigt insbesondere Tabelle 5.19 auf, dass der *Brute Force Score* bei zusammengesetzten *Atomen* deutlich sinkt. So kann bspw. davon ausgegangen werden, dass bei einer Substantiv-Verb-Verbindung das gemeinsame Auftreten in der Regel höchstens einem Prozent der Wortfrequenz des häufigeren Wortes entspricht, wodurch das *Linking* für diese Verbindung auf unter etwa 10^{-4} reduziert wird.

5.5 *Noisy Channel Mining*: Extraktion paradigmatischer und historischer Schreibweisen

Historische Texte waren im Laufe der Zeit zahlreichen Einflüssen ausgesetzt. Einerseits veränderte ein Editor mehr oder weniger stark einen Text. Auf der anderen Seite wurden die Texte durch die Textüberlieferung, bspw. durch abschreibende Mönche, bereits in der Vergangenheit modifiziert. Da es das intrinsische Interesse einer *Text Re-use Analysis* ist, unabhängig vom Grad der Veränderung bzw. im Sinne der *Conditional Kolmogorov Complexity* unabhängig von der Größe des minimalsten Programmes $\mathcal{P}_{min}$ eine *Source* auf jedes mögliche *Target* zu linken, stellen insbesondere stärkere Veränderungen eine Herausforderung dar. Einerseits kann zwar, wie in der *System Evaluation* auf verschiedene Ressourcen zurückgegriffen werden, jedoch ist oftmals die nötige Qualität bzw. das *Coverage* auf einer zu untersuchenden *Digital Library* fraglich (vgl. die verschiedenen *Coverage* aus der *Component Evaluation* in Abschnitt 5.4).

Das *Noisy Channel Theorem* von Shannon kann jedoch nicht nur dazu verwendet werden, wie in Kapitel 4 definiert und in der *System Evaluation* aus Abschnitt 5.3 eingesetzt, die *Mining Ability* eines Verfahrens auf einer bestimmten *Digital Library* zu bestimmen. Vielmehr kann der *Noisy Channel* auf systematische Veränderungen, dem *Noisy Channel Mining* (vgl. *Text Re-use Tasks* in Abschnitt 2.7 auf Seite 79) untersucht werden. Bezugnehmend auf das initiale Beispiel aus dem Buch Genesis, Kapitel 1, Vers 1 (vgl. Tabelle 5.2 auf Seite 177) sollen paradigmatische Relationen, wie *make*, *prepare* und *create* oder auch *heavens* und *heaven*, extrahiert werden.

Dem *Noisy Channel Mining* liegt zuvor eine *Text Re-use Analysis* zugrunde, so dass ein *Text Re-use Graph* als gegeben angesehen werden kann, der paarweise *Re-use Units* aufeinander verlinkt, die bzgl. des *Scoring*-Schwellwertes t als hinreichend ähnlich bezeichnet werden können. Für das *Noisy Channel Mining* empfiehlt sich ein höherer *Scoring*-Schwellwert von $t \geq 0.8$, da sonst die genauen Grenzen einer Veränderung meist nicht durch den Computer nachvollziehbar sind.

Das *Noisy Channel Mining* kann sich methodisch, wie in der *Microview* aus Abb. 1.1 (vgl. Seite 35), vorgestellt werden, wobei gezielt Veränderungen, wie *Substitutions*, *Deletion* und *Insertion*, Gegenstand der Analyse sind. Jede dieser Veränderungen vergrößert im Kontext der *Conditional Kolmogorov Complexity* das minimalste Programm $\mathcal{P}_{min}$. In Anlehnung an die *Microview* aus Abb. 1.1 werden nachfolgend durch ein *Noisy Channel Mining* all diejenigen *Substitutions* zweier *Re-use Units* extrahiert, welche sowohl links als auch rechts ein gemeinsames *Feature* besitzen. Diese Form des *Sequence Alignment* ist sehr passiv gewählt, da bspw. eine Mehrwort-Ersetzung unberücksichtigt und damit der *Recall* vergleichsweise niedrig bleibt. Jedoch ist eine hohe *Precision* durch dieses Vorgehen gegeben, so dass die Möglichkeiten des *Noisy Channel Mining* besser eingeschätzt werden können.

Tabelle 5.22 reflektiert die Ergebnisse des zuvor geschilderten *Noisy Channel Mining*. Insgesamt werden auf diese Weise knapp 12000 Kandidaten für eine paradigmatische Relation extrahiert. 8193 sind aufgrund signifikanter Merkmale auf die neun Relationstypen in Tabelle 5.22 semiautomatisch klassifiziert. Die restlichen knapp 4000 ungetypten Beziehungen sind jedoch nicht nur Rauschen, sondern entsprechen vielmehr dem *Long Tail*. So sind in dieser Menge noch zahlreiche Synonyme, wie *punishment* und *torment* enthalten, die nicht in *WordNet* annotiert worden sind. Weiterhin enthält diese Menge zahlreiche Kohyponyme. Stichprobenweise kann festgehalten werden, dass von den 4000 ungetypten Relationen etwa 50% der Kanten einer erwartbaren tieferen semantischen Relation unterliegen.

Relations-typ	**Anzahl der extrahierten Relationen**
Synonym	3066
Inflected variant	989
Similar written word	1245
Hyphen	451
Prefix	545
Suffix	84
Composition	512
Archaic inflected variant	669
Archaic synonym	632
Sum	8193

Tabelle 5.22: Ergebnisse des *Noisy Channel Mining* auf sieben englischsprachigen Bibelversionen. Die Ergebnisse einer *Noisy Channel Mining*-Analyse sind insgesamt in neun verschiedene Relationstypen klassifiziert worden. Insgesamt sind 8193 der etwa 12000 extrahierten Assoziationen klassifiziert.

Tabelle 5.22 reflektiert, dass von den insgesamt 8193 klassifizierten Relationen mit 4055 extrahierten Assoziationen in etwa die Hälfte auf diverse *Synonyme* und morphologischen Varianten entfallen, die durch Datenbasen, wie *WordNet* bzw. morphologischen Tools, abgedeckt werden können. Die restlichen 4138 extrahierten Datensätze entfallen im Wesentlichen auf historische bzw. editorspezifische Varianten.

Archaic inflected variants ist eine Klasse, welche eine archaische auf eine moderne Wortform, wie *couldeth* auf *could*, transformieren kann. Grundlage für diese Klasse sind einige zusätzliche morphologische Regeln, auf deren Basis überprüft wird, ob eine entsprechende moderne Grundform ableitbar ist. Ähnlich verhält es sich mit der Klasse *Archaic synonym*, welche Einträge enthält, die durch das Trimmen einer archaischen Morphologie zu einem in *WordNet* enthaltenen *Synonym*-Relation abgeleitet werden kann. Diese beiden Klassen sind vor allem für die *King James Version* signifikant.

Die Klasse *Composition* enthält diejenigen paradigmatischen Relationen, die zwei Kriterien erfüllen. Erstens, es muss ein Bindestrich enthalten sein. Zweitens, mindestens ein Teilwort, welches durch den Bindestrich getrennt ist, muss ein Kandidatenpaar gemeinsam haben. Beispiel hierfür sind u. a. *sea-beast* vs. *sea-monster*, *sea-gull* vs. *sea-mew* vs. *sea-hawk* oder auch *apple-tree* vs. *citron-tree*. In diesen Klassen befinden sich unabhängig von den einzelnen Bibelversionen hauptsächlich *Kohyponyme*.

Die Klasse *Hyphen* fasst diejenigen Kandidatenpaare zusammen, die einen eingefügten Bindestrich enthalten. Beispiele hierfür sind u. a. *birth-day* vs. *birthday*, *back-bone* vs. *backbone*, *zareth-shahar* vs. *zarethshahar*. Diese Klasse ist bezogen auf die sieben unterschiedlichen *Bibelversionen* von besonderem Interesse, da jeder Editor bei anderen Wörtern derartige Einfügungen von Bindestrichen hinzufügt bzw. je nach Vorlage einer anderen Version auch entfernt. In *Webster's Revision* werden bspw. *birth-day* und *back-bone* im Vergleich zu den anderen Bibelversionen benutzt. In der *WEB*-Version wird *Bethlehem* 31-mal durch *Beth-lehem* beschrieben. Für die *Young Literal Translation* wird *first-fruits* insgesamt 29-mal benutzt.

Die beiden Klassen *Prefix* und *Suffix* decken paradigmatische Relationen ab, zu welchen ein Wort Teil des anderen Wortes ist, wie bspw. *ambush* vs. *ambushment*, *shimite* vs. *shimites* oder auch *bearing* vs. *childbearing*.

Alle weiteren stringähnlichen Wortpaare sind in der Klasse *Similar written word* zusammengefasst. Diese Klasse umfasst in erster Linie verschiedene Schreibweisen eines Wortes, wie *anathothite* vs. *anethothite* vs. *anetothite* vs. *annethothite* vs. *antothite*. Der Umfang von 1245 verschiedenen Wortpaaren (vgl. Tabelle 5.22) zeigt deutlich, dass dieser drittgrößten Klasse ein besonderer Stellenwert zugesprochen werden muss.

Trotz des relativ einfachen *Sequence Alignment* werden bereits verschiedene Veränderungen im *Noisy Channel* festgestellt, die entweder historisch oder editorspezifisch sind. Dennoch bleibt an dieser Stelle auch festgehalten, dass das Potential des *Noisy Channel Mining* mit dem hier vorgeschlagenen Ansatz noch nicht voll ausgeschöpft wird, da nur Einwort-Ersetzungen Gegenstand der Analyse sind. Oftmals kann jedoch das Gegenteil beobachtet werden. Ein Beispiel hierfür ist durch das Buch Genesis, Kapitel 34, Vers 19 gegeben. In *ASV*, *KJV* und *WBS* wird der temporale Ausdruck *not defer* benutzt. In *DBY* sowie *YLT* wird *not delay* und in der *WEB*-Version *not wait* verwendet. Auch wenn es noch möglich scheint, diese drei temporalen Ausdrücke aufeinander zu mappen, so wird es ohne *Noisy Channel Mining* nahezu unmöglich, den Ausdruck *without loss of time* in *BBE* auf die zuvor genannten Varianten abzubilden. Auch wenn der vorgestellte Algorithmus das derzeit nicht kann, so soll das *Noisy Channel Mining* als ein zukünftiges Forschungsfeld des *Historical Text Re-use* verstanden werden. Einerseits wird dadurch direkt der fachwissenschaftliche Schwerpunkt der Textkritik bedient. Andererseits hilft das Wissen darüber, dass *not wait*, *not delay*, *not defer* sowie *without loss of time* den gleichen Ausdruck beschreiben, diese im Kontext der *Digital Signature* als ein mehrgliedriges *Feature* zu verstehen sind. So kann nicht nur die Geschwindigkeit sondern auch die *Precision* verbessert werden.

Abschließend stellt sich das *Noisy Channel Mining* im Kontext einer *Text Re-use Analysis* zukünftig als ein Teil eines iterativen Prozesses dar. Hierzu wird in einem ersten Schritt eine *Text Re-use Analysis* mit einem hohen *Scoring*-Schwellwert t durchgeführt, welche sich dem Schritt des *Noisy Channel Mining* anschließt. Das *Noisy Channel Mining* kann hierbei als ein *Postprocessing*-Schritt verstanden werden. Nach dem *Noisy Channel Mining* werden die extrahierten Daten in einer zweiten Iteration als Erweiterung des *Preprocessing* einer *Text Re-use Analysis* eingesetzt, um die in Texten enthaltenen Varianten zu normalisieren. Dieser Prozess kann mehrfach wiederholt werden, wodurch eine *Text Re-use Analysis* nicht nur als ein automatisches Instrument der *Intertextuality*, sondern auch als eine Methode des vorwissensfreien Extrahierens von historischen Varianten verstanden werden kann.

5.6 Zusammenfassung

Dieses Kapitel zeigt, dass eine *Re-use Analysis* komplex ist. Durch die *7-Level-Architektur* des *Historical Text Re-use* (vgl. Kapitel 3, Seite 87 ff.) ist es möglich, die verschiedenen Level, wie *Preprocessing*, *Featuring* und *Selection*, auf Abhängigkeiten zu untersuchen. In diesem Kapitel wurde eine *System Evaluation* von zwölf Experimenten zu unterschiedlichen *Preprocessing*- und *Featuring*-Techniken durchgeführt. Die Wahl von sieben verschiedenen Bibelversionen als Datenbasis war nicht zufällig. Jede Bibel entstand aus einem Archetyp, wobei hinter jeder Version eine von dem Autor spezifische Intention steht. Bedingt vom Interesse des jeweiligen Editors sind sich die gleichen Verse in zwei Bibeln ähnlicher bzw. unähnlicher als andere, da nur kleinere Änderungen, wie in der *Webster's Revision* der *King James Version*, vorgenommen worden sind. Andere Interessen, wie die zur *Bibel in Basic English*, welche in einer sehr einfachen Sprache geschrieben wurde, oder auch die zur *Young Literal Translation*, die sich an der hebräischen Syntax orientiert, verändern Verse deutlich stärker, so dass bereits eine *Paraphrase* vorliegt.

Es konnte während der *System Evaluation* gezeigt werden, dass diese Änderungen des minimalsten Programmes im Sinne der *Conditional Kolmogorov Complexity* $\mathcal{P}_{min}$ einen enormen Einfluss auf das Ergebnis einer *Text Re-use Analysis* haben. Im Ergebnis kann daher festgehalten werden, dass in Abhängigkeit von der Intention bei der Erstellung einer neuen Bibelversion, teilweise stark abweichende Techniken und Parameter für die *Text Re-use Analysis* nötig sind. So konnte während der *System Evaluation* gezeigt werden, dass ein *Scoring*-Schwellwert t von der ausgewählten *Featuring*-Technik abhängig ist. Auf der anderen Seite konnte auch herausgearbeitet werden, dass der *Scoring*-Schwellwert t in einer paarweisen *Text Re-use Analysis* zweier Bibelversionen bei unterschiedlichen und sogar stark abweichenden Werten das Maximum im *F-Measure* erreicht, was wiederum in erster Linie in der Nähe zweier Bibelversionen begründet liegt. So konnte der höchste Schwellwert für *Webster*'s Revision der *King James Version* mit ebendieser bestimmt werden, während eine stark paraphrasierte Version, wie die *Bibel in Basic English*, im Vergleich mit *KJV* einen deutlich geringeren *Scoring*-Schwellwert hat, zu welchem F maximal wird. Weiterhin ist es ein Ergebnis dieses Kapitels, dass das *F-Measure* ebenfalls zu unterschiedlichen *Featuring*-Techniken maximal wird.

All diese *Diversity* natürlicher Sprache im Gebrauch geben der Forschung im Bereich des *Text Re-use* eine umfangreiche Komplexität auf. Vielmehr kann aus diesem Kapitel entnommen werden, dass es kaum eine generelle Gesetzmäßigkeit gibt. Weiterhin gibt es sehr viele autor- und editorspezifische Abhängigkeiten. Das größte Problem, welches sich aus der *Diversity* ergibt, ist jedoch, wie in einem solchen Umfeld evaluiert werden kann. Letztlich müsste ein entsprechender *Gold Standard* nicht nur die sprachliche *Diversity* reflektieren, sondern vielmehr auch die verschiedenen *Meme* und *Re-use Styles* (beide vgl. Abschnitt 2.6) in sich vereinen. Offensichtlich wird kein Evaluierungsstandard diese Eigenschaften erfüllen können.

Aus diesem Grund bedarf es zumindest für die Evaluierungen und Optimierungen einer *Text Re-use Analysis* einer rein quantitativen Evaluierung, die keinen *Gold Standard* benötigt. Hierzu wurde in diesem Kapitel die *Text Re-use Compression* sowie die *Noisy Channel Evaluation* getestet. Im Ergebnis konnte erarbeitet werden, dass es eine nahezu perfekte Korrelation zwischen der *Text Re-use Compression* und dem *Recall* gibt, wobei ersteres Maß keinen *Gold Standard* benötigt, während letzteres von einer Evaluierungsbasis abhängig ist. Weiterhin konnte eine hinreichende Korrelation nach Pearson zwischen dem *F-Measure* und der *Noisy Channel Evaluation* festgestellt werden. Insbesondere das Verhalten beider Evaluierungstechniken in Abhängigkeit des *Scoring*-Schwellwertes t ist auffällig gleich. So lieferte eine Evaluierung durch beide Techniken ein nahezu identisches Ergebnis,

welches auch in Abwesenheit eines *Gold Standards* bei der *Noisy Channel Evaluation* nicht selbstverständlich ist. Für das Optimieren von Parametern scheint die *Noisy Channel Evaluation* besonders gut geeignet zu sein. Nichtsdestotrotz bedarf es für die *Precision* einer nicht vollautomatischen Technik. Hierzu wurde vorgeschlagen, dass das Ergebnis durch eine manuelle Durchsicht einer zufälligen Stichprobe evaluiert wird.

Bei der grundlegenden Frage nach einer derartigen *System Evaluation* bleibt jedoch oftmals unklar, was genau verbessert werden kann. Hierzu wurde im Rahmen einer *Component Evaluation*, die aus der *Biometrie* adaptiert worden ist, aufgezeigt, wo genau Schwachstellen während der *System Evaluation* liegen, so dass punktuelle Verbesserungen möglich sind. Das umfasst einerseits die Fähigkeit, speziell historische Varianten lemmatisieren zu können, andererseits jedoch auch den Umgang mit historischen Schreibweisen. Weiterhin wurde aufgezeigt, wie die *Digital Signature* einer *Re-use Unit* optimiert werden kann. Das ist einerseits durch ein Herabsetzen der *Feature Density* möglich, wobei dadurch ein sinkender *Recall* zu erwarten ist. Andererseits konnte durch ein wortartspezifisches *Selection* weiterführend gezeigt werden, wie sich zu einer *Text Re-use Analysis* bei vergleichbaren Evaluierungsergebnissen die Geschwindigkeit deutlich verbessert.

Schließlich zeigen die Ergebnisse dieses Kapitels auch, dass eine *Text Re-use Analysis* immer einen Kompromiss aus gewünschtem *Recall*, und somit Quantität, sowie dem Laufzeitverhalten darstellt. Anhand des *Brute Force Score* konnte dargestellt werden, dass selbst bei einem *Pruning* der häufigsten Wörter schnell ein Vielfaches an *Linking*-Kosten gegenüber eines *Brute Force Linking* erzeugt werden kann, so dass relativ einfach Laufzeiten von mehreren Prozessormonaten bzw. -jahren nötig sind.

Um das Laufzeitverhalten verbessern zu können, wurde weiterhin vorgeschlagen, dass sich auf Basis eines Tests zur Näherung des *Re-use Nucleus* zukünftig von einer statischen Größe eines *Atoms* entfernt werden sollte. Spezielle Muster, wie Substantiv-Verb-Verbindungen, empfehlen sich in besonderer Weise für zusammengesetzte *Features*, wodurch die *Linking*-Kosten aufgrund einer kleineren *Feature*-Frequenz reduziert werden können und sich somit das Laufzeitverhalten verbessert.

Abschließend wurden Ergebnisse eines *Noisy Channel Mining* vorgestellt. Hierbei werden Varianten eines Wortes in Form von verschiedenen Schreibweisen, Synonymen aber auch Kohyponymen oder editorspezifischen Varianten durch eine unterschiedliche Benutzung von Bindestrichen extrahiert.

Zusammenfassend kann konstatiert werden, dass dieses Kapitel trotz seines hohen Umfangs bei Weitem nicht alle Aspekte ausleuchten konnte. Dennoch ist aufgezeigt worden, dass die *Text Re-use Analysis* durch die große Volatilität von Menschen in der Benutzung von Texten nicht nur eine Anwendung im Sinne der *Intertextuality* darstellt, sondern vielmehr ein Forschungsbereich ist, welchem zukünftig wesentlich mehr Aufmerksamkeit gewidmet werden muss. Insbesondere auf dem Erforschen der *Minutiae*, den relevanten Hauptkomponenten einer *Digital Signature*, muss der Fokus dieser wissenschaftlichen Arbeiten liegen, wodurch sowohl die Qualität als auch die Geschwindigkeit optimiert werden können.

Kapitel 6

Zusammenfassung

Contents

6.1 Ziele und Ergebnisse dieser Arbeit 218
6.2 Lessons Learnt 225
6.3 Weiterführende Aspekte und zukünftige Arbeiten 227

Aus vagen Vorstellungen wurden klare Ziele,

aus Kampf – Gelassenheit,

aus Fremd – Verständnis.

Swetlana Reiche, (1971-)

Die Arbeit stellt die Grundlagen des *Historical Text Re-use* zusammen. Im Rahmen dieses Kapitels werden die Ergebnisse und Fortschritte sowohl der Dissertation als auch deren Bedeutung für die Wissenschaft in der *Computer Science*, den *eHumanities* aber auch den *Humanities* reflektiert. Das Kapitel ist in drei Abschnitte gegliedert, welche ausgehend von den Ergebnissen dieser Dissertation hin zu Perspektiven und weiteren Arbeiten führen.

6.1 Ziele und Ergebnisse dieser Arbeit

Die Ausgangslage vor dieser Dissertation ist einfach zu definieren. Die Geisteswissenschaften haben manuell Parallelstellen und Zitate in Datenbanken oder gar Excel-Tabellen gesammelt. Die Informatik hat meist relativ einfache Techniken, wie das *Bigram Shingling*, eingesetzt, um Plagiarismus zu erkennen. Insbesondere durch zahlreiche Plagiarismusvorwürfe der letzten Jahre wird zunehmend genauer auf die widerrechtliche Nutzung fremder Texte oder Textabschnitte geachtet.

Während in den fachwissenschaftlichen Datenbanken zwar qualitativ gute Datensätze enthalten sind, so mangelt es oftmals an der nötigen Quantität, um komplexere Zusammenhänge beschreiben zu können. Derartige Beziehungen können Zitierabhängigkeiten und damit auch Wissen über Existenzen von Büchern zu bestimmten Zeitpunkten aber auch Akzeptanz und Ablehnung von Theorien und Wissen sein.

In der Informatik dagegen ist es für einen hinreichenden Plagiarismusvorwurf nicht nötig, alle plagiierten Textstellen aufzudecken, sondern es reicht bereits 60% oder 70% des möglichen *Text Re-use* aufzuzeigen, da letztlich Plagiarismus nicht von der Menge selbst, sondern in erster Linie von der Straftat als solche abhängig ist. Dementsprechend sind die Methoden oftmals sehr rudimentär und können insbesondere stärker veränderte aber dennoch kopierte Argumentationsstrukturen bereits nicht mehr erkennen.

Gravierender neben methodischen Problemen auf beiden Seiten ist jedoch, dass beide Welten bisher unabhängig von einander koexistiert haben. Ein guter Indikator hierfür sind benutzte Terminologien. Während *Text Re-use* ein aus den *Computer Science* geprägter Begriff ist, welcher das Wiederverwenden einer textbasierten Information beschreibt, ist der vergleichbarste Begriff aus den *Humanities* zu diesem Thema die *Intertextuality* (vgl. [Allen 2011]). Vielmehr existiert eine ganze Taxonomie von Abhängigkeiten zwischen Textstellen, so dass bspw. auch die Terminologien *Intertextuality* und *Hypertextuality* unterschieden werden (vgl. [Riffaterre 1994]).

Mit den *eHumanities* sind insbesondere durch großzügige Förderungen beide Welten in den letzten zehn Jahren deutlich näher zusammengerückt. In der *Scientific Community* (vgl. Abb. 2.1 auf Seite 58) gibt es immer noch vier verschiedene Sichtweisen, aus denen das Thema des *Text Re-use* betrachtet werden kann. In den *Humanities* steht die Textkritik im Vordergrund, welche wiederum voraussetzt, dass möglichst alle Parallelstellen, also *Text Re-use*, gefunden werden. Dies ist jedoch weder mit manuell erstellten Datenbanken aufgrund der geringen Datenmenge möglich, noch hilft eine Suchmaschine dabei, dies effizient tun zu können, da letztlich jedes einzelne Vorkommen eines Wortes inklusive aller morphologischen Flektionen für einen *Text Re-use Candidate* vom Fachwissenschaftler manuell durchgearbeitet werden muss. In den *Digital Humanities* steht die Frage im Vordergrund, wie unabhängig von der Methodik, also manuell oder automatisch, der *Text Re-use* gespeichert und archiviert werden kann. Die *Computer Science* muss sich letztlich mit oftmals einfachen und pragmatischen sowie meist mehr Forschungssoftware als auf ernsthafte Produkte aufsetzenden Lösungsansätzen der Anforderung einer *Data Diversity* stellen, die die entsprechend frühen und unausgereiften Softwareversionen bei Weitem nicht erfüllen können.

Insofern kann das Thema des *Historical Text Re-use* im Rahmen der *eHumanities* in besonderer Weise durch die verschiedendsten Interessen bedient werden, so dass, wenn bisher auch nicht ernsthaft geschehen, eine natürliche Überlappung im Interesse am *Text Re-use* ausgemacht werden kann. Die Grundsteinlegung des *Historical Text Re-use* als eigenständigen Forschungsbereich im Rahmen dieser Arbeit ist daher nicht zufällig mit zwei weiteren und die *Community* zusammenführenden Instrumenten verbunden. Auf der einen Seite wurde die zentrale Rolle dieser Arbeit für die *Scientific Community* durch zahlreiche An-

fragen insbesondere aus den *Humanities* und *Digital Humanities* deutlich. Als Konsequenz dessen hat der Autor im November 2011 eine *Google Group* mit dem Namen *Historical Text Re-use*[1] gegründet, um ein offenes und transparentes Kommunikationsforum für die zusammenwachsende *Community* bereitzustellen, der zum Zeitpunkt dieser Textsetzung etwa 50 nationale und internationale Forscher mit gemeinsamen Interessen am Thema angehören. Auf der anderen Seite hat diese Arbeit die Erstellung des ersten *Historical Text Re-use Glossary* begleitet. Die eingangs bereits genannten und an einem Beispiel aufgezeigten terminologischen Schwierigkeiten haben in der Vergangenheit eine fächerübergreifende Kommunikation deutlich erschwert, da oftmals Gleiches gemeint war, jedoch unterschiedliche Terminologien genutzt worden sind. Das ausgewiesene Ziel dieses Glossars ist es, nicht nur ein Fundament für eine gemeinsame Terminologie zu legen, sondern in Anlehnung an die Erfahrungen aus der Biometrie (vgl. [BIMA 2012, NSTC 2006]) auch verschiedene und standardisierte Tests zu formulieren. In Kapitel 5 wurde hierzu bereits, wenn auch noch unvollständig, eine Evaluierungstaxonomie, bestehend aus den zwei Hauptklassen, der *System Evaluation* sowie der *Component Evaluation*, vorgestellt (vgl. Abb. 5.1 auf Seite 169). Die Version 1.0 des *Historical Text Re-use Glossary* wird hierzu kurz nach Fertigstellung dieser Arbeit am 5./6.-April-2013 im Rahmen einer Konferenz der *Digital Classics Association*[2] in Buffalo, NY, USA, erstmals öffentlich vorgestellt.

Sowohl das *Historical Text Re-use Glossary* als auch die *Google Group* stellen, wenn auch im Rahmen dieser Arbeit nur in der Zusammenfassung erwähnt, das Fundament dar, da so der Fokus dieser Dissertation immer mit dem Blick auf eine *Scientific Community*, bestehend aus mehreren Disziplinen und Forschungsinteressen, gegeben ist. Ergänzend zu diesen beiden Instrumenten wurde im Rahmen dieser Arbeit das *ACID for the eHumanities* Paradigma formuliert (vgl. Abschnitt 1.5 auf Seite 38 ff.). Ziel des Paradigmas, ähnlich dem Wasserfall-Modell oder der *SWOT*-Analyse, ist es, durch ein gezieltes und systematisches Formulieren von Fragen interdisziplinäre Spannungen zu vermeiden. Beim *ACID for the eHumanities* Paradigma wird explizit nach der *Acceptance* in den Fachwissenschaften für eine IT-Methode, nach der *Complexity* einer Aufgabe, der *Interoperability* von Daten untereinander sowie Daten und Verfahren als auch nach der *Diversity* von fachwissenschaftlichen Daten gefragt.

Das Zusammenbringen der *Humanities* mit der *Computer Science* birgt aber über interdisziplinäre Spannung (vgl. Abschnitt 1.3 auf Seite 31 ff.) auch systematische Gefahren, die insbesondere im Abschnitt 1.4 (vgl. ab Seite 32) ausgeführt worden sind. Auf der einen Seite generieren *Mining*-Methoden im Allgemeinen und eine *Text Re-use Analysis* im Speziellen eine Fülle an Daten, die faktisch kein Geisteswissenschaftler systematisch bearbeiten kann, so dass ein fachwissenschaftlicher *Information Overload* erzeugt wird. Hierzu sind spezielle Visualisierungen, die teilweise auch Daten aggregieren, wie in Plots und Verläufen, nötig. In Abschnitt 1.4 werden hierzu insgesamt vier verschiedene Visualisierungen und ihre Einsatzgebiete in den *Humanities* vorgestellt. Auf der anderen Seite geht mit den *Distant Reading* Visualisierungen einher, dass sie die Hermeneutik bzw. die Interpretation eines Sachverhaltes in eine falsche Richtung beeinflussen können. Insbesondere bei zeitabhängigen Visualisierungen, wie Abb. 1.4 auf Seite 38, muss immer die Frage gestellt werden, ob ein *Peak* auch bei einem vollständig erhaltenen Textkanon gegeben sein würde. Aufgrund dessen, dass über die Jahrzehnte und für die Antike gar Jahrtausende Texte immer wieder abgeschrieben werden mussten, um vor dem Verfall aber auch menschlicher Zerstörung geschützt zu sein, existieren viele Werke nicht mehr. Das führt zu einer *Information Poverty*. Die *Information Poverty* stellt insbesondere für *Distant Reading* Visualisierungen des

[1] vgl. `http://groups.google.com/group/historical-text-re-use/`

[2] Die Konferenz *Word, Space, Time - Digital Perspectives on the Classical World* wird von Neil Coffee organisiert. Link: `http://classics.buffalo.edu/events/dcaconference/`

Historical Text Re-use die Gefahr der systematischen Fehlinterpretation von Daten dar.

In Kapitel 2 sind grundlegende Definitionen des *Historical Text Re-use* festgelegt. In Anlehnung an die biometrische Analyse werden in Abschnitt 2.4 Qualitätskriterien definiert, die für die folgenden Abschnitte und Kapitel dieser Arbeit Grundlage sind.

In den Abschnitten 1.7 und 2.8 wird die für diese Arbeit grundlegende Einbettung des *Historical Text Re-use* in Shannon's *Noisy Channel Model* fundiert und erklärt. Die Einbettung ist dadurch gegeben, dass ein Werk eines zitierten bzw. älteren Autors als *Source* und eines jüngeren Werkes als *Target* angesehen werden kann. Aufgrund der enormen Zeitausprägungen von mehreren Jahrhunderten oder gar Jahrtausenden haben sich Konzepte semantisch verändert, Wörter sind einer Sprachevolution ausgesetzt gewesen oder durch regionale Dialekte unterschiedlich geschrieben worden. Weiterhin war es nicht selten der Fall, dass ein abschreibender Mönch durch fehlende Lesbarkeit von Buchstaben oder Wörter aber auch durch ein fehlendes Verständnis eines Wortes, Fehler während des Kopierprozesses eingefügt hat.

Dieses teilweise hochvolatile System kann durch den *Noisy Channel* modelliert werden (vgl. Abschnitt 2.8 ab Seite 83). Hierbei stellt sich die grundlegende Frage, wie viele Änderungen von einem Original ausgehend vorgenommen werden können, so dass eine *Mining*-Methode *Source* und *Target* miteinander verlinken kann. Im genannten Abschnitt wird insbesondere auf das Zusammenspiel zwischen dem Grad des Rauschens im *Noisy Channel* sowie des minimalsten Programmes $\mathcal{P}_{min}$ im Kontext der *Conditional Kolmogorov Complexity* eingegangen. Die *Conditional Kolmogorov Complexity* kann als ein Pseudoprogramm verstanden werden, welches systematische Veränderungen im *Noisy Channel*, wie eine *Substitution*, ein *Insertion* oder einem *Deletion*, aufzeichnet.

Neben der grundlegenden Frage, wie groß das entsprechende minimalste Programm bzw. wie groß die *Data Diversity* sein darf, so dass der *Text Re-use*, also die Beziehung zwischen einem *Source* und einem *Target*, erkannt werden kann, ergeben sich zwei für die Arbeit fundamentale Aspekte.

Erstens, kann, wie in Abschnitt 4.4 eingeführt, das *Noisy Channel Model* dazu eingesetzt werden, ein künstliches Störsignal in den *Noisy Channel* einzuführen, welches in Abschnitt 4.5 ab Seite 149 in insgesamt fünf verschiedene Randomisierungsklassen eingeteilt ist. Die Klassen können als verschiedene Abstraktionen im Sinne eines *Turing*-Tests verstanden werden. Während ein *T1*-Randomisierer sehr einfach von natürlicher Sprache zu unterscheiden ist, imitiert ein *T5*-Randomisierer bestmöglich natürliche Sprache. Je nachdem welcher Randomisierer ausgewählt wird, kann unterschiedlich gut zwischen Zufall und natürlichsprachlicher Struktur durch eine *Mining*-Methode, wie die *Text Re-use Analysis*, unterschieden werden. In Abschnitt 4.4 wird weiterführend ein Versuchsaufbau vorgestellt, welcher durch einen Randomisier in einem ersten Schritt eine *Randomised Digital Library* erzeugt (vgl. Abb. 4.3 auf Seite 143). Anschließend wird eine *Text Re-use Analysis* zu einem Parameterraum Θ sowohl auf der *Digital Library* als auch der *Randomised Digital Library* ausgeführt und beide Ergebnisse quantitativ durch die *Mining Ability* $\mathcal{L}_\Theta$ miteinander verglichen (vgl. *Mining Ability* aus Definition 21 auf Seite 144). Der *Score* der *Mining Ability* kann somit als die Fähigkeit einer *Text Re-use Analysis* angesehen werden, um zwischen Ergebnissen auf realen Daten und dem zufälligen Grundrauschen einer Methode unterscheiden zu können, so dass dieser Methode in Abschnitt 4.4 der Name *Noisy Channel Evaluation* gegeben wurde. Die Besonderheit dieser Form der Evaluierung ist, dass keine Evaluierungsbasis bzw. kein *Gold Standard* notwendig ist.

Zweitens, es kann aus der Einbettung des *Historical Text Re-use* in das *Noisy Channel Model* auch ein Formalismus geschaffen werden, durch welchen die Änderungen im Sinne des minimalsten Programmes $\mathcal{P}_{min}$ der *Conditional Kolmogorov Complexity* nicht nur aufgezeichnet werden können, sondern wiederum auch Gegenstand der Forschung sind. In

Abschnitt 5.5 wird hierzu die Methode des *Noisy Channel Mining* eingeführt, welche u. a. paradigmatische Relationen von historischen Schreibweisen und Ersetzungen untersuchen kann. Vielmehr ermutigen die Ergebnisse des *Noisy Channel Mining* in Abschnitt 5.5 dazu, dieses in den *eHumanities* bisher unberücksichtigte Forschungsfeld zukünftig weiter auszubauen. Weiterhin kann das wahre Potenzial in einem Vergleich mit *WordNet* aufgezeigt werden. Während *WordNet* semantische Beziehungen zwischen Konzepten der Gegenwart repräsentiert, stellen sich Konzeptbeziehungen in ihrer Art über größere Zeiträume als dynamisch dar. Als Beispiel wurde in dieser Arbeit die Beziehung zwischen *Bier* und *Wein* gebracht. Während beide Konzepte in der Gegenwart kohyponym zu verstehen sind, sind sie in der Antike synonym benutzt worden. Neben solchen Veränderungen in der semantischen Beziehung, die zwangsläufig mit großen Zeitperioden von in einer *Digital Library* enthaltenen Texten einhergeht, kann insbesondere keine Datenbasis alle Varianten oder auch nur eine annähernd repräsentative Menge bereitstellen. Aus diesem Grund, und wie Prof. Crane auch einmal darauf verwiesen hat, dass wir für die antiken Sprachen keine Muttersprachler haben, die wir einfach fragen können, ist das *Noisy Channel Mining* nicht nur Gegenstand, die sprachliche Vielfalt einer *Digital Library* zu bestimmen und während des *Preprocessing* dementsprechend zu normalisieren, sondern auch ein automatisches Werkzeug, um entsprechende *WordNets* für antike Sprachen aufzubauen. Hierbei ist die *Diversity* der Sprache nicht nur Segen, sondern vielmehr helfen uns zitierende Autoren, die teilweise bereits vor mehreren Jahrhunderten gestorben sind, dabei, antike *WordNets* in der Gegenwart systematisch aufzubauen. Weiterhin kann das *Noisy Channel Mining* als das naheliegendste automatisierte Modell angesehen werden, welches die fachwissenschaftliche Textkritik umsetzt.

Die eben bereits erwähnte *Data Diversity* von historischen Daten in einer *Digital Library* stellt nicht nur eine enorme Herausforderung für eine *Text Re-use Analysis* dar, da entsprechende Modifikationen das Erkennen von Gleichem erschwert, sondern ist auch zentraler Gegenstand dieser Arbeit. Das Problem der *Diversity* ist neben den bereits genannten Aspekten der Sprachevolution von Wörtern, verschiedenen Dialekten, historischen Schreibvarianten aber auch Rechtschreibfehlern noch mächtiger. Im Rahmen dieser Arbeit wurden hierzu zwei Systematiken entwickelt, die einen *Re-use Graph* typisieren.

Erstens, in den Tabelle 2.1 bis 2.5 wurden insgesamt 50 verschiedene *Meme Types* als ein Ergebnis dieser Arbeit zusammengetragen (vgl. Abschnitt 2.6 Seite 68 ff.). *Text Re-use* geschieht nicht zufällig und willkürlich. Die verschiedenen *Meme Types*, wie *Proverb*, *Definition*, *Legend* oder *Battle Cry*, zeichnen sich alle durch eine unterschiedliche Intention aus. Sie können als Klassen von Instanzen für *Text Re-use* verstanden werden. Jede dieser Klassen besitzt bzgl. Länge, Art der Kommunikation, wie mündlich oder schriftlich, oder auch die literarische Klassifikation, in welcher das *Meme* vorzugsweise oder alleinig auftritt, spezifische Eigenschaften. *Battle Cries* tendieren dazu, sehr kurz, dafür aber stark wortwörtlich, wiedergegeben zu werden. *Meme* vom Typ *Definition* sind durchschnittlich deutlich länger als *Battle Cries*, haben aber dennoch gemeinsam, dass sie meist sehr wortwörtlich benutzt werden. Ein *Proverb* hingegen hat eine vergleichbare Länge, wie das *Meme Definition*, jedoch wird ein *Proverb* stärker in einen Kontext eingebaut und ggf. auch verändert.

Zweitens, es wurde eine Taxonomie der Kanten für einen *Re-use Graph* aufgestellt (vgl. Abb. 2.3 auf Seite 77). Diese umfasst einen wortwörtlichen Zitierstil, wie *Verbatim*, bis hin zu stark verändernden *Re-use Styles*, wie einer *Allusion* oder *Paraphrase*.

Die hier dargestellte *Diversity* stellt eine *Text Re-use Analysis* im Kontext des *Historical Text Re-use* vor zwei entscheidende Fragestellungen, die im Rahmen dieser Dissertation bearbeitet und gelöst wurden.

Einerseits, zeigt sich deutlich, dass der *Data Diversity* während einer *Text Re-use Analysis* nicht mit einem Algorithmus entgegengetreten werden kann. Während kurze *Meme*,

wie ein *Battle Cry*, eine Segmentierung zu kurzen *Re-use Units* durch einen überlappenden Ansatz, wie einem *Moving Window*, benötigen, ist bei einer Analyse von Bibeln eine versweise Segmentierung zielführend. Weiterhin hängt das ausgewählte *Atom*, wie ein *Bigram* oder ein *Word*, vom *Re-use Style* eines Autors ab. All diese Parameter haben zu der *7-Level-Architektur* für eine *Text Re-use Analysis* auf historischen Texten geführt, welche in Kapitel 3 ausführlich dargestellt wurde. Das im Rahmen dieser Arbeit erstellte *TRACER*-Werkzeug implementiert diese *7-Level-Architektur* und bietet zum Zeitpunkt der Setzung dieses Textes über eine Million Möglichkeiten der Kombination von Implementierungen der einzelnen Level an, wodurch auf einen Großteil aller durch die *Diversity* notwendigen Möglichkeiten eine spezifische *Text Re-use Analysis* zusammengestellt werden kann. Auch wenn zahlreiche Einsatzszenarien dadurch möglich werden, so sind im Kontext eines *Syntactic Re-use*, *Semantic Re-use* bzw. eines *Cognitive Re-use* (vgl. Abschnitt 2.5 ab Seite 64) auch dem Computer Grenzen gesetzt. Während der Computer dafür eingesetzt wird, um die Quantität abzudecken (vgl. Ergebnisse der *System Evaluation* aus Abschnitt 5.3 ab Seite 176), sind die Möglichkeiten bei einem *Cognitive Re-use*, wie dem *Text Re-use* zwischen *like will to like* und *birds of same feather flock together*, deutlich eingeschränkt bis gar nicht möglich. Die Ergebnisse der *System Evaluation* aus Abschnitt 5.3 zeigen jedoch auf, dass bereits eine Paraphrase, wie die Verse der *Bible in Basic English*, gegenüber anderen Versionen, wie denen der *King James Version*, vergleichsweise schlecht aufdeckbar sind. Hier sind insbesondere weiterführende Ergebnisse des *Noisy Channel Mining* nötig, um die *Performance* sukzessive zu verbessern.

Andererseits, wie kann im Kontext der *Diversity* von *Text Re-use* das Ergebnis einer *Text Re-use Analysis* evaluiert werden? In der Vergangenheit sind zwar bereits zahlreiche kleinere und spezifische Datenbanken entstanden, die zur Evaluierung eingesetzt werden könnten, jedoch sind diese oftmals nicht frei zugänglich. Weiterhin repräsentieren sie meist nur ein kleines Partikularinteresse, so dass für eine Evaluierung zwar eine Aussage für die Evaluierungsbasis getroffen werden kann, jedoch die Ergebnisse oftmals durch einen anderen *Gold Standard* meist nicht verifiziert werden können. Diese und weitere Probleme sind Gegenstand von Abschnitt 4.3, in welchem signifikante Nachteile einer Evaluierung gegen einen *Gold Standard*, bspw. durch *Precision*, *Recall* und *F-Measure*, aufgezeigt werden. Insbesondere wird in diesem Abschnitt der Fakt deutlich gemacht, dass die Evaluierungsergebnisse wesentlich von der Qualität der Überlappung aus *Gold Standard* und *Digital Library* abhängig sind. Ist die Überlappung gering, so wird insbesondere der *Recall* schlecht ausfallen. Bei einer starken inhaltlichen Überlappung sind bei gleicher Technik zumindest ein größerer *Recall* und damit verbunden auch ein besseres *F-Measure* die Konsequenz, auch wenn das Verfahren das Gleiche geblieben ist, so dass die wahre Aussagekraft einer auf einem *Gold Standard* basierten Evaluierung immer infrage gestellt sein muss. Weiterhin ist es sowohl systematisch für den Forscher als auch unabsichtlich durch eine zu geringe Überlappung möglich, das Evaluierungsergebnis stark zu beeinflussen. Nicht selten können Ergebnisse selbst auf gleichen oder gar identischen Daten nicht reproduziert werden, was im Sinne der *Circumvention*, einem der in Abschnitt 2.4 definierten Qualitätskriterien für eine *Text Re-use Analysis*, nachteilig ist. Aus dem Grund der *Diversity* von historischen Daten sowie dem aufgezeigten *Circumvention* einer *Text Re-use Analysis* wurde im Abschnitt 3.10 die *Text Re-use Compression* sowie im Abschnitt 4.4 die *Noisy Channel Evaluation* (vgl. Seite 141 ff.) eingeführt. Beide Techniken können eine Aussage über die Qualität des Ergebnisses einer *Text Re-use Analysis* treffen, ohne dabei eine Evaluierungsbasis zu benötigen.

So wird in Kapitel 5 aufgezeigt, dass für einen Vergleich von Ergebnissen einer *Text Re-use Analysis* auch weitestgehend kein *Gold Standard* nötig ist. Das ist insofern auch prinzipiell gegeben, da es für eine Evaluierung meist nicht darauf ankommt, eine Aussage darüber zu treffen, wie hoch eine *Precision* oder ein *Recall* genau ist, sondern wie sich das

Ergebnis zu einer Vergleichsanalyse verhält. Letztlich bedeutet eine gemessene *Precision* von $P_{measure} = 0.8$ nicht, dass die reale *Precision* diesen Wert annimmt. Vielmehr sind bestimmte Werte für *Precision*, *Recall* und *F-Measure* bei einer Evaluierung gegen einen *Gold Standard* als untere Schranken anzusehen, da immer Datensätze existieren, die *true negatives*, welche zwar richtigerweise gefunden wurden, jedoch nicht Teil der *Evaluierungsbasis* sind, so dass die Evaluierungsmetriken prinzipiell immer geringer ausfallen. Dem Vergleich zwischen zwei oder mehreren Verfahren genügen auch die ohne Evaluierungsbasis auskommenden und das quantitative Ergebnis einer *Text Re-use Analysis* evaluierenden Methoden der *Text Re-use Compression* sowie der *Noisy Channel Evaluation*.

Im Abschnitt 5.3 konnten hierzu zwei direkte und starke Zusammenhänge bestimmt werden. Zum einen konnte sowohl auf Basis der Ergebnisse einer *System Evaluation* als auch einer Analyse in Abhängigkeit vom *Scoring*-Schwellwert t mit $t \in [0, 1]$ gezeigt werden, dass mit einer Korrelation von über $\rho \geq 0.97$ eine Abhängigkeit zwischen dem *Recall* R sowie der *Text Re-use Compression* besteht. Zum anderen zeigt der gleiche Versuchsaufbau auf, dass es einen starken Zusammenhang und in vielen Belangen auch ein sehr ähnliches Verhalten zwischen dem *F-Measure* sowie der *Mining Ability* der *Noisy Channel Evaluation* gibt. Dies umfasst neben einer starken Korrelation auch gleiche Eigenschaften, wie das Verändern des maximalen *F-Measure* F_{max} und der maximalen *Mining Ability* bzgl. dem *Scoring*-Schwellwert t als Variable (vgl. Abbildungen 5.8 und 5.9 auf den Seiten 196 und 197).

Im Detail haben die Evaluierungen aus Kapitel 5 gezeigt, dass bezogen auf die *7-Level-Architektur* nichts als konstant angenommen werden kann. Im Rahmen der *System Evaluation* in Abschnitt 5.3 wurde gezeigt, dass je nachdem wie und wie stark sich verschiedene Bibelversionen unterscheiden, das minimalste Programm $\mathcal{P}_{max}$ deutlich variiert und somit auch große Einflüsse auf die optimalen Einstellungen der einzelnen Level der *7-Level-Architektur* haben. So konnte ein maximales *F-Measure* zwischen den beiden Bibelversionen *King James Version* und *Webster's Revision* bei einem *Featuring* durch ein *Bigram Shingling* bestimmt werden, während sonst das *Word based Featuring* für maximale *F-Measure* und eine *Mining Ability* gesorgt hat. Auf der anderen Seite ist es ein Ergebnis, dass der *Scoring*-Schwellwert t von der Art und Weise des *Preprocessing* abhängig ist. Weiterhin zeigen die Ergebnisse dieser Arbeit, dass je näher sich zwei Bibelversionen sind, desto höher der *Scoring*-Schwellwert zu wählen ist.

Genau diese *Diversity* und das damit verbundene Erkennen der besten Algorithmen eines jeden Level der *7-Level-Architektur* des *Historical Text Re-use* für eine *Digital Library* setzt eine ganzheitliche Evaluierung des Ergebnisses voraus. Dies gilt insbesondere auch dann, wenn es keinen *Gold Standard* bzw. keine geeignete Evaluierungsbasis für die jeweilige Sprache bzw. Domäne gibt. Aus diesem Grund ist die evaluierungsbasisunabhängige Methodik der *Text Re-use Compression* sowie der *Noisy Channel Evaluation* von besonderer Bedeutung. In der Zusammenfassung zum Kapitel 5 wurde aber auch deutlich gemacht, dass die *Precision* derzeit durch keine automatische Methode approximiert werden kann. Deshalb wurde auf Basis der Erfahrungen einer *Text Re-use Analysis* auf der *Perseus Digital Library* der Vorschlag unterbreitet, die *Precision* durch eine manuelle Evaluierung der Ergebnisse zu bestimmen. Das hat zwar den offensichtlichen Nachteil, dass nur ein sehr geringer Teil des Ergebnisses auf dessen Genauigkeit untersucht werden kann, dennoch wird durch eine geeignete Stichprobe die realistische *Precision* besser gemessen, die immer größer sein wird als die *Precision* aus einer Evaluierung gegen einen *Gold Standard*, da nicht in den Evaluierungsdaten enthaltene relevante Daten als positiv gewertet werden können, so dass sich der wahren bzw. realistischen Genauigkeit dennoch deutlich besser genähert werden kann.

Was bedeutet diese Arbeit für die *Scientific Community* bzw. welcher Mehrwert wird durch sie generiert? Ausgehend von dem *4-Sichten-Modell* des *Historical Text Re-use* für die *Humanities*, *Digital Humanities*, *eHumanities* und *Computer Science* gibt es dementsprechend nicht eine sondern mindestens vier Antworten.

Für die *Humanities* bedeutet eine automatische *Text Re-use Analysis* in erster Linie ein Umstellen der Methodik als auch eine Verschiebung von Aufgabenbereichen. Während, wie in Abb. 1.5 auf Seite 41, das manuelle Sammeln von *Text Re-use* Teil der Tradition geisteswissenschaftlicher Arbeit ist, bedeuten automatisierte Techniken, dass wenn auch nicht gänzlich vollautomatische, so aber dennoch mit einer großen Quantität, Erstellen entsprechender Parallelstellen. Das führt zweifelsohne nicht zu einer Ersetzung der Fachwissenschaften, sondern ermöglicht einen schnelleren Zugriff auf Informationen, ohne dass erst teilweise jahrelang Daten zusammentragen werden müssen[3]. Dadurch ergeben sich für die Geisteswissenschaften zwei direkte Vorteile. Erstens kann in einer fachwissenschaftlichen Anwendung, wie der *Textkritik*, schneller und auf mehr Parallelstellen zurückgegriffen werden, so dass der Forschungsschwerpunkt stärker auf die Textkritik und weniger auf dem Sammeln der Daten liegt. Zweitens, es wird durch die quantitative Methode der *Serendipity Effect* in den Fachwissenschaften unterstützt (vgl. [Büchler 2013b]). Kein Algorithmus muss einem Universitätsprofessor in den Geisteswissenschaften das erklären, was bereits als Grundwissen im Bachelorstudium gelehrt wird. Da der Mensch jedoch nach der *Theorie der selektiven Wahrnehmung* niemals alle Informationen sowohl gleichzeitig als auch in ihrer Gesamtheit erfassen und verarbeiten kann, tendiert jede Person dazu, nach eigenen Selektionsstrategien dieses Wissen zu reduzieren, was bspw. in Anlehnung an eine *facettierte Suche* auf Basis von zu erwartenden Metainformationen gemäß des Grundwissens geschehen kann. Hierbei werden etwaige und vielleicht auch besonders wichtige Informationen durch eine menschliche, facettierte Selektion aus dem Ergebnis auf Basis einer falschen Erwartung entfernt. Insbesondere ein *Text Re-use Graph* kann dazu eingesetzt werden, die nicht zu erwartenden Beziehungen zu extrahieren. Die wissenschaftlichen Grundlagen hierzu hat bereits Granovetter mit seiner Theorie der *weak ties* gelegt (vgl. [Granovetter 1983]), welche durch die *Contrastive Semantics* (vgl. [Büchler 2010f]) eine technische Umsetzung erfahren haben. Im Detail bedeutet dies, dass der *Serendipity Effect* nicht nur durch ein einfaches graphbasiertes *Browsing*, sondern insbesondere durch eine Analyse auf *weak ties* unterstützt werden kann. Das heißt, dass hierfür nicht nur etwaige offensichtliche und bekannte Zitierabhängigkeiten, wie die zwischen Platon und Galen, angezeigt werden, sondern in besonderer Weise auf die *weak ties* hingewiesen werden kann, so dass die Chance vergrößert wird, neue bzw. bisher nicht aufgedeckte Zusammenhänge zu entdecken.

Für die *Digital Humanities* stellt diese Arbeit die Grundlage für ein *Typsystem* sowohl der Knoten als auch der Kanten eines *Text Re-use Graph* dar (vgl. Abschnitt 2.6). Während der Bearbeitung des Themas ist bereits sehr früh deutlich geworden, dass jeder *Text Chunk*, welcher wiederverwendet wird, letztlich bereits einen Namen hat. So kann ein *Text Re-use* durch ein *Meme Definition*, *Proverb*, *Abstract* aber auch *Idiom* beschrieben werden. Weiterhin gibt es auch aus Sicht der *Edges* verschiedene *Re-use Styles*, wie *Verbatim* oder *Paraphrase*, welche teilweise stark unterschiedliche Techniken während einer *Text Re-use Analysis* benötigen. Somit stellen die beiden *Typsystems* für die *Nodes* und *Edges* nicht nur eine Grundlage für das Speichern und Konservieren eines *Text Re-use Graph* dar, sondern sind, wie bereits ausgeführt, Indikatoren für die *Diversity*, welche die *7-Level-Architektur* nach sich gezogen hat.

Aus Sicht der *Computer Science* kann gezeigt werden, dass das quadratische Verhalten einer *Text Re-use Analysis* insbesondere bei größeren *Digital Libraries* sukzessive zu einem

[3] Das gilt natürlich nur solange, wie die Daten auch digital vorliegen. Die gemachte Aussage kann ganz klar nicht für gedruckte Ausgaben gelten.

Problem wird. Entsprechende *Linking*-Analysen in Abschnitt 5.4.5 ab Seite 209 zeigen, dass selbst durch eine *Selection* nur auf Basis von Substantiven, Verben und Adjektiven als *Features*, welche im Rahmen einer manuellen Analyse durch 24 Testpersonen als besonders in einer *Digital Signature* erhaltenswert angesehen worden sind (vgl. Abschnitt 5.4.4), das Laufzeitverhalten noch schlechter als ein *Brute Force Linking* ist, welches jede *Re-use Unit* mit jeder anderen vergleicht. Ziel und Forschungsgegenstand zukünftiger Arbeiten muss somit für die *Computer Science* auf der Erforschung der *Minutiae* sowie der bestmöglichen Bestimmung des *Re-use Nucleus* liegen.

Für die *eHumanities* und in starker Anlehnung an die Methodik der *Computer Science* stellt die *Diversity* eine Herausforderung sowohl für die *Text Re-use Analysis* als auch der Evaluation von entsprechenden Ergebnissen dar. Einerseits hat die *Diversity* im Rahmen dieser Arbeit gezeigt, dass sehr spezifische Analysen mit unterschiedlichen Algorithmen und Parametern der *7-Level-Architektur* nötig sind. Andererseits resultiert daraus automatisch auch, dass der *Text Re-use* in einer *Digital Library* nicht nur durch einen *Algorithmus* berechnet werden kann, sondern je nach enthaltenen *Meme* und *Re-use Styles* auf mehrere Analysen aufgesplittet und anschließend die Teilergebnisse zu einem im Sinne des *Hybrid Text Re-use* zusammengesetzten *Text Re-use Graph* zusammengeführt werden müssen.

6.2 Lessons Learnt

Text Mining auf historischen Dokumenten gestaltet sich deutlich schwieriger als auf moderneren Texten. Neben den bereits vielfach diskutierten historischen Schreibvarianten und Dialekten betrifft ein erstes *Lessons Learnt* den Umgang mit Komposita.

Insbesondere bei deutschen Texten können mit den bisherigen *Preprocessing*-Techniken nicht alle Varianten normalisiert werden. Nur ein Beispiel hierfür sind die biblischen Anspielungen *Schwefelregen* und *schwefliger Regen*. Es ist zwar einerseits möglich, durch das *Noisy Channel Mining* diese Varianten zu extrahieren. Andererseits scheint eine Kompositazerlegung sowie eine Reduktion der Wörter auf einzelne Morpheme naheliegender.

Neben solchen eher einfachen Aspekten des *Preprocessing* kommt das grundlegendste *Lessons Learnt* direkt aus den Ergebnissen der *System Evaluation* in Abschnitt 5.3 ab Seite 176. Hier konnte gezeigt werden, dass sich die Vielfalt, mit der Text wiederverwendet wird, deutlich auf die Parameter auswirkt. So konnten bei Bibelversionen, wie der *King James Version* und der *Webster's Revision*, zwischen welchen ein naher Zusammenhang besteht, andere Algorithmen mit unterschiedlichen Parametern als optimal bestimmt werden, als beim Vergleich der *King James Version* mit der *Bible in Basic English*. Als Konsequenz dessen, stellt sich die Frage, ob die bisherige Methode, einen Algorithmus auf eine *Digital Library* anzuwenden, die Zukunft sein kann. Letztlich ist auch ein *Hybrid Text Re-use* nur eine Teillösung des Problems. Im Rahmen der Evaluierungen wurde gezeigt, dass zukünftig Methoden der Korpuslinguistik und der Informatik näher zusammengebracht werden müssen.

Dies bedeutet im Detail, dass die Methode des werkweisen Vergleiches aus der Korpuslinguistik (vgl. Tabelle 6.1) zukünftig auch für die *Text Re-use Analysis* adaptiert werden muss. Dies resultiert sowohl aus der *Diversity* durch unterschiedliche *Meme* als auch verschiedene *Re-use Styles*, die jeweils als autorspezifisch anzusehen sind. So zitiert ein Autor, ähnlich dem Vergleich zwischen der *King James Version* und *Webster's Revision*, näher das Original, während ein anderer Autor dazu tendiert, stärker zu paraphrasieren. Durch einen werkweisen Vergleich kann insbesondere auch durch die *Text Re-use Compression* und die *Noisy Channel Evaluation* vollautomatisch das im Sinne der *7-Level-Architektur* beste Sprachmodell gefunden werden. Vielmehr verrät das Sprachmodell, bei welchem die Evalu-

ierung maximal wird, etwas über die Gewohnheiten eines zitierenden Autors, wodurch das werkweise Analysieren auf *Text Re-use* entgegen dem bisherigen Anwenden und Evaluieren eines Algorithmus in der Informatik einem Paradigmenwechsel gleichkommt.

	Pro	**Kontra**
Corpuslinguistik	meist autor- bzw. werkspezifische *Text Re-use Analysis*	*Big Scope* bzw. *Distant View* fehlt oftmals
Informatik	globale Zitiergewohnheiten und Abhängigkeiten können analysiert werden	schlechte Qualität, da nur ein Algorithmus als Kompromiss angewandt wird

Tabelle 6.1: Vergleich durch Pro und Kontra für eine *Text Re-use Analysis* in der Korpuslinguistik sowie der Informatik.

Mit diesem Paradigmenwechsel geht im Sinne eines *Distributed Computing* eine starke Vereinfachung der Parallelisierbarkeit einher. Während aus der Sicht einer ganzheitlichen *Text Re-use Analysis* einer *Digital Library* eine Parallelisierung aufgrund intensiver Rechenzeiten lediglich auf dem *Linking*- bzw. damit verbunden auch auf dem *Scoring*-Level sinnvoll wäre, kann durch das paarweise Vergleichen eine nahezu beliebig verteilte oder lokale Parallelisierbarkeit erreicht werden. Das sei am Beispiel der *Perseus Digital Library* verdeutlicht. Die *Perseus Digital Library* umfasst derzeit etwa 500 griechische Werke. Etwa 200 weitere antike Werke stehen durch *OCR-Processing* zur Verfügung. Aus diesem Grund sei von insgesamt 700 Werken ausgegangen. Jedes Werk mit jedem zu vergleichen würde bedeutet, insgesamt knapp 49.000 werkweise Vergleiche durchführen zu müssen. Da diese Vergleiche symmetrisch sind, und nicht Werk A mit B sowie Werk B mit A verglichen werden muss, reichen bereits knapp 44.500 werkweise Vergleiche aus. Ausgehend von einem verteilenden Knoten könnten somit insgesamt bis zu 44.500 *Computational Nodes* in einer *Distributed Environment* mit jeweils einem Vergleich zeitgleich ausgelastet werden, wobei jeweils unterschiedliche Einstellungen berechnet werden, um sich letztlich für das Modell zu entscheiden, welches die besten Evaluierungsergebnisse liefert. Sollten nur 100 *Computational Nodes* zur Verfügung stehen, würde dementsprechend jeder dieser Knoten im Durchschnitt rund 445 paarweise Vergleiche berechnen müssen. Anschließend werden die Einzelergebnisse im Sinne einer *Divide & Conquer* Strategie zu einem *Text Re-use Graph* zusammengesetzt. Mit einer ganzheitlichen Analyse einer *Digital Library* wäre eine derartige Parallelisierbarkeit nicht einmal ansatzweise möglich, so dass der Paradigmenwechsel auch im Kontext der Infrastruktur positiv zu werten ist.

Ein weiteres *Lessons Learnt* umfasst die *Re-use Technik* selbst. Die im Rahmen dieser Arbeit vorgestellte *Text Re-use Analysis* setzt implizit immer voraus, dass sowohl *Source*- als auch *Target*-Werke in einer *Digital Library* enthalten sind. Das ist jedoch eine stark einschränkende Bedingung, die in vielen Fällen nicht erfüllt werden kann. In Abb. 2.3 auf Seite 77 ist der *Incomplete Text Re-use* als ein spezieller *Edge Type* eingeführt worden. Ohne dass die Quelle in der *Digital Library* enthalten ist, kann jedoch ungleich schwerer entschieden werden, ob ein *Text Chunk* von dem Autor eines Werkes stammt. Dennoch können die Methoden der Autorenerkennung, wie in [Tschuggnall 2012], verwendet werden, um diejenigen Sätze zu bestimmen, welche nicht mit dem typischen Satzbau eines Autors übereinstimmen. Hierfür eignen sich insbesondere syntaktische *Grammar Trees* (vgl. [Tschuggnall 2012]). Auch wenn deren Einsatz auf großen *Digital Libraries* derzeit noch nicht erwiesen ist, so zeigen die *Fragmentary Authors* (vgl. [Berti 2009, Berti 2012]) die Notwendigkeit auf. Auf Basis der im Rahmen dieser Arbeit erzielten Ergebnisse kann nur geschätzt werden, dass etwa 60% des *Text Re-use* unvollständig ist.

6.3 Weiterführende Aspekte und zukünftige Arbeiten

Neben den *Lessons Learnt* haben sich auch weiterführende Aspekte zukünftiger Arbeiten herauskristallisiert. Im Rahmen dieser Arbeit wurden die Terminologien *Minutiae* und *Re-use Nucleus* für den *Text Re-use* eingeführt. Die dazugehörigen Ergebnisse aus Abschnitt 5.4.4 ab Seite 205 zur Evaluierung einer *Digital Signature* können nur als ein allererster Schritt einer völlig neuartigen Denkweise angesehen werden. Mit den Arbeiten zu den *Minutiae* als auch dem *Re-use Nucleus* werden zwei Dinge grundlegend verändert. Erstens, die Forschung wird sich zukünftig der Frage stellen müssen, was und vor allem warum wird etwas wiederverwendet. Der derzeitige Forschungsstand wendet oftmals lediglich nur eine Technik auf Text an. Warum eine bestimmte Technik im Sinne der *Minutiae* als relevant angesehen werden kann, bleibt zuweilen offen. Vielmehr sensibilisiert das Denken in *Minutiae* in Anlehnung an die Biometrie, wie und was Teil des *Re-use Nucleus* sein kann und muss. Vielmehr stellt sich die Forschungsfrage, ob es überhaupt für den *Text Re-use* eine relevante Form von *Minutiae* gibt und wie diese aussehen. Selbst in der Biometrie gibt es je nach Nachschlagewerk zwischen sieben und zwölf *Minutiae*. Jedoch setzen praxisnahe Institutionen, wie das *FBI*, lediglich zwei Arten von *Minutiae* ein. Neben technischen Aspekten des *Feature Extraction* stehen hierbei insbesondere die Qualitätskriterien, wie *Collectability* und *Circumvention* (vgl. Abschnitt 2.4), im Vordergrund. Selbst wenn auf linguistischer Ebene ein theoretisches Konstrukt für die *Minutiae* und damit auch der Formung eines *Re-use Nucleus* existiert, muss im Sinne der *Collectability* das bei Weitem nicht gut erfassbar sein. Zweitens, mit dem Testen und Evaluieren der *Minutiae* und damit verbunden auch mit der *Digital Signature* geht die Denkweise einer Evaluierung von einer *System Evaluation* hin zu einer *Component Evaluation*, wodurch einzelne Komponenten, wie die *Digital Signature*, separat auf deren Qualität untersucht werden.

Abschließend zu den *Minutiae* scheint es eine Parallele zum Spracherwerb bei Kleinkindern zu geben, wie die Ergebnisse in Abschnitt 5.4.4 ab Seite 205 zeigen. Letztlich impliziert die Frage nach den *Minutiae*, welche Wörter und Wortgruppen als Kern des *Text Re-use* zwingend in die *Digital Signature* übernommen werden müssen. Vielmehr kann man auch sagen, dass nur diejenigen Wörter entfernt werden, welche eine Redundanz bzw. unnötige Information darstellen. Die Gemeinsamkeit mit dem Spracherwerb von Kleinkindern kommt damit einher, dass sie anfangs auch nicht in ganzen Sätzen sprechen, so dass Phrasen, wie *spielen gehen*, bereits ausreichen, um ein Interesse deutlich zu machen. Das Ergebnis einer *Minutiae*-Analyse mit 24 Probanden auf deutschen Redewendungen hat gezeigt, dass sich die Phrasen von Kindern sowie die reduzierten aber dennoch zusammenhängenden Sprachkonstrukte der Redewendungen deutlich überlappen. Auch wenn es nicht Gegenstand dieser Arbeit war, scheint es sinnvoll, diese Beziehung näher zu beleuchten und etwaige Schlüsse aus dem Spracherwerb zu adaptieren.

Wenngleich es in erster Linie Gegenstand der *7-Level-Architektur* gewesen ist, aus Forschungssicht auf die *Diversity* historischer Daten bzw. verschiedener *Re-use Styles* eine passende technische Antwort zu formulieren, so kann diese Architektur dabei helfen, die Akzeptanz zu erhöhen. Dies geschieht dadurch, dass jedes Ergebnis eines Level auf der Festplatte gespeichert wird. Neben dem Fakt, dass auf diese Weise einfacher auf bereits berechnete Daten zurückgegriffen werden kann und somit nicht mehr alles neu berechnet werden muss, können die gespeicherten Zwischenergebnisse im Sinne eines *Debuggers* eingesetzt werden. Grundidee ist hierbei nicht nur das Ergebnis, also die Verlinkung zweier *Re-use Units*, dem Nutzer aufzuzeigen, sondern ähnlich dem in der Informatik hinlänglich bekannten *Debug*-Modus, dem Nutzer die Zwischenschritte und damit auch Ursachen für etwaige Fehler zu zeigen. In der Konsequenz kann auf diese Weise die *Black Box* transparent gemacht werden, wodurch die Identifikation mit den Methoden sowie damit verbunden

die *Acceptance* erhöht werden kann. Für den *Historical Text Re-use* ist auf Basis der *7-Level-Architektur* und den Daten der *Text Re-use Analysis* aus Abschnitt 5.3 bereits ein Demonstrator entwickelt worden[4].

Auf der anderen Seite kann die *Acceptance* auch dadurch erreicht werden, indem eine *Text Re-use Analysis* in einem praxisnahen Umfeld eingesetzt bzw. evaluiert wird. Im Rahmen des *eTRACES*-Projektes wird hierzu gerade der Einsatz sowie der Nutzen von *Text Re-use* beim Erstellen von *Online Editionen* erforscht (vgl. [Geßner 2013]). Der Arbeitsschwerpunkt liegt auf dem Kommentieren der Texte. Das bedeutet, dass in entsprechenden Fußnoten Hinweise auf Querbezüge zu anderen Texten erstellt werden sollen. Bisher ist diese Aufgabe weitestgehend manuell geschehen, so dass durch ein Beschleunigen im Finden von Parallelstellen berechtigte Motivation besteht, dass automatische Methoden in solchen Umgebungen akzeptiert werden. Weiterhin ist es Gegenstand dieser Arbeiten, zu prüfen, inwiefern der *Serendipity Effect* angenommen wird. Es besteht allerdings auch zeitgleich die Gefahr, dass durch eine zu starke *Acceptance*, Ergebnisse einer *Text Re-use Analysis* ungeprüft bzw. schlecht verifiziert übernommen werden, so dass auf Quantität anstelle von Qualität gesetzt wird. Ein grundlegendes Problem für einen derartigen Einsatz bleibt jedoch die Mindestgröße der *Digital Library* bzw. deren Repräsentativität bzgl. des zu bearbeitenden Werkes. Der erwünschte Effekt der *Acceptance* kann auch in das Gegenteil umschlagen, wenn nicht sichergestellt wird, dass durch eine *Digital Library* hinreichend viele Treffer generiert werden, so dass ein Mehrwert ausbleibt.

Während die bisherigen Aspekte entweder auf die *Acceptance* oder die Grundlagenforschung abzielen, wird abschließend zu dieser Arbeit auf einige relevante und weiterführende Anwendungen von *Text Re-use* in den *eHumanities* eingegangen. Grundlegend gilt für alle nachfolgenden Themen, dass *Text Re-use* nicht mehr als wissenschaftliche Methode sondern als Hilfswissenschaft in einem größeren Kontext betrachtet wird.

Mit jedem *Text Re-use* geht implizit eine Gewichtung einher. Niemand zitiert jemand anderen zufällig. Dementsprechend kann ein *Text Re-use Graph* einer *Digital Library* als eine komplexe Datenstruktur zur Gewichtung verstanden werden. Mit der durch einen *Text Re-use Graph* gegebenen *Hypertextualität* kann Google's *PageRanking*-Technik auf einen solchen Graphen angewandt werden (vgl. [Büchler 2012b]). Somit deckt der *Historical Text Re-use* nicht nur Abhängigkeitsverhältnisse auf, sondern liefert auch Gewichtungen auf Basis von bereits seit Jahrhunderten verstorbenen antiken Autoren. Diese konkrete Anwendung resultiert aus dieser Arbeit und wird derzeit im Rahmen von *eTRACES* in Zusammenarbeit mit der *Perseus Digital Library* umgesetzt. Entgegen dem *PageRanking*, welcher quasi das vertrauensvollste Dokument durch möglichst viele hoch gewichtete eingehende Links bestimmt, ist der Fokus dieser Zusammenarbeit in zwei Dimensionen aufgeteilt. Einerseits soll nicht nur immer das gewichtigste Dokument angezeigt werden, sondern auch im Sinne des *Serendipity Effects* gegenteilige Dokumente, welche selten zitiert worden sind. Die Arbeitshypothese hierbei ist, dass Studenten und junge Forscher tendenziell eher den Dokumenten mit einem hohen *Text Re-use* und damit mit einer hohen Vertrauenswürdigkeit folgen, während reifere Forscher diese Textstellen bereits hinreichend kennen und zu den niederfrequent zitierten Texten bzw. Textstellen tendieren. Die zweite Dimension umfasst, ob eine hohe *Text Re-use Temperature* oder eine hohe *Text Re-use Coverage* (beide vgl. Abb. 1.3) für eine derartige Ranking-Strategie vom Nutzer bevorzugt wird.

In Kapitel 4 wurde aufgezeigt, dass bedingt durch das *Zipfsche Gesetz* bereits rund 95% aller *Co-occurrences* und *Bigrams*, welche die bzgl. des *Re-use Overlaps* kleinsten *Text Re-use Candidates* darstellen, fünfmal und seltener in einer *Digital Library* beobachtet werden. Eine *Text Re-use Analysis* kann nun vielmehr helfen, eine *Digital Library* durch ein *Text*

[4] vgl. `http://roedel.e-humanities.net:8080/webdebugger/webdebugger/word_input_form`

Decontamination so aufzubereiten, dass *Text Re-use* für das maschinelle Lernen entfernt wird. Insbesondere für die sehr vielen seltenen Ergebnisse verbessert sich der maschinelle Lernerfolg so deutlich, da nicht aus den Redundanzen, welche durch den *Text Re-use* erzeugt werden, etwas Falsches gelernt bzw. das Ergebnis nicht verfälscht wird.

Abschließend sei diese Arbeit als ein Beitrag zum *Re-use* im Allgemeinen verstanden. In Kapitel 1 wurden einige einführende Beispiele, wie in Abschnitt 1.2 ab Seite 28, gegeben. Weiterhin wurden zahlreiche Verbindungen zur Biometrie hergestellt. Grundlegend haben diese Forschungsbereiche alle gemeinsam, dass sie nicht nur *Re-use* aufdecken wollen, sondern in erster Linie ihn erst einmal versuchen, zu verstehen. Auch wenn die Frage nach dem generellen und globalen *Re-use Pattern*, welches die Forschungen der Biometrie, Biologie, DNA-Analyse oder auch des *Text Re-use* umfasst, derzeit nicht beantwortbar ist, so ist es dennoch angenehm über die generellen Strukturen des Universums nachzudenken. Vielmehr ist die Erforschung des globalen *Re-use Pattern* auch Antrieb für weitere Forschungen.

A new scientific truth does not triumph by convincing its opponents

and making them see the light,

but rather because its opponents eventually die,

and a new generation grows up that is familiar with it.

Max Planck, (1858-1947)

Anhang A

χ^2-Tabelle

Ich traue keiner Statistik, die ich nicht selbst gefälscht habe.

Winston Churchill, (1874-1965)[1]

Dieser Anhang enthält χ^2-Tabellen für verschiedene α-Fehlerniveaus sowie für bis zu 250 Freiheitsgrade.

[1] Diese Aussage stammt möglicherweise auch vom deutschen Reichspropagandaminister Joseph Goebbels, der Churchill diesen Spruch andichten wollte. [Wikipedia 2011]

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
1	2.706	2.875	3.065	3.284	3.538	3.842	4.218	4.710	5.412	6.635	7.880	10.828	15.137
2	4.606	4.816	5.052	5.319	5.627	5.992	6.438	7.014	7.825	9.211	10.597	13.816	18.421
3	6.252	6.492	6.759	7.061	7.407	7.815	8.312	8.948	9.838	11.345	12.839	16.267	21.108
4	7.780	8.044	8.337	8.667	9.045	9.488	10.026	10.712	11.668	13.277	14.861	18.467	23.513
5	9.237	9.522	9.837	10.192	10.597	11.071	11.645	12.375	13.389	15.087	16.750	20.516	25.745
6	10.645	10.948	11.284	11.660	12.090	12.592	13.198	13.968	15.034	16.812	18.548	22.458	27.857
7	12.018	12.338	12.692	13.088	13.540	14.068	14.704	15.510	16.623	18.476	20.278	24.322	29.878
8	13.362	13.698	14.069	14.484	14.957	15.508	16.171	17.011	18.169	20.091	21.955	26.125	31.828
9	14.684	15.035	15.422	15.854	16.346	16.919	17.609	18.480	19.680	21.666	23.590	27.878	33.720
10	15.988	16.352	16.754	17.203	17.714	18.308	19.021	19.922	21.161	23.210	25.189	29.589	35.565
11	17.276	17.653	18.069	18.534	19.062	19.676	20.413	21.342	22.618	24.725	26.757	31.265	37.367
12	18.550	18.940	19.370	19.849	20.394	21.027	21.786	22.742	24.054	26.217	28.300	32.910	39.135
13	19.812	20.215	20.657	21.151	21.712	22.363	23.143	24.125	25.472	27.689	29.820	34.529	40.871
14	21.065	21.478	21.934	22.441	23.017	23.685	24.486	25.494	26.873	29.142	31.320	36.124	42.580
15	22.308	22.732	23.200	23.721	24.311	24.996	25.817	26.848	28.260	30.578	32.802	37.698	44.264
16	23.542	23.978	24.457	24.991	25.595	26.297	27.136	28.191	29.634	32.000	34.268	39.253	45.925
17	24.770	25.215	25.706	26.252	26.871	27.588	28.445	29.523	30.996	33.409	35.719	40.791	47.567
18	25.990	26.446	26.947	27.505	28.138	28.870	29.746	30.845	32.347	34.806	37.157	42.313	49.190
19	27.204	27.670	28.182	28.752	29.397	30.144	31.037	32.158	33.688	36.191	38.583	43.821	50.796
20	28.412	28.888	29.410	29.991	30.649	31.411	32.321	33.463	35.020	37.567	39.997	45.315	52.386
21	29.616	30.100	30.633	31.225	31.895	32.671	33.598	34.760	36.344	38.933	41.402	46.798	53.963
22	30.814	31.308	31.850	32.453	33.135	33.925	34.868	36.050	37.660	40.290	42.796	48.268	55.525
23	32.007	32.510	33.062	33.676	34.370	35.173	36.132	37.333	38.969	41.639	44.182	49.729	57.075
24	33.197	33.708	34.270	34.894	35.599	36.416	37.390	38.610	40.271	42.980	45.559	51.179	58.613
25	34.382	34.902	35.473	36.107	36.824	37.653	38.642	39.881	41.567	44.315	46.928	52.620	60.141

Tabelle A.1: χ^2_α-Signifikanzwerte für 1 bis 25 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
26	35.564	36.092	36.672	37.316	38.044	38.886	39.890	41.147	42.856	45.642	48.290	54.052	61.658
27	36.742	37.278	37.867	38.521	39.260	40.114	41.132	42.407	44.140	46.963	49.645	55.477	63.165
28	37.916	38.461	39.058	39.722	40.471	41.338	42.370	43.663	45.419	48.279	50.994	56.893	64.663
29	39.088	39.640	40.246	40.919	41.679	42.557	43.604	44.914	46.693	49.588	52.336	58.302	66.152
30	40.257	40.817	41.431	42.113	42.884	43.773	44.834	46.160	47.962	50.893	53.672	59.704	67.633
31	41.422	41.990	42.612	43.304	44.084	44.986	46.060	47.403	49.227	52.192	55.003	61.099	69.106
32	42.585	43.160	43.791	44.491	45.282	46.195	47.282	48.642	50.487	53.486	56.329	62.488	70.572
33	43.746	44.328	44.967	45.676	46.476	47.400	48.501	49.876	51.743	54.776	57.649	63.871	72.030
34	44.904	45.494	46.140	46.858	47.668	48.603	49.716	51.108	52.996	56.061	58.964	65.248	73.482
35	46.059	46.656	47.311	48.037	48.857	49.802	50.929	52.336	54.244	57.343	60.275	66.619	74.927
36	47.213	47.817	48.479	49.213	50.042	50.999	52.138	53.560	55.489	58.620	61.582	67.986	76.365
37	48.364	48.975	49.644	50.387	51.226	52.193	53.344	54.782	56.731	59.893	62.884	69.347	77.798
38	49.513	50.131	50.808	51.559	52.407	53.384	54.547	56.000	57.969	61.163	64.182	70.703	79.225
39	50.660	51.285	51.969	52.729	53.585	54.573	55.748	57.216	59.204	62.429	65.476	72.055	80.647
40	51.806	52.437	53.128	53.896	54.761	55.759	56.946	58.428	60.437	63.691	66.766	73.402	82.063
41	52.949	53.587	54.286	55.061	55.935	56.943	58.142	59.638	61.666	64.951	68.053	74.745	83.474
42	54.091	54.735	55.441	56.224	57.107	58.125	59.335	60.846	62.892	66.207	69.336	76.084	84.880
43	55.231	55.882	56.594	57.385	58.276	59.304	60.526	62.051	64.116	67.460	70.616	77.419	86.281
44	56.369	57.026	57.746	58.544	59.444	60.481	61.715	63.254	65.337	68.710	71.893	78.750	87.678
45	57.506	58.169	58.896	59.702	60.610	61.657	62.902	64.454	66.556	69.957	73.167	80.077	89.070
46	58.641	59.311	60.044	60.857	61.774	62.830	64.086	65.652	67.772	71.202	74.437	81.401	90.458
47	59.775	60.451	61.191	62.011	62.936	64.002	65.268	66.848	68.986	72.444	75.705	82.721	91.842
48	60.907	61.589	62.336	63.164	64.096	65.171	66.449	68.042	70.197	73.683	76.969	84.038	93.221
49	62.038	62.726	63.479	64.314	65.255	66.339	67.628	69.234	71.407	74.920	78.231	85.351	94.597
50	63.168	63.862	64.621	65.463	66.412	67.505	68.804	70.423	72.614	76.154	79.490	86.661	95.969

Tabelle A.2: χ^2_α-Signifikanzwerte für 26 bis 50 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
51	64.296	64.996	65.762	66.611	67.568	68.670	69.979	71.611	73.819	77.386	80.747	87.968	97.338
52	65.423	66.129	66.901	67.757	68.722	69.833	71.153	72.798	75.022	78.616	82.001	89.273	98.702
53	66.549	67.260	68.039	68.902	69.874	70.994	72.324	73.982	76.223	79.844	83.253	90.574	100.064
54	67.673	68.391	69.176	70.045	71.025	72.154	73.494	75.164	77.422	81.069	84.502	91.872	101.422
55	68.797	69.520	70.311	71.188	72.175	73.312	74.663	76.345	78.620	82.293	85.749	93.168	102.776
56	69.919	70.648	71.445	72.328	73.323	74.469	75.830	77.524	79.815	83.514	86.994	94.461	104.128
57	71.040	71.774	72.578	73.468	74.470	75.624	76.995	78.702	81.009	84.733	88.237	95.751	105.476
58	72.160	72.900	73.709	74.606	75.616	76.778	78.159	79.878	82.201	85.951	89.477	97.039	106.821
59	73.279	74.025	74.840	75.743	76.760	77.931	79.321	81.053	83.392	87.166	90.716	98.325	108.164
60	74.398	75.148	75.969	76.879	77.903	79.082	80.482	82.226	84.580	88.380	91.952	99.608	109.503
61	75.515	76.271	77.098	78.014	79.045	80.233	81.642	83.397	85.768	89.592	93.187	100.888	110.840
62	76.631	77.392	78.225	79.147	80.186	81.382	82.801	84.567	86.953	90.802	94.419	102.167	112.174
63	77.746	78.513	79.351	80.280	81.326	82.529	83.958	85.736	88.138	92.011	95.650	103.443	113.505
64	78.860	79.632	80.477	81.412	82.464	83.676	85.114	86.904	89.320	93.217	96.879	104.717	114.834
65	79.974	80.751	81.601	82.542	83.602	84.821	86.268	88.070	90.502	94.423	98.106	105.989	116.160
66	81.086	81.868	82.724	83.672	84.738	85.965	87.422	89.235	91.682	95.626	99.331	107.258	117.484
67	82.198	82.985	83.846	84.800	85.874	87.109	88.574	90.398	92.860	96.828	100.555	108.526	118.805
68	83.308	84.101	84.968	85.928	87.008	88.251	89.726	91.561	94.037	98.029	101.776	109.792	120.124
69	84.418	85.216	86.088	87.054	88.141	89.392	90.876	92.722	95.213	99.228	102.997	111.056	121.441
70	85.528	86.330	87.208	88.180	89.274	90.532	92.025	93.882	96.388	100.426	104.215	112.317	122.755
71	86.636	87.444	88.327	89.305	90.405	91.671	93.173	95.041	97.561	101.622	105.433	113.577	124.067
72	87.744	88.556	89.445	90.429	91.536	92.809	94.319	96.198	98.734	102.817	106.648	114.836	125.377
73	88.850	89.668	90.562	91.552	92.665	93.946	95.465	97.355	99.905	104.010	107.862	116.092	126.685
74	89.957	90.779	91.678	92.674	93.794	95.082	96.610	98.511	101.074	105.203	109.075	117.347	127.991
75	91.062	91.889	92.794	93.795	94.922	96.217	97.754	99.665	102.243	106.393	110.286	118.600	129.294

Tabelle A.3: χ^2_α-Signifikanzwerte für 51 bis 75 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
76	92.167	92.999	93.909	94.916	96.049	97.351	98.897	100.819	103.411	107.583	111.496	119.851	130.596
77	93.271	94.108	95.023	96.035	97.175	98.485	100.039	101.971	104.577	108.771	112.704	121.100	131.895
78	94.374	95.216	96.136	97.154	98.300	99.617	101.180	103.123	105.742	109.959	113.911	122.348	133.193
79	95.477	96.323	97.248	98.273	99.425	100.749	102.320	104.273	106.907	111.145	115.117	123.595	134.489
80	96.579	97.430	98.360	99.390	100.548	101.880	103.459	105.423	108.070	112.329	116.322	124.840	135.783
81	97.680	98.536	99.471	100.507	101.671	103.010	104.598	106.571	109.232	113.513	117.525	126.083	137.075
82	98.781	99.641	100.582	101.623	102.793	104.139	105.735	107.719	110.393	114.695	118.727	127.325	138.366
83	99.881	100.746	101.692	102.738	103.915	105.268	106.872	108.866	111.553	115.877	119.927	128.565	139.654
84	100.980	101.850	102.801	103.852	105.035	106.395	108.008	110.012	112.713	117.057	121.127	129.804	140.941
85	102.079	102.954	103.909	104.966	106.155	107.522	109.143	111.157	113.871	118.236	122.325	131.042	142.226
86	103.178	104.057	105.017	106.080	107.275	108.648	110.277	112.301	115.028	119.414	123.522	132.278	143.510
87	104.276	105.159	106.124	107.192	108.393	109.774	111.410	113.444	116.185	120.592	124.718	133.513	144.792
88	105.373	106.261	107.231	108.304	109.511	110.899	112.543	114.587	117.340	121.768	125.913	134.746	146.072
89	106.469	107.362	108.337	109.415	110.628	112.022	113.675	115.728	118.495	122.943	127.107	135.978	147.351
90	107.566	108.462	109.442	110.526	111.745	113.146	114.806	116.869	119.649	124.117	128.299	137.209	148.628
91	108.661	109.562	110.547	111.636	112.861	114.268	115.937	118.009	120.802	125.290	129.491	138.438	149.903
92	109.756	110.662	111.651	112.745	113.976	115.390	117.066	119.149	121.954	126.462	130.682	139.667	151.178
93	110.851	111.761	112.755	113.854	115.091	116.512	118.195	120.287	123.105	127.633	131.871	140.894	152.450
94	111.945	112.859	113.858	114.962	116.205	117.632	119.324	121.425	124.256	128.804	133.060	142.119	153.721
95	113.038	113.957	114.960	116.070	117.318	118.752	120.451	122.562	125.405	129.973	134.247	143.344	154.991
96	114.131	115.054	116.062	117.177	118.431	119.871	121.578	123.699	126.554	131.142	135.434	144.567	156.259
97	115.224	116.151	117.164	118.284	119.543	120.990	122.705	124.834	127.702	132.309	136.619	145.790	157.526
98	116.316	117.247	118.264	119.390	120.655	122.108	123.830	125.969	128.850	133.476	137.804	147.011	158.792
99	117.407	118.343	119.365	120.495	121.766	123.226	124.955	127.104	129.996	134.642	138.987	148.231	160.056
100	118.499	119.438	120.465	121.600	122.876	124.343	126.080	128.237	131.142	135.807	140.170	149.450	161.319

Tabelle A.4: χ^2_α-Signifikanzwerte für 76 bis 100 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
101	119.589	120.533	121.564	122.704	123.986	125.459	127.204	129.370	132.287	136.972	141.352	150.668	162.581
102	120.679	121.628	122.663	123.808	125.096	126.575	128.327	130.502	133.432	138.135	142.533	151.884	163.841
103	121.769	122.721	123.761	124.912	126.205	127.690	129.449	131.634	134.575	139.298	143.713	153.100	165.100
104	122.858	123.815	124.859	126.015	127.313	128.804	130.571	132.765	135.718	140.460	144.892	154.315	166.358
105	123.947	124.908	125.957	127.117	128.421	129.918	131.693	133.895	136.861	141.621	146.070	155.528	167.615
106	125.036	126.000	127.054	128.219	129.528	131.032	132.813	135.025	138.002	142.781	147.248	156.741	168.870
107	126.124	127.093	128.150	129.320	130.635	132.145	133.934	136.154	139.143	143.941	148.424	157.952	170.124
108	127.212	128.184	129.246	130.421	131.741	133.257	135.053	137.283	140.284	145.099	149.600	159.163	171.377
109	128.299	129.275	130.342	131.521	132.847	134.369	136.172	138.411	141.423	146.257	150.775	160.373	172.629
110	129.386	130.366	131.437	132.621	133.952	135.481	137.291	139.538	142.562	147.415	151.949	161.581	173.880
111	130.472	131.457	132.532	133.721	135.057	136.592	138.409	140.665	143.701	148.571	153.122	162.789	175.129
112	131.558	132.547	133.626	134.820	136.161	137.702	139.526	141.791	144.838	149.727	154.295	163.996	176.378
113	132.644	133.636	134.720	135.919	137.265	138.812	140.643	142.917	145.975	150.883	155.467	165.202	177.625
114	133.729	134.726	135.814	137.017	138.369	139.921	141.760	144.042	147.112	152.037	156.638	166.407	178.871
115	134.814	135.814	136.907	138.115	139.472	141.030	142.876	145.166	148.248	153.191	157.808	167.611	180.116
116	135.899	136.903	137.999	139.212	140.574	142.139	143.991	146.290	149.383	154.344	158.978	168.814	181.360
117	136.983	137.991	139.092	140.309	141.676	143.247	145.106	147.414	150.518	155.497	160.146	170.016	182.603
118	138.067	139.079	140.184	141.405	142.778	144.354	146.220	148.536	151.652	156.649	161.315	171.218	183.845
119	139.150	140.166	141.275	142.501	143.879	145.461	147.334	149.659	152.786	157.800	162.482	172.418	185.086
120	140.233	141.253	142.366	143.597	144.980	146.568	148.448	150.781	153.919	158.951	163.649	173.618	186.326
121	141.316	142.339	143.457	144.692	146.080	147.674	149.561	151.902	155.051	160.101	164.815	174.817	187.565
122	142.398	143.426	144.547	145.787	147.180	148.780	150.673	153.023	156.183	161.250	165.980	176.015	188.804
123	143.480	144.512	145.637	146.882	148.280	149.885	151.785	154.143	157.315	162.399	167.145	177.212	190.041
124	144.562	145.597	146.727	147.976	149.379	150.990	152.897	155.263	158.445	163.547	168.309	178.409	191.277
125	145.643	146.682	147.816	149.070	150.478	152.094	154.008	156.383	159.576	164.695	169.472	179.604	192.512

Tabelle A.5: χ^2_α-Signifikanzwerte für 101 bis 125 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
126	146.725	147.767	148.905	150.163	151.576	153.198	155.119	157.501	160.706	165.842	170.635	180.799	193.746
127	147.805	148.852	149.994	151.256	152.674	154.302	156.229	158.620	161.835	166.988	171.797	181.994	194.979
128	148.886	149.936	151.082	152.349	153.772	155.405	157.339	159.738	162.964	168.134	172.958	183.187	196.212
129	149.966	151.019	152.170	153.441	154.869	156.508	158.448	160.855	164.092	169.279	174.119	184.380	197.443
130	151.046	152.103	153.257	154.533	155.966	157.610	159.557	161.972	165.220	170.424	175.279	185.571	198.674
131	152.125	153.186	154.344	155.624	157.062	158.712	160.666	163.089	166.347	171.568	176.438	186.763	199.903
132	153.204	154.269	155.431	156.715	158.158	159.814	161.774	164.205	167.474	172.711	177.597	187.953	201.132
133	154.283	155.352	156.518	157.806	159.254	160.915	162.882	165.321	168.600	173.854	178.756	189.143	202.360
134	155.362	156.434	157.604	158.897	160.349	162.016	163.989	166.436	169.726	174.997	179.913	190.332	203.587
135	156.440	157.516	158.690	159.987	161.444	163.117	165.096	167.551	170.851	176.139	181.070	191.520	204.814
136	157.518	158.597	159.775	161.077	162.539	164.217	166.203	168.665	171.976	177.280	182.227	192.708	206.039
137	158.596	159.679	160.860	162.166	163.633	165.316	167.309	169.779	173.101	178.421	183.383	193.895	207.264
138	159.673	160.760	161.945	163.255	164.727	166.416	168.414	170.893	174.225	179.562	184.538	195.081	208.488
139	160.751	161.840	163.030	164.344	165.821	167.515	169.520	172.006	175.348	180.701	185.693	196.266	209.711
140	161.827	162.921	164.114	165.433	166.914	168.613	170.625	173.119	176.471	181.841	186.847	197.451	210.933
141	162.904	164.001	165.198	166.521	168.007	169.712	171.729	174.231	177.594	182.980	188.001	198.636	212.154
142	163.980	165.081	166.282	167.609	169.099	170.810	172.834	175.343	178.716	184.118	189.154	199.819	213.375
143	165.057	166.160	167.365	168.696	170.192	171.907	173.938	176.455	179.838	185.256	190.307	201.002	214.595
144	166.132	167.240	168.448	169.784	171.284	173.005	175.041	177.566	180.959	186.393	191.459	202.184	215.814
145	167.208	168.319	169.531	170.871	172.375	174.101	176.144	178.677	182.080	187.530	192.611	203.366	217.032
146	168.283	169.398	170.614	171.957	173.466	175.198	177.247	179.787	183.201	188.667	193.762	204.547	218.250
147	169.358	170.476	171.696	173.044	174.557	176.294	178.349	180.897	184.321	189.803	194.912	205.727	219.466
148	170.433	171.554	172.778	174.130	175.648	177.390	179.451	182.007	185.441	190.938	196.062	206.907	220.683
149	171.507	172.632	173.860	175.216	176.738	178.486	180.553	183.116	186.560	192.074	197.212	208.086	221.898
150	172.582	173.710	174.941	176.301	177.829	179.581	181.655	184.225	187.679	193.208	198.361	209.265	223.113

Tabelle A.6: χ^2_α-Signifikanzwerte für 126 bis 150 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
151	173.656	174.787	176.022	177.386	178.918	180.676	182.756	185.334	188.798	194.342	199.509	210.443	224.327
152	174.729	175.864	177.103	178.471	180.008	181.771	183.856	186.442	189.916	195.476	200.657	211.621	225.540
153	175.803	176.941	178.184	179.556	181.097	182.865	184.957	187.550	191.033	196.609	201.805	212.797	226.752
154	176.876	178.018	179.264	180.640	182.186	183.959	186.057	188.657	192.151	197.742	202.952	213.974	227.964
155	177.949	179.094	180.344	181.724	183.274	185.053	187.157	189.764	193.268	198.875	204.098	215.149	229.175
156	179.022	180.171	181.424	182.808	184.363	186.146	188.256	190.871	194.384	200.007	205.245	216.324	230.386
157	180.095	181.247	182.503	183.892	185.451	187.239	189.355	191.978	195.500	201.138	206.390	217.499	231.596
158	181.167	182.322	183.583	184.975	186.539	188.332	190.454	193.084	196.616	202.269	207.535	218.673	232.805
159	182.239	183.398	184.662	186.058	187.626	189.425	191.552	194.190	197.732	203.400	208.680	219.847	234.013
160	183.311	184.473	185.741	187.141	188.713	190.517	192.651	195.295	198.847	204.531	209.824	221.019	235.221
161	184.383	185.548	186.819	188.223	189.800	191.609	193.748	196.400	199.962	205.661	210.968	222.192	236.428
162	185.454	186.623	187.897	189.306	190.887	192.701	194.846	197.505	201.076	206.790	212.112	223.364	237.635
163	186.525	187.697	188.975	190.388	191.973	193.792	195.943	198.609	202.190	207.919	213.254	224.535	238.841
164	187.596	188.771	190.053	191.469	193.059	194.883	197.040	199.713	203.304	209.048	214.397	225.706	240.046
165	188.667	189.845	191.131	192.551	194.145	195.974	198.137	200.817	204.417	210.176	215.539	226.876	241.251
166	189.738	190.919	192.208	193.632	195.231	197.064	199.233	201.921	205.530	211.304	216.681	228.046	242.455
167	190.808	191.993	193.285	194.713	196.316	198.155	200.329	203.024	206.643	212.432	217.822	229.215	243.658
168	191.878	193.066	194.362	195.794	197.401	199.245	201.425	204.127	207.755	213.559	218.963	230.384	244.861
169	192.948	194.139	195.439	196.875	198.486	200.334	202.521	205.229	208.867	214.686	220.103	231.552	246.063
170	194.018	195.212	196.515	197.955	199.571	201.424	203.616	206.332	209.979	215.812	221.243	232.720	247.265
171	195.087	196.285	197.592	199.035	200.655	202.513	204.711	207.434	211.090	216.938	222.382	233.887	248.466
172	196.157	197.358	198.668	200.115	201.739	203.602	205.805	208.535	212.201	218.064	223.522	235.054	249.667
173	197.226	198.430	199.743	201.194	202.823	204.691	206.900	209.637	213.312	219.189	224.660	236.220	250.866
174	198.295	199.502	200.819	202.274	203.906	205.779	207.994	210.738	214.422	220.314	225.799	237.386	252.066
175	199.364	200.574	201.894	203.353	204.990	206.867	209.088	211.839	215.532	221.439	226.937	238.551	253.264

Tabelle A.7: χ^2_α-Signifikanzwerte für 151 bis 175 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
176	200.432	201.646	202.969	204.432	206.073	207.955	210.181	212.939	216.642	222.563	228.074	239.716	254.463
177	201.500	202.717	204.044	205.510	207.156	209.043	211.274	214.039	217.751	223.687	229.211	240.881	255.660
178	202.568	203.788	205.119	206.589	208.238	210.130	212.368	215.139	218.860	224.811	230.348	242.044	256.857
179	203.636	204.859	206.193	207.667	209.321	211.217	213.460	216.239	219.969	225.934	231.484	243.208	258.054
180	204.704	205.930	207.268	208.745	210.403	212.304	214.553	217.338	221.078	227.057	232.620	244.371	259.250
181	205.772	207.001	208.342	209.823	211.485	213.391	215.645	218.437	222.186	228.179	233.756	245.534	260.445
182	206.839	208.071	209.416	210.900	212.567	214.478	216.737	219.536	223.294	229.301	234.891	246.696	261.640
183	207.906	209.142	210.489	211.978	213.648	215.564	217.829	220.635	224.401	230.423	236.026	247.857	262.834
184	208.973	210.212	211.563	213.055	214.729	216.650	218.920	221.733	225.509	231.545	237.160	249.019	264.028
185	210.040	211.282	212.636	214.132	215.810	217.735	220.012	222.831	226.616	232.666	238.295	250.179	265.222
186	211.107	212.351	213.709	215.208	216.891	218.821	221.103	223.929	227.722	233.787	239.428	251.340	266.414
187	212.173	213.421	214.782	216.285	217.972	219.906	222.193	225.026	228.829	234.907	240.562	252.500	267.607
188	213.240	214.490	215.855	217.361	219.052	220.991	223.284	226.124	229.935	236.027	241.695	253.659	268.798
189	214.306	215.559	216.927	218.437	220.132	222.076	224.374	227.221	231.041	237.147	242.828	254.818	269.990
190	215.372	216.628	217.999	219.513	221.212	223.161	225.464	228.317	232.146	238.267	243.960	255.977	271.180
191	216.437	217.697	219.071	220.589	222.292	224.245	226.554	229.414	233.252	239.386	245.092	257.135	272.371
192	217.503	218.766	220.143	221.665	223.372	225.329	227.644	230.510	234.357	240.505	246.224	258.293	273.560
193	218.568	219.834	221.215	222.740	224.451	226.413	228.733	231.606	235.462	241.624	247.355	259.450	274.750
194	219.634	220.903	222.287	223.815	225.530	227.497	229.822	232.702	236.566	242.742	248.486	260.607	275.938
195	220.699	221.971	223.358	224.890	226.609	228.580	230.911	233.797	237.670	243.860	249.616	261.764	277.127
196	221.764	223.039	224.429	225.965	227.688	229.664	231.999	234.892	238.774	244.978	250.747	262.920	278.314
197	222.828	224.106	225.500	227.039	228.766	230.747	233.088	235.987	239.878	246.095	251.877	264.076	279.502
198	223.893	225.174	226.571	228.114	229.845	231.830	234.176	237.082	240.981	247.212	253.006	265.231	280.688
199	224.957	226.241	227.642	229.188	230.923	232.912	235.264	238.176	242.085	248.329	254.136	266.386	281.875
200	226.022	227.308	228.712	230.262	232.001	233.995	236.352	239.271	243.187	249.446	255.265	267.541	283.061

Tabelle A.8: χ^2_α-Signifikanzwerte für 176 bis 200 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
201	227.086	228.375	229.782	231.336	233.079	235.077	237.439	240.365	244.290	250.562	256.393	268.695	284.246
202	228.150	229.442	230.852	232.409	234.156	236.159	238.527	241.458	245.392	251.678	257.522	269.849	285.431
203	229.213	230.509	231.922	233.483	235.233	237.241	239.614	242.552	246.495	252.793	258.650	271.003	286.616
204	230.277	231.576	232.992	234.556	236.311	238.323	240.701	243.645	247.596	253.909	259.777	272.156	287.800
205	231.340	232.642	234.062	235.629	237.388	239.404	241.787	244.738	248.698	255.024	260.905	273.308	288.983
206	232.404	233.708	235.131	236.702	238.464	240.485	242.874	245.831	249.800	256.139	262.032	274.461	290.167
207	233.467	234.774	236.200	237.775	239.541	241.566	243.960	246.924	250.901	257.253	263.159	275.613	291.349
208	234.530	235.840	237.269	238.847	240.617	242.647	245.046	248.016	252.002	258.368	264.285	276.764	292.532
209	235.593	236.906	238.338	239.919	241.694	243.728	246.132	249.109	253.102	259.482	265.411	277.916	293.713
210	236.655	237.972	239.407	240.992	242.770	244.808	247.217	250.201	254.203	260.595	266.537	279.067	294.895
211	237.718	239.037	240.475	242.064	243.846	245.888	248.303	251.292	255.303	261.709	267.663	280.217	296.076
212	238.780	240.102	241.544	243.135	244.921	246.968	249.388	252.384	256.403	262.822	268.788	281.367	297.256
213	239.843	241.167	242.612	244.207	245.997	248.048	250.473	253.475	257.502	263.935	269.913	282.517	298.437
214	240.905	242.232	243.680	245.279	247.072	249.128	251.558	254.566	258.602	265.047	271.038	283.667	299.616
215	241.967	243.297	244.748	246.350	248.147	250.208	252.642	255.657	259.701	266.160	272.162	284.816	300.796
216	243.029	244.362	245.816	247.421	249.222	251.287	253.727	256.748	260.800	267.272	273.286	285.965	301.975
217	244.090	245.426	246.883	248.492	250.297	252.366	254.811	257.838	261.899	268.384	274.410	287.113	303.153
218	245.152	246.491	247.951	249.563	251.372	253.445	255.895	258.929	262.997	269.495	275.533	288.261	304.331
219	246.213	247.555	249.018	250.634	252.446	254.524	256.979	260.019	264.096	270.607	276.657	289.409	305.509
220	247.274	248.619	250.085	251.704	253.520	255.602	258.063	261.108	265.194	271.718	277.780	290.556	306.686
221	248.335	249.683	251.152	252.775	254.595	256.681	259.146	262.198	266.292	272.829	278.902	291.703	307.863
222	249.396	250.747	252.219	253.845	255.669	257.759	260.229	263.288	267.389	273.939	280.025	292.850	309.039
223	250.457	251.810	253.286	254.915	256.742	258.837	261.312	264.377	268.487	275.049	281.147	293.997	310.215
224	251.518	252.874	254.352	255.985	257.816	259.915	262.395	265.466	269.584	276.160	282.269	295.143	311.391
225	252.579	253.937	255.419	257.055	258.889	260.993	263.478	266.555	270.681	277.269	283.390	296.288	312.566

Tabelle A.9: χ^2_α-Signifikanzwerte für 201 bis 225 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Freitheits-grade	α-Signifikanzniveau												
	$\chi^2_{0.1}$	$\chi^2_{0.09}$	$\chi^2_{0.08}$	$\chi^2_{0.07}$	$\chi^2_{0.06}$	$\chi^2_{0.05}$	$\chi^2_{0.04}$	$\chi^2_{0.03}$	$\chi^2_{0.02}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	$\chi^2_{0.001}$	$\chi^2_{0.0001}$
226	253.639	255.001	256.485	258.124	259.963	262.070	264.560	267.643	271.778	278.379	284.512	297.434	313.741
227	254.699	256.064	257.551	259.193	261.036	263.148	265.643	268.732	272.874	279.488	285.633	298.579	314.916
228	255.759	257.127	258.617	260.263	262.109	264.225	266.725	269.820	273.971	280.597	286.754	299.724	316.090
229	256.819	258.189	259.683	261.332	263.182	265.302	267.807	270.908	275.067	281.706	287.874	300.868	317.263
230	257.879	259.252	260.749	262.401	264.254	266.379	268.889	271.996	276.163	282.815	288.994	302.012	318.437
231	258.939	260.314	261.814	263.470	265.327	267.455	269.970	273.084	277.259	283.923	290.114	303.156	319.610
232	259.999	261.377	262.880	264.538	266.399	268.532	271.052	274.171	278.354	285.032	291.234	304.300	320.782
233	261.058	262.439	263.945	265.607	267.471	269.608	272.133	275.258	279.449	286.139	292.353	305.443	321.955
234	262.118	263.501	265.010	266.675	268.543	270.684	273.214	276.345	280.544	287.247	293.473	306.586	323.127
235	263.177	264.563	266.075	267.743	269.615	271.760	274.295	277.432	281.639	288.355	294.592	307.729	324.298
236	264.236	265.625	267.140	268.812	270.687	272.836	275.376	278.519	282.734	289.462	295.710	308.871	325.469
237	265.295	266.687	268.204	269.879	271.758	273.912	276.456	279.606	283.829	290.569	296.829	310.013	326.640
238	266.354	267.748	269.269	270.947	272.830	274.988	277.537	280.692	284.923	291.676	297.947	311.155	327.810
239	267.413	268.810	270.333	272.015	273.901	276.063	278.617	281.778	286.017	292.782	299.065	312.296	328.980
240	268.471	269.871	271.398	273.082	274.972	277.138	279.697	282.864	287.111	293.889	300.183	313.437	330.150
241	269.530	270.932	272.462	274.150	276.043	278.213	280.777	283.950	288.205	294.995	301.300	314.578	331.320
242	270.588	271.994	273.526	275.217	277.114	279.288	281.857	285.036	289.298	296.101	302.417	315.719	332.489
243	271.646	273.054	274.590	276.284	278.185	280.363	282.936	286.121	290.391	297.206	303.534	316.859	333.657
244	272.705	274.115	275.653	277.351	279.255	281.438	284.016	287.207	291.485	298.312	304.651	317.999	334.826
245	273.763	275.176	276.717	278.418	280.326	282.512	285.095	288.292	292.577	299.417	305.768	319.139	335.993
246	274.820	276.237	277.780	279.485	281.396	283.586	286.174	289.377	293.670	300.522	306.884	320.278	337.161
247	275.878	277.297	278.844	280.551	282.466	284.660	287.253	290.461	294.763	301.627	308.000	321.417	338.328
248	276.936	278.357	279.907	281.617	283.536	285.734	288.332	291.546	295.855	302.731	309.116	322.556	339.495
249	277.993	279.418	280.970	282.684	284.606	286.808	289.410	292.630	296.947	303.836	310.231	323.695	340.662
250	279.051	280.478	282.033	283.750	285.675	287.882	290.489	293.715	298.039	304.940	311.347	324.833	341.828

Tabelle A.10: χ^2_α-Signifikanzwerte für 226 bis 250 Freiheitsgrade zu einem akzeptierten α-Fehler (Typ-1-Fehler).

Anhang B

Wissenschaftlicher Werdegang

Marco Büchler
Erich-Zeigner-Allee 51
04229 Leipzig

geb. am 24. Juli 1978
geb. in Eilenburg, Sachsen

Automatische Sprachverarbeitung
Institut für Informatik
Universität Leipzig
Augustusplatz 10/11
04109 Leipzig

Kontakt:
Raum : P(aulinum)818
eMail : mbuechler@e-humanities.net
Phone : 0341/97-32257

B Wissenschaftlicher Werdegang **245**
B.1 Wissenschaftlicher Lebenslauf 246
B.2 Praxiserfahrungen in der Informatik sowie in den *Humanities* 246
B.3 Auszeichnungen und Preise 247
B.4 Aktivitäten in *Advisory Boards*, als *Reviewer* und *Initiator*, in *Programmkomitees* sowie Unterstützung anderer Forscher und Projekte 247
B.5 Bearbeitete wissenschaftliche Projekte 248
B.6 Wissenschaftliche Akquise 249
B.7 Organisierte Workshops 249
B.8 Expert Talks, Invited Talks und Interviews 250
B.9 Besuchte Veranstaltungen 251
B.10 Bücher, Whitepaper und strategische Dokumente 252
B.11 Publikationen 252
B.12 Vorträge 254
B.13 Poster und Posterdemonstrationen 256
B.14 Lehrveranstaltungen 256
B.15 Betreute Abschlussarbeiten 257

B.1 wissenschaftlicher Lebenslauf

04/2011 – 02/2013	Promotion im Bereich der *eHumanities* am Institut für Informatik zum Thema *Informationstechnische Aspekte des Historical Text Re-use*
seit 07/2011	*Principle Investigator* des *eTRACES*-Projektes zum Thema *Text Re-use* in Kooperation mit den Partnern *GESIS* (Köln), dem *Göttingen Centre for Digital Humanities* (Göttingen) sowie dem *Deutschen Textarchiv an der Berlin-Brandenburgischen Akademie der Wissenschaften* (Berlin)
10/2009 – 04/2010	Gastwissenschaftler beim *Perseus*-Project an der Tufts University, Boston, MA, USA
04/2008 – 03/2011	Technischer Koordinator des *eAQUA*-Projektes in Kooperation mit dem *Geisteswissenschaftlichen Zentrum* (Leipzig), der *Bundeswehruniversität Hamburg* und der *Universität Heidelberg*
seit 07/2006	wissenschaftlicher Mitarbeiter am Lehrstuhl für *Automatische Sprachverarbeitung* an der *Universität Leipzig*
07/2006	Diplom der Informatik am *Institut für Informatik* der *Universität Leipzig*
08/2002	Vordiplom der Informatik am *Institut für Informatik* der *Universität Leipzig*
10/1999 – 07/2006	Studium der Informatik an der *Universität Leipzig*

B.2 Praxiserfahrungen in der Informatik sowie in den *Humanities*

10/2011 – 07/2012	Arabisch-Kurs an der Universität in Halle/S.
01/2012 – 06/2012	Altgriechisch-Kurs an der Universität in Leipzig
2011	universitärer IT-Consultant im Bereich *Semantischer Suchmaschinen* für *Interface:Projects AG*, Dresden
10/2009 - 04/2010	universitärer IT-Consultant im Bereich *Semantisches Retrieval zur Semiautomatischen Beantwortung von eingehenden eMails an einen technischen Support* für *T-Systems MMS*, Dresden
10/2008	archäologische Fachexkursion nach Kleinasien (Westtürkei) mit Besuchen von u. a. Pergamon, Troja, Ephesos, Milet und Didyma
03/2008 – 08/2010	universitärer und freiberuflicher IT-Consultant im Bereich *Graphbasiertes Browsing in großen Textkollektionen* bei der *NetCon Solutions AG*, Leipzig
07/2006 – 04/2007	freiberuflicher IT-Consultant für *QuBit GbR*
03/2000 – 05/2002	Team- und Abteilungsleiter *Quality Management* bei der *Virbus AG, Leipzig* mit der Spezialisierung auf Last- und Performancetests

B.3 Auszeichnungen und Preise

01/2011 NVIDIA Award (Hardware-Sponsoring) für den Beitrag *Thinking in Signals: Measuring Text Re-use with Signal Processing Algorithms*

07/2010 2010 Alliance of Digital Humanities Organizations Award at Digital Humanities Conference für das Papier *Detection of Citations and Textual Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project* at 2010 Digital Humanities Conference (London, UK).

07/2010 Notable Mentions at Developer Challenge of THATCamp, London für den Beitrag *Bringing Modern Spell Checking Approaches to Ancient Greek Papyri - Completing Fragmentary Texts* at 2010 Digital Humanities Conference (London, UK).

11/2009 Receiver of a Student Award für den Beitrag über *Discovering Latent Relations of Concepts by Graph Mining Approaches on Ancient Greek Texts* at 2009 Chicago Colloquium on Digital Humanities and Computer Science (Chicago, USA)

B.4 Aktivitäten in *Advisory Boards*, als *Reviewer* und *Initiator*, in *Programmkomitees* sowie Unterstützung anderer Forscher und Projekte

04/2013 *Scientific host* für Elton Barker (Open University) im Rahmen eines Forschungsaufenthaltes bei der *Leipzig eHumanities Research Group* zwischen April und September 2013

01/2013 *Empfehlungsschreiben* an Eleni Bozia (*Department of Classics, University of Florida, USA*) für die Projektantragstellung im Bereich der archäologischen Digitalisierung von Münzen, Abklatschen und Artefakten

11/2012 *Empfehlungsschreiben* an Bridget Almas und Gregory Crane (*Alphaios.net & Perseus Project am Department of Classics an der Tufts University, Boston, USA*) für die Projektantragstellung im Rahmen des *Bamboo Projects* bei der *Mellon Foundation*

11/2012 *Empfehlungsschreiben* an Neil Coffee (*Tesserae Project am Department of Classics an der Buffalo University, Buffalo, NY, USA*) für die Projektantragstellung *The Tesserae Project: Enhanced Automatic Detection of Allusion* bei der *Loeb Classical Library Foundation* an der *Harvard University, USA*

10/2012 *Scientific host* für *Angelo Del Grosso* (Institute of Computational Linguistics CNR of Pisa, Italy) im Rahmen eines Forschungsaufenthaltes bei der *Leipzig eHumanities Research Group* zwischen Oktober und Dezember 2012

seit 2012 *Contributor Reviewer* für das *Digital Humanities Quarterly*

seit 2012 *Initiator* und *Programmkomitee* für den *Annual eHumanities Innovation Award*

seit 2012	*Initiator* und *Programmkomitee* für das *Leipzig eHumanities Research Seminar*
11/2011	Gründer der Google Gruppe *Historical Text Re-use*, welcher bis Januar 2013 knapp 50 nationale und internationale Forscher angehören
seit 2011	*Advisory Board* für das *Tesserae project: Intertextual Phrase Matching*
seit 2011	*Paper* und *Poster Reviewer* für die *Digitial Humanities Conference*
seit 2011	*Grant Proposal Reviewer* für das *Dutch Council for the Humanities* NWO
09/2011	*Empfehlungsschreiben* an Neil Coffee (*Tesserae Project am Department of Classics an der Buffalo University, Buffalo, NY, USA*) für die Projektantragstellung *The Tesserae Project* im Rahmen des *NEH ODH START UP GRANT PHASE II*

B.5 Bearbeitete wissenschaftliche Projekte

seit 09/2011	*Digitalisierung der Daten der Deutschen Morgenländischen Gesellschaft*
09/2012 – 09/2013	*Optical Character Recognition and Text Re-use Detection of Polytonic Greek in the Google Books Ancient Greek and Latin Corpus* im Rahmen des Compute Canada RAC Proposal, PI: Bruce Robertson, Canada
seit 07/2011	*eTRACES*: Recherche und Analyse von Zitationsspuren und Wissenstransfer in sozialwissenschaftlichen Texten und deutschsprachiger Literatur
09/2010 – 12/2012	*TRACER*: A Java based Software Library for the Detection of *Historical Text Te-use*
seit 03/2009	*GnomeDB*: Datenbank der Syrischen und Arabischen Gnomologien (CASG)
04/2008 – 03/2011	*eAQUA*: Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft
09/2005 – 03/2006	*MPI Linguistic Services*: Enabling Max Planck Institute for Accessing Leipzig Linguistic Services
seit 2005	*Medusa*: A Statistical Framework for High Speed Text Processing
seit 2004	*Leipzig Linguistic Services*: Enabling SOAP-based Access to Wortschatz Databases for a Public Community

B.6 Wissenschaftliche Akquise

2013 *Reading Texts Spatially: Deep Maps for the Humanities Chorography.* Beantragt bei *ERC Starting Grant. Principle Investigator*: Elton Barker, Open University, Oxford, UK. (eingereicht und im Reviewprozess)

2012 Forschungsstipendium für *Angelo Del Grosso* (Institute of Computational Linguistics CNR of Pisa, Italy) für einen Forschungsaufenthalt bei der *Leipzig eHumanities Research Group* zwischen Oktober und Dezember 2012

2012 Hardwareförderung: *Optical Character Recognition and Text Re-use Detection of Polytonic Greek in the Google Books Ancient Greek and Latin Corpus.* Gefördert durch *Compute Canada RAC Initiative. Principle Investigator*: Bruce Robertson, Mount Alison, Kanada

2011 *eTRACES*[1]: *Recherche und Analyse von Zitationsspuren und Wissenstransfer in sozialwissenschaftlichen Texten und deutschsprachiger Literatur* beim Bundesministerium für Bildung und Forschung

2007 *eAQUA*[2]: *Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft* beim Bundesministerium für Bildung und Forschung

B.7 Organisierte Workshops

2012 Workshop: Marco Büchler, Gregory Crane: *Please Mind the Gap: How to Bridge Recent Infrastructure Activities* an der Universität Leipzig.

2012 Hackathon: Monica Berti, Marco Büchler, Gregory Crane: *The Banquet of the Digital Scholars: Humanities Hackathon on editing Athenaeus and on the Reinvention of the Edition in a Digital Space* an der Universität Leipzig.

2011 Workshop: Marco Büchler: *Medusa Technical Teaching* bei *Interface:Projects AG*, Dresden.

2011 Workshop: Tom Brughmans, Marco Büchler: *Thinking through networks: generating, visualising and analysing complex re-use graphs in the Humanities* auf der *2011 Interface Conference*, London, UK.

2010 Workshop: Gerhard Heyer, Marco Büchler: *eHumanities - How does computer science benefit?*. Auf dem *44. Jahrestag der Gesellschaft für Informatik e.V.: Service Science - Neue Perspektiven für die Informatik.* Leipzig, 2010.

2010 Workshop: Gerhard Heyer, Marco Büchler, Thomas Eckart, Charlotte Schubert: *Text Mining in the Digital Humanities* auf der *2010 Digital Humanities Conference* London, UK, 2010.

2009 Workshop: Marco Büchler, Andre Bünte: *Workshop on Text Mining for Humanities.* At *Gerhard Heyer: Text Mining Services Conference*, Leipzig, 2009.

2008 eAQUA-interner Workshop zur Qualifizierung der Geisteswissenschaftler im Bereich Text Mining am Beispiel der geisteswissenschaftlichen Teilprojekte: Marco Büchler, Gerhard Heyer: *Text Mining for Classical Studies.* Leipzig, 2008-9.

[1] Antragsteller und Projektkoordinator: Gerhard Heyer
[2] Antragsteller und Projektkoordinator: Gerhard Heyer und Charlotte Schubert

B.8 Expert Talks, Invited Talks und Interviews

09/2013 Expert Talk: *Historical Text Re-use Analysis: Wie kann die Informatik die geisteswissenschaftliche Forschung unterstützen?* auf der Konferenz *<philtag n="11"/>* an der Universität Würzburg, September 25/26, 2013.

04/2013 Invited Talk: *eTRACES: Basics and Applications of Historical Text Re-use Detection* im Rahmen der Konferenz *Word, Space, Time: Digital Perspectives on the Classical World* veranstaltet von der *Digital Classics Association*, Buffalo, NY, USA, April 5/6, 2013.

03/2013 Expert Talk: *Text Re-use on Noisy Data: Using Text Re-use Graphs for Page-Ranking on OCR'ed Digital Libraries* im Rahmen der Veranstaltung *Forschung im Netz von Marbach-Weimar-Wolfenbüttel* an der *Herzog-August-Bibliothek* in Wolfenbüttel, März 25/26, 2013.

03/2013 Expert Talk: *Text Re-use and Bibliometrics - About Applications of Formal Text Re-use Graphs* auf dem *13th International Symposium of Information Science* in Potsdam, März 21, 2013.

12/2011 Invited Talk: *Interdisciplinary Work between Computer Science, eHumanities, Digital Humanities, and Humanities: Four Different Views to the Topic of Text Re-use* im Rahmen der Veranstaltung *Perspektiven einer corpusbasierten historischen Linguistik und Philologie* an der Berlin-Brandenburgischen Akademie der Wissenschaften zu Berlin, Dezember 12, 2011.

11/2011 Invited Talk: *Commercial and Humanities Text Mining Revealing Techniques in Ancient Greek Literature* für die *Textual Analysis Working Group* der *Digital Humanities Initiative Buffalo* am *Classics Department* der *Buffalo University, Buffalo , NY, USA*, November 22, 2011.

11/2011 Invited Talk: *Semantic Exploration of Massive Data in a Digital Library - About Semantics in Historical Texts* im Rahmen der Veranstaltung *Progresso scientifico e trasformazione della cultura del libro – l'illustrazione come strumento delle scienze dell'antichità* am Deutschen Archäologischen Institut in Rom, Italien, November 18, 2011.

10/2011 Expert Talk: *Textvervollständigung, OCR- und Rechtschreibkorrektur - Drei Sichten auf gleiche Methoden* im Rahmen der Veranstaltung *Erfahrungen aus der Digitalisierungspraxis: OCR, Volltexte und Präsentationsformen* des *IMPACT*-Projektes an der *Bayerischen Staatsbibliothek* in München, Oktober 12, 2011.

07/2011 Invited Talk: Marco Büchler: *Text Mining in den eHumanities - Aspects of eAQUA and eTRACES* an der Technischen Universität Darmstadt.

02/2011 Expert Talk: *Gnomology Database of Arabic and Syriac - Bridging Ancient Texts and Modern Language Models* im Rahmen des Workshops *Exploring Formulaic Knowledge through Languages, Cultures and Time* an der Universität Trier, Februar 02, 2011.

02/2011 Expert Talk: *About Text Re-use, Knowledge Transfer, and Applications* im Rahmen eines Workshops vom Projekt *Sharing Ancient Wisdoms*, Wien, Österreich, Februar 16, 2011.

02/2011 Invited Talk: *Text Mining in den Fachwissenschaften - About the Gap between Search and Find* bei den Religionswissenschaften an der Universität Leipzig.

09/2010 Plenarvortrag: *eAQUA: Quantitative Computer-Modelle in der qualitativen geisteswissenschaftlichen Forschung - Interdisziplinäre Arbeit zwischen Informatik und Altertumswissenschaften* im Rahmen der *D-Spin Sommerschule*, Bad Homburg, September 02, 2010

07/2010 Expert Talk: *„Fragments“ of eAQUA - Bridging Classics and Computer Science, Infrastructure and Graph Mining* am *Centre for Computing in the Humanities* am *King's College London*, London, UK, Juli 31, 2009

07/2010 Interview with Marion Lame auf der *2010 Digital Humanities Conference* in London, UK, Juli 16, 2010.

06/2010 Invited Talk: *Gnomology, Databases, and Text Re-use. Having a Database of Gnomai: What are the Next Steps?* am *Institut National des langues et Civilisations Orientales*, Paris, Frankreich, Juni 11, 2010.

B.9 Besuchte Veranstaltungen

02/2013 Deutsches Textarchiv: *Historische Textkorpora für die Geistes- und Sozialwissenschaften. Fragestellungen und Nutzungsperspektiven* an der *Berlin-Brandenburgischen Akademie der Wissenschaften*, Berlin, Februar 18/19, 2012.

06/2012 Hildelies Balk, Clemens Neudecker: *IMPACT OCR Workshop* an der *National library of the Netherlands*, Den Haag, Niederlande, Juni 26, 2012.

05/2012 Gerhard Heyer: *eHumanities an der Universität Leipzig*, Mai 2, 2012.

07/2012 Gerhard Lauer, Juan Garces: *Eröffnung des Göttingen Centre of Digital Humanities*, Göttingen, Juli 12, 2012.

04/2012 Gerhard Lauer: *Text and Opinion Mining in Humanities and Religious Studies Research* am *Göttingen Centre for Digital Humanities*, Göttingen, April 25, 2012.

11/2011 Gregory Crane: *Infrastructure for Historical Languages* am *Institut für deutsche Sprache und Linguistik*, *Humboldt-Universität zu Berlin*, Dezember 14/15, 2011.

07/2010 John Bradly, Gabriel Bodard: *Developers Challenge @ ThatCamp London (Humanities and Technology Camp)* am *King's College London*, London, UK , Juli 6, 2010.

04/2010 CLARIN: *Infrastructure and Security Workshop* am *National Institute for Subatomic Physics* in Amsterdam, Niederlande, April 27, 2010.

01/2010 Gregory Crane, Anke Lüdeling: *Workshop on Historical Texts* an der *Tufts University*, Boston, MA, USA, Januar 13-14, 2010.

11/2009 Volker Boehlke, Gerhard Heyer: *CLARIN-Workshop: Web Service and Workflow Aspects*, Leipzig, November 19/20, 2009.

09/2009 Martin Potthast, Benno Stein: *SEPLN´09 Workshop PAN. Uncovering Plagiarism, Authorship and Social Software Misuse.*, San Sebastian, Spanien, September 10, 2009.

B.10 Bücher, Whitepaper und strategische Dokumente

09/2010	Marco Büchler: *Bericht zur 2010 Digital Humanities Conference* in London für das Bundesministerium für Bildung und Forschung	[Büchler 2010a]
08/2010	Marco Büchler: *Reisebericht eines Informatikers durch die nationalen und internationalen Digital Humanities und eHumanities - April 2008 bis August 2010* für das Bundesministerium für Bildung und Forschung	[Büchler 2010b]
04/2008	Marco Büchler: *Medusa: Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung.* VdM Verlag Dr. Müller, 2008.	[Büchler 2008a]
07/2005	Marco Büchler: *Analyse der XML-Performance für Suchanfragen*	[Büchler 2005]

B.11 Publikationen

2013	Marco Büchler, Gregory Crane: *Uncovering Serendipity from Historical Data – About Usage of Network Analysis in Humanities.* in Tom Brughmans: *The Connected Past: People, Networks and Complexity in Archaeology and History*, to be published, Oxford University Press, Oxford, UK 2013.	[Büchler 2013b]
2013	Marco Büchler, Maria Moritz: *Historical Bibliometrics: Historical Relevance Feedback Detection by Text Re-use Mining.* to be published, In: *Bibliometrie - Praxis und Forschung. Beitrag zur Konferenz Bibliometrie 2012.* Regensburg, 2013.	[Büchler 2013d]
2013	Marco Büchler, Annette Geßner, Monica Berti und Thomas Eckart: *Measuring the Influence of a Work by Text Re-Use.* In: Stuart Dunn und Simon Mahony eds., *Digital Classicist Supplement: Bulletin of the Institute of Classical Studies*, Wiley-Blackwell, 2013.	[Büchler 2013c]
2012	Marco Büchler, Gregory Crane, Maria Moritz und Alison Babeu: *Increasing Recall for Text Re-use in Historical Documents to Support Research in the Humanities.* In: *Proceedings of Theory and Practice of Digital Libraries 2012*, 2012	[Büchler 2012c]
2012	Marco Büchler, Sebastian Kruse und Thomas Eckart: *Bringing Modern Spell Checking Approaches to Ancient Texts - Automated Suggestions for Incomplete Words.* In: *Proceedings of Digital Humanities 2012*, Hamburg, Germany, 2012.	[Büchler 2012d]
2012	Marco Büchler, Gregory Crane und Gerhard Heyer: *Historical Relevance Feedback Detection by Text Re-use Mining.* In: Maximilian Schich, Roger Malina, Isabel Meirelles, Christian Huepe: *Arts, Humanities, and Complex Networks Living Companion at Arts, Humanities, and Complex Networks* — 3rd Leonardo Satellite Symposium hosted by NetSci2012, Evanston, IL, USA, 2012	[Büchler 2012b]

2012	Marco Büchler, Ute Pietruschka und Norman Wetzig: *Transmission of Greek Gnomologia in Syriac and Arabic: The Corpus of Arabic and Syriac Gnomologia.* In: *International Workshop "Exploring Formulaic Knowledge through Languages, Cultures and Time"*, Trier, 2012.	[Büchler 2012e]
2011	Marco Büchler, Gregory Crane, Martin Mueller, Philip Burns und Gerhard Heyer: *One Step Closer To Paraphrase Detection On Historical Texts: About The Quality of Text Re-use Techniques and the Ability to Learn Paradigmatic Relations*, In: *Proceedings of the 2011 Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, IL, USA, 2011.	[Büchler 2011c]
2011	Gerhard Heyer, Marco Büchler und Volker Boehlke: *Aspects of an Infrastructure for eHumanities.* In: *2011 Supporting Digital Humanities*, Copenhagen, Denmark, 2011.	[Heyer 2011b]
2011	Gerhard Heyer, Marco Büchler, Thomas Eckart und Maria Moritz: *eAQUA - Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaften: Technologien und Ansätze zu Infrastruktur, Text Mining und Knowledge Transfer.* Leipzig, Leipzig University, 2011.	[Heyer 2011c]
2011	Thomas Eckart, David Pansch und Marco Büchler: *Integration of Distributed Text Resources by Using Schema Matching Techniques.* In: *Proceedings of Digital Humanities 2011*, Stanford, USA, 2011.	[Eckart 2011]
2011	Marco Büchler, Stefan Beyer, Thomas Eckart und Ute Pietruschka: *Collecting Ancient Greek, Arabic, and Syriac Sapiential Statements - The Gnomology Database of Arabic and Syriac.* In: Marie-Christine Bornes-Varol and Marie-Sol Ortola: *2nd International Colloquium Aliento - Corpus anciens et Bases de donnees*, Press Universitaires de Nancy, 2011.	[Büchler 2011b]
2010	Gerhard Heyer, Marco Büchler: *Some Challenges Posed to Computer Science by the eHumanities.* GI Jahrestagung, Leipzig, 2010.	[Heyer 2010]
2010	Marco Büchler, Gerhard Heyer: *Salton und Wittgenstein in den Humanities: Über die Semantik in Philosophischen Texten.* GI Jahrestagung, Leipzig, 2010.	[Büchler 2010f]
2010	Markus Deufert, Judith Blumenstein, Andreas Trebesius, Stefan Beyer, Marco Büchler: *Objective Detection of Plautus' Rules by Computer Support.* In: *Proceedings of Digital Humanities 2010*, London, 2010.	[Deufert 2010]
2010	Marco Büchler, Annette Geßner, Gerhard Heyer, Thomas Eckart: *Detection of Citations and Text Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project.* In: *Proceedings of Digital Humanities 2010*, London, 2010.	[Büchler 2010e]
2010	Thomas Eckart, Reinhard Förtsch, Sebastian Kruse, Marco Büchler: *Accessing, Visualizing and Annotating Geographical Information in Archeology.* In: *Proceedings of CAA 2010*, Granada, Spain, 2010.	[Eckart 2010]

2010	Marco Büchler, Frederik Baumgardt, Thomas Eckart: *Von Platon zu Alexander dem Großen – Automatische Extraktion von Topic-Maps-basierten Assoziationsketten aus Sozialen Netzwerken der Antike.* In Detlef Reineke (Hrsg. eDITion): *Terminologie und Text Mining.* Ausgabe 1/2010. ISSN 1862-023X, Las Palmas de Gran Canaria, 2010.	[Büchler 2010c]
2010	Marco Büchler, Annette Geßner, Thomas Eckart: *Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts.* In: *Proceedings of the 2009 Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, 2010.	[Büchler 2010d]
2009	Marco Büchler, Lutz Maicher, Benjamin Bock, Frederick Baumgardt: *Automatic Extraction of Topic Maps based Argumentation Trails.* In: *Text Mining Services – Building and applying text mining based service infrastructures in research and industry.* Proceedings of the Conference on *Text Mining Services – TMS 2009* at Leipzig University, 8, 2009.	[Büchler 2009b]
2009	Marco Büchler, Gerhard Heyer: *Leipzig Linguistic Services - A 4 Years Summary of Providing Linguistic Web Services.* In: *Proceedings of TMS 2009 Conference*, Augustusplatz 10/11, 04109 Leipzig, Germany, 2009.	[Büchler 2009a]
2008	Marco Büchler, Gerhard Heyer und Sabine Gründer: *eAQUA - Bringing modern Text Mining approaches to two thousand years old ancient texts.* In: *Proceedings of the 4th IEEE International Conference on e-Science*, 2008.	[Büchler 2008c]
2007	Marco Büchler, Gerhard Heyer: *Kookurenzberechnungen mit UIMA und Medusa.* In: *UIMA Workshop at the GLDV 2007* in Tübingen/Germany, 2007.	[Büchler 2007]
2006	Marco Büchler: *Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten.* Diplomarbeit. Augustusplatz 10/11, 04109 Leipzig, Germany, 2006.	[Büchler 2006a]

B.12 Vorträge

2012 Marco Büchler: *eTRACES- Winged words, quotations and our cultural heritage - About Text Re-use Graphs and Their Applications* auf der *Culture & Technology: European Summer School in Digital Humanities*, Leipzig, Juli 25, 2012.

2012 Marco Büchler, Gregory Crane: *Historical Text Re-use Detection on Perseus Digital Library - About Text Re-use Graphs and Their Application* im *Digital Classicist & Institute of Classical Studies Seminar 2012*, London, UK, Juni 29, 2012.

2012 Marco Büchler: *ACID for the eHumanities - A new paradigm for successful eHumanities projects* auf der CLARIN-M12-Workshop, Leipzig, Juni 28, 2012.

2012 Marco Büchler: *Insights into Recent Research Activities of the Leipzig eHumanities Research Group - About Components of Next Generation Search Engines in Humanities* auf dem Workshop *eHumanities an der Universität Leipzig*, Leipzig, Mai 02, 2012.

2012 Marco Büchler: *Generation of Text Graphs and Text Re-use Graphs from Massive Digital Data - About Uncovering the Unexpected* auf der Konferenz *The Connected Past - People, Networks, and Complexity in Archaeology and History*, Southampton, UK, März 25, 2012.

2011 Marco Büchler: *Interdisciplinary Work between Computer Science, eHumanities, Digital Humanities, and Humanities: Four Different Views to the Topic of Text Re-use* auf dem Workshop *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, Berlin, Dezember 12, 2011.

2011 Marco Büchler: *Bringing Modern Spell Checking Approaches to Ancient Texts: Automatized Suggestions for Incomplete Words* im *Digital Classicist & Institute of Classical Studies Seminar 2012*, London, UK, Juli 29, 2011.

2010 Monica Berti, Marco Büchler: *Fragmentary Texts and Digital Collections of Fragmentary Authors* im *Digital Classicist & Institute of Classical Studies Seminar 2012*, London, UK, Juli 30, 2010.

2010 Marco Büchler: *eAQUA -Benefits of Interdisciplinary Work: Text Reuse & Knowledge Transfer* auf der *Culture & Technology: European Summer School in Digital Humanities*, Leipzig, Juli 26, 2010

2010 Marco Büchler: *Graph Mining in the Humanities - Viewing on Semantic Spaces* auf der Konferenz des *Hestia*-Projektes, Oxford, UK, Juli 03, 2010.

2010 Marco Büchler (*Lecture*): *An Introduction to eAQUA* beim geisteswissenschaftlichen Exzellenzcluster *Asien und Europa im globalen Kontext: Die Dynamik der Transkulturalität* an der *Ruprecht-Karls-Universität Heidelberg*, Heidelberg, Mai 26, 2010.

2009 Marco Büchler: *Discovering Latent Relations of Concepts by Graph Mining Approaches* auf dem *Chicago Colloquium on Digital Humanities and Computer Science*, Chicago, IL, USA, November 15, 2009.

2009 Marco Büchler: *„Fragments“ of Natural Language Processing* bei *T-Systems Multimedia Solutions*, Dresden, September 29, 2009.

2009 Marco Büchler, Annette Geßner (damals Loos): *Textual Re-use of Ancient Greek Texts: A case study on Plato's works* im *Digital Classicist & Institute of Classical Studies Seminar 2012*, London, UK, Juni 26, 2009.

2008 Marco Büchler: *Leipzig Linguistic Services - A 4 Years Summary of Providing Linguistic Web Services* im Rahmen der Veranstaltung *Web services architecture in Clarin*, München, November 10/11, 2008.

2008 Marco Büchler: *Techniques from Natural Language Processing for Detection and Protection of Kids against Pornography and Racism in the Web* bei der *Cybits AG* im Rahmen von Gesprächen zum EU-Antrag *web4kids - Innocence is in Danger*, Frankfurt/M., Februar 18, 2008.

2007 Marco Büchler: *Automatische Sprachverarbeitung an der Universität Leipzig* beim *Apache UIMA Projekt* bei *IBM Deutschland Research & Development GmbH*, Böblingen, Oktober 15, 2007.

B.13 Poster und Posterdemonstrationen

03/2011 Stefan Jänicke, Ralf Stockmann, Marco Büchler and Gerik Scheuermann: *Europeana4D – Visualizing And Exploring Geospatial-Temporal Data.* At *The Connected Past – People, Networks and Complexity in Archaeology and History Symposium* hosted at 2012 Computer Applications and Quantitative Methods in Archaeology, Southampton, UK, March 2012.

12/2009 Marco Büchler, Annette Geßner, Gerhard Heyer, Chris Walther, Thomas Eckart: *Citation Detection and Visualisation by Service Oriented Components.* At 5th IEEE International Conference on e-Science, Oxford, UK, Dec. 2009.

10/2008 Marco Büchler, Thomas Eckart, Chris Walther, Reinhold Scholl: *Completing Ancient Greek Papyri*, 2008.

10/2008 Marco Büchler, Thomas Eckart, Chris Walther, Gerhard Heyer: *The Social Network of the Ancient World*, 2008.

10/2008 Marco Büchler, Gerhard Heyer, Chris Walther, Charlotte Schubert: *Interaction between Humanities and Natural Science*, 2008.

10/2008 Marco Büchler, Gerhard Heyer, Chris Walther, Charlotte Schubert: *Wechselwirkung zwischen den Geistes- und Naturwissenschaften*, 2008.

10/2008 Marco Büchler, Stefan Beyer, Chris Walther, Marcus Deufert: *The Metric of Plautus*, 2008.

10/2008 Marco Büchler, Thomas Eckart, Annette Geßner, Chris Walther: *The Reception of Plato's Text in the Ancient World*, 2008.

B.14 Lehrveranstaltungen

WS 2012 Seminar: *Leipzig eHumanities Seminar* an der Universität Leipzig.

WS 2012 Praxisseminar: *Software-gestütztes Arbeiten mit Historischen Texten - Text Mining in den Geisteswissenschaften* an der Martin-Luther Universität Halle/S., Germany.

SS 2011 Praxisseminar: *Software-gestütztes Arbeiten mit Historischen Texten - Text Mining in den Geisteswissenschaften* an der Martin-Luther Universität Halle/S., Germany.

WS 2010 Vorlesung: *Text Mining in den eHumanities* an der Martin-Luther Universität Halle/S., Germany.

SS 2010 Seminar: Betreuung von Studenten im Seminar *Anwendungen Linguistischer Informatik*: Themenbereich: *Spell Checking*

SS 2009 Seminar: Betreuung von Studenten im Seminar *Anwendungen Linguistischer Informatik*

WS 2007 Vorlesung: *Text Mining* (vertretungsweise), Universität Leipzig.

SS 2007 Praktikum: *Linguistische Webservices*

WS 2006 Vorlesung: *Text Mining* (vertretungsweise), Universität Leipzig.

B.15 Betreute Abschlussarbeiten

08/2013	Elmar Voigtländer: *Entwurf eines hochperformanten und XML-basierten Protokolls für den Distributed Text Re-use (Arbeitstitel)* Diplomarbeit. Einzureichen im dritten Quartal 2013.	[Voigtländer 2013]
04/2013	Markus Ackermann: *Construction of a Multi-Lingual Lexical Database for Ancient Languages from Machine-Readable Dictionaries (Arbeitstitel)*. Bachelorarbeit. Einzureichen im zweiten Quartal 2013.	[Ackermann 2013]
04/2013	Frederik Baumgardt: *Evaluation of semantic measures in text reconstruction from OCR data (Arbeitstitel)*. Masterarbeit. Einzureichen im zweiten Quartal 2013.	[Baumgardt 2013]
07/2011	Frederik Baumgardt: *Visualisierung von Kookkurrenzgraphen*. Bachelorarbeit, 2011.	[Baumgardt 2011]
06/2011	Daniel Müller: *Local Text Reuse Detection mittels Diskreter Kosinustransformation auf Grafik-Hardware* Diplomarbeit, 2011.	[Müller 2011]
05/2011	Maria Moritz: *Fragmentarische Autoren - Extraktion altgriechischer Eigennamen und Belegstellen auf antiken Texten*, Masterarbeit, 2011.	[Moritz 2011]
01/2011	David Stange: *Mental Maps: Aufbau von orts- und zeitabhängigen Bedeutungsräumen zum automatischen Erkennen politischer und gesellschaftlicher Zäsuren im historischen Kontext*, Masterarbeit, 2011.	[Stange 2011]
08/2010	David Pansch: *Datenintegration heterogener Quellen im Kontext der eHumanities*, Bachelorarbeit, 2010.	[Pansch 2010]
06/2010	Sebastian Sander: *Performanceanalyse von SOAP- und REST-basierten Services in einer Linguistic Resources Umgebung*, Diplomarbeit, 2010.	[Sander 2010]
09/2009	Sebastian Kruse: *Textvervollständigung auf antiken Texten*, Bachelorarbeit, 2009.	[Kruse 2009]
05/2009	Marcus Puchalla: *Termextraktion auf antiken Texten*, Bachelorarbeit, 2009.	[Puchalla 2009]
01/2009	Christine Voigtländer: *Morphologische Analyse von antiken Texten im sprach-evolutionären Wandel*, Diplomarbeit, 2009.	[Voigtländer 2009]
08/2008	Konstantin Sveds: *Language Independent Sentence Boundary Detection*, Diplomarbeit, 2008.	[Sveds 2008]

Literaturverzeichnis

[Ackermann 2013] Markus Ackermann. *Construction of a Mulit-Lingual Lexical Database for Ancient Languages from Machine-Readable Dictionaries (Arbeitstitel). Eingereicht im zweiten Quartal 2013*, 2013.

[Aggarwal 2010a] Charu C. Aggarwal and Haixun Wang. *An Introduction to Graph Data.* In Managing and Mining Graph Data [Aggarwal 2010b], pages 1–11.

[Aggarwal 2010b] Charu C. Aggarwal and Haixun Wang, editeurs. Managing and mining graph data, volume 40 of *Advances in Database Systems.* Springer, 2010.

[Allan 2002] J. Allan. *Topic Detection and Tracking: Event-Based Information Organization.* Kluwer International Series on Information Retrieval. Kluwer Academic Publishers, 2002.

[Allen 2011] G. Allen. Intertextuality. Routledge, 2011.

[Alsleben 2007] Brigitte Alsleben, Werner Schoze-Stubenrecht and Dudenredaktion. *Das grosse Buch der Zitate und Redewendungen.* Dudenverlag, Mannheim, Germany, 2007.

[An 2004] Yuan An, Jeannette Janssen and Evangelos E. Milios. *Characterizing and Mining the Citation Graph of the Computer Science Literature.* Knowl. Inf. Syst., vol. 6, no. 6, pages 664–678, November 2004.

[Apfel 2010] Willi Apfel. *Die schönsten Zitate und Weisheiten der Welt.* Nikol Verlagsgesellschaft mbH, July 2010.

[Archambault 2004] Eric Archambault and Etienne Vignola Gagne. *Science Metrix (Final Report): The Use of Bibliometrics in the Social Sciences and Humanities*, 2004. Prepared for the Social Sciences and Humanities Research Council of Canada (SSHRCC), August, 2004.

[Asratian 1998] A.S. Asratian, T.M.J. Denley and R. Häggkvist. *Bipartite Graphs and their Applications.* Cambridge Tracts in Mathematics. Cambridge University Press, 1998.

[Babeu 2011] Alison Babeu. *"Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classicists.* Rapport technique, August 2011.

[Bamman 2008] David Bamman and Gregory Crane. *The logic and discovery of textual allusion.* In Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008), Marrakesh, 2008.

[Barabási 1999] Albert-László Barabási and Réka Albert. *Emergence of Scaling in Random Networks.* Science, vol. 286, no. 5439, pages 509–512, 1999.

[Barabási 2003] Albert-László Barabási. *Linked - how Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life.* Plume, 2003.

[Barabási 2012] Albert-László Barabási, Chaoming Song and Dashun Wang. *Publishing: Handful of papers dominates citation.* Nature, vol. 491, no. 7422, page 40, November 2012.

[Baroni 2004] Marco Baroni and Stefano Vegnaduzzo. *Identifying Subjective Adjectives through Web-based Mutual Information.* In Proceedings of the 7th Konferenz zur Verarbeitung Natürlicher Sprache (German Conference on Natural Language Processing – KONVENS'04, pages 613–619, 2004.

[Barrón-Cedeño 2010a] Alberto Barrón-Cedeño. *On the mono- and cross-language detection of text reuse and plagiarism.* In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pages 914–914, New York, NY, USA, 2010. ACM.

[Barrón-Cedeño 2010b] Alberto Barrón-Cedeño, Chiara Basile, Mirko Degli Esposti and Paolo Rosso. *Word length n-grams for text re-use detection.* In Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing, CICLing'10, pages 687–699, Berlin, Heidelberg, 2010. Springer-Verlag.

[Basile 2009] Chiara Basile, Dip Matematica, Dario Benedetto, Emanuele Caglioti, Giampaolo Cristadoro and Mirko Degli Esposti. *Caglioti E.: A plagiarism detection procedure in three steps: selection, matches and 'squares.* In SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09, pages 1–9, 2009.

[Baumgardt 2011] Frederik Baumgardt. *Visualisierung von Kookkurrenzgraphen*, 2011.

[Baumgardt 2013] Frederik Baumgardt. *Evaluation of semantic measures in text reconstruction from OCR data*, 2013.

[Beaulieu 2012] Marie-Claire Beaulieu and Bridget Almas. *Digital Humanities in the Classroom: Introducing a New Editing Platform for Source Documents in Classics.* In Digital Humanities 2012, June 2012.

[Believer's Resource 2011] Believer's Resource. *XML encoded versions of several English language Bible translations*, 2011. URL: `http://www.believersresource.com/categories/bible-raw-data.html` last accessed Nov. 11th, 2011.

[Bendersky 2009] Michael Bendersky and W. Bruce Croft. *Finding text reuse on the web.* In Proceedings of the Second ACM International Conference on Web Search and Data Mining, WSDM '09, pages 262–271, New York, NY, USA, 2009. ACM.

[Berti 2009] Monica Berti, Matteo Romanello, Alison Babeu and Gregory Crane. *Collecting fragmentary authors in a digital library.* In Fred Heath, Mary Lynn, Rice-Lively, Richard Furuta: Proceedings of the 2009 Joint International Conference on Digital Libraries, JCDL 2009, Austin, TX, USA, June 15-19, 2009, pages 259–262, 2009.

[Berti 2012] Monica Berti. *Fragmentary Texts - Collecting and representing fragments of lost authors and works*, 2012. URL: <http://www.fragmentarytexts.org/> last accessed Sept. 17th, 2012.

[Biemann 2007a] C. Biemann, G. Heyer, U. Quasthoff and M. Richter. *The Leipzig Corpora Collection - Monolingual corpora of standard size.* In Proceedings of Corpus Linguistic 2007, Birmingham, UK, 2007.

[Biemann 2007b] S. Biemann. *Unsupervised and Knowledge-Free Natural Language Processing in the Structure Discovery Paradigm.* PhD thesis, Universität Leipzig, 2007.

[Bigwood 1983] J. M. Bigwood. *The Ancient Accounts of the Battle of Cunaxa.* The American Journal of Philology, vol. 104, no. 4, pages pp. 340–357, 1983.

[BIMA 2012] BIMA. *Biometrics Glossary - Version 6.0.* Rapport technique, Biometrics Identity Management Agency in conjunction with the United States Department of Defense, April 2012.

[Bird 2009] S. Bird, E. Klein and E. Loper. *Natural Language Processing with Python.* Oreilly Series. O'Reilly Media, Incorporated, 2009.

[Blei 2006] David M. Blei and John D. Lafferty. *Dynamic topic models.* In Proceedings of the 23rd international conference on Machine learning, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[Bloom 1973] D. M. Bloom and W. Knight. *A birthday problem.* American Mathematical Monthly, vol. 80, no. 10, pages 1141–1142, December 1973.

[Bobenhausen 2009] Klemens Bobenhausen and Günter Geh. *Automatisches metrisches Markup deutschsprachiger Gedichte.* Jahrbuch für Computerphilologie 9. mentis Verlag, Paderborn 2009, S. 61-85., 2009.

[Bocek 2007] T. Bocek, E. Hunt and B. Stiller. *Fast Similarity Search in Large Dictionaries.* Rapport technique ifi-2007.02, Department of Informatics, University of Zurich, April 2007. http://fastss.csg.uzh.ch/.

[Bolelli 2006] Levent Bolelli, Seyda Ertekin and C. Lee Giles. *Clustering scientific literature using sparse citation graph analysis.* In PKDD, pages 30–41, 2006.

[Bordag 2007] S. Bordag. *Elements of Knowledge-free and Unsupervised Lexical Acquisition.* PhD thesis, Universität Leipzig, 2007.

[Bordag 2008] Stefan Bordag. *Cooccaccess - Fast Access to Cooccurrence Data in Java*, 2008. URL: `http://wortschatz.uni-leipzig.de/~sbordag/cooccaccess/index.html` last accessed Jul. 21th, 2010.

[Boschetti 2009] Federico Boschetti, Matteo Romanello, Alison Babeu, David Bamman and Gregory Crane. *Improving OCR Accuracy for Classical Critical Editions.* In Maristella Agosti, José Luis Borbinha, Sarantos Kapidakis, Christos Papatheodorou and Giannis Tsakonas, editeurs, ECDL, volume 5714 of *Lecture Notes in Computer Science*, pages 156–167. Springer, 2009.

[Boyer 1976] R.S. Boyer and J.S. Moore. *A Fast String Searching Algorithm.* AD/A-022. Defense Technical Information Center, 1976.

[Boyer 1977] Robert S. Boyer and J. Strother Moore. *A fast string searching algorithm.* Commun. ACM, vol. 20, no. 10, pages 762–772, October 1977.

[Brin 1998] Sergey Brin and Lawrence Page. *The anatomy of a large-scale hypertextual Web search engine.* Comput. Netw. ISDN Syst., vol. 30, no. 1-7, pages 107–117, April 1998.

[Broder 1997a] Andrei Z. Broder. *On the Resemblance and Containment of Documents.* In In Compression and Complexity of Sequences (SEQUENCES'97, pages 21–29. IEEE Computer Society, 1997.

[Broder 1997b] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse and Geoffrey Zweig. *Syntactic clustering of the Web.* Comput. Netw. ISDN Syst., vol. 29, no. 8-13, pages 1157–1166, 1997.

[Broder 1998] Andrei Z. Broder, Moses Charikar, Alan M. Frieze and Michael Mitzenmacher. *Min-wise Independent Permutations.* Journal of Computer and System Sciences, vol. 60, pages 327–336, 1998.

[Büchler 2005] Marco Büchler. *Analyse der XML-Performanz für Suchanfragen.* Seminararbeit im Rahmen der Veranstaltung Angewandte Linguistische Informatik, 2005.

[Büchler 2006a] Marco Büchler. *Flexibles Berechnen von Kookkurrenzen auf strukturierten und unstrukturierten Daten*, 2006.

[Büchler 2006b] Marco Büchler. *Medusa Release Homepage*, 2006. URL: http://mbuechler.e-humanities.net/medusa/ last accessed Feb. 14th, 2010.

[Büchler 2007] Marco Büchler and Gerhard Heyer. *Kookurenzberechnungen mit UIMA und Medusa.* In GLDV, editeur, UIMA Workshop at the GLDV 2007 in Tübingen/Germany, 2007.

[Büchler 2008a] Marco Büchler. *Medusa: Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung*. VDM Verlag Dr. Müller, 2008.

[Büchler 2008b] Marco Büchler and Gerhard Heyer. *Text Mining for Classical Studies*, 2008. URL: http://www.eaqua.net/e3.php last accessed Aug. 12th, 2010.

[Büchler 2008c] Marco Büchler, Gerhard Heyer and Sabine Gründer. *Bringing Modern Text Mining Approaches to Two Thousand Years Old Ancient Texts.* In e-Humanities – an emerging discipline: Workshop in the 4th IEEE International Conference on e-Science, 2008.

[Büchler 2009a] Marco Büchler and Gerhard Heyer. *Leipzig Linguistic Services – A 4 Years Summary of Providing Linguistic Web Services.* In Gerhard Heyer (Editor): Text Mining Services – Building and applying text mining service infrastructures in research and industry. Proceedings of the Conference on Text Mining Services 2009. Leipziger Beiträge zur Informatik: Band XIV Leipzig, Germany, 2009.

[Büchler 2009b] Marco Büchler, Lutz Maicher, Frederik Baumgardt and Benjamin Bock. *Automatic Extraction of Topic Maps based Argumentation Trails.* In Gerhard Heyer (Editor): Text Mining Services – Building and applying text mining service infrastructures in research and industry. Proceedings of the Conference on Text Mining Services 2009. Leipziger Beiträge zur Informatik: Band XIV Leipzig, Germany, 2009, 2009.

[Büchler 2010a] Marco Büchler. *Bericht zur 2010 Digital Humanities Conference in London für das Bundesministerium für Bildung und Forschung*, 2010.

[Büchler 2010b] Marco Büchler. *Reisebericht eines Informatikers durch die nationalen und internationalen Digital Humanities und eHumanities - April 2008 bis August 2010 für das Bundesministerium für Bildung und Forschung*, 2010.

[Büchler 2010c] Marco Büchler, Frederik Baumgardt and Thomas Eckart. *Von Platon zu Alexander dem Großen – Automatische Extraktion von Topic-Maps-basierten Assoziationsketten aus Sozialen Netzwerken der Antike.* In Detlef Reineke (Hrsg. eDITion): Terminologie und Text Mining. Ausgabe 1/2010. ISSN 1862-023X, Las Palmas de Gran Canaria, 2010.

[Büchler 2010d] Marco Büchler, Annette Geßner and Thomas Eckart. *Unsupervised Detection and Visualisation of Textual Reuse on Ancient Greek Texts.* In Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science, 2010.

[Büchler 2010e] Marco Büchler, Annette Geßner, Gerhard Heyer and Thomas Eckart. *Detection of Citations and Text Reuse on Ancient Greek Texts and its Applications in the Classical Studies: eAQUA Project.* In Digital Humanities 2010 – Conference Abstracts, King's College London, London, UK, 2010.

[Büchler 2010f] Marco Büchler and Gerhard Heyer. *Salton and Wittgenstein in the Humanities: About Semantics in Philosophical Texts.* In eHumanities Workhop on Informatik 2010 Conference. In Proceeding to 44. Jahrestag der Gesellschaft für Informatik e.V.: Service Science - Neue Perspektiven für die Informatik. Springer-Verlag. Leipzig, 2010.

[Büchler 2011a] Marco Büchler. *Thinking in Signals: Measuring Text Re-use with Signal Processing Algorithms*, 2011.

[Büchler 2011b] Marco Büchler, Stefan Beyer, Thomas Eckart and Ute Pietruschka. *Collecting Ancient Greek, Arabic, and Syriac Sapiential Statements - The Gnomology Database of Arabic and Syriac*, pages 283–295. Press Universitaires de Nancy, Marie-Christine Bornes-Varol and Marie-Sol Ortola: 2nd International Colloquium Aliento - Corpus anciens et Bases de donnees édition, Apr 2011.

[Büchler 2011c] Marco Büchler, Philip R. Burns, Gregory Crane, Martin Mueller and Gerhard Heyer. *One Step Closer To Paraphrase Detection On Historical Texts: About The Quality of Text Re-use Techniques and the Ability to Learn Paradigmatic Relations.* In Proceedings of the 2011 Chicago Colloquium on Digital Humanities and Computer Science. Chicago, 2012, 2011.

[Büchler 2012a] Marco Büchler. *ACID for the eHumanities - A new paradigm for successful eHumanities projecs.* In Presentation at CLARIN M12 Workshop, Leipzig, Germany, June 27-28, 2012, 2012.

[Büchler 2012b] Marco Büchler, Gregory Crane and Gerhard Heyer. *Historical Relevance Feedback Detection by Text Re-use Mining.* In Maximilian Schich, Roger Malina, Isabel Meirelles, Christian Huepe: Arts, Humanities, and Complex Networks Living Companion at Arts, Humanities, and Complex Networks — 3rd Leonardo satellite symposium hosted by NetSci2012, Evanston, IL, USA, 06 2012.

[Büchler 2012c] Marco Büchler, Gregory Crane, Maria Moritz and Alison Babeu. *Increasing Recall for Text Re-use in Historical Documents to Support Research in the Humanities.* In George Buchanan, Edie Rasmussen and Fernando Loizides, editeurs, Theory and Practice of Digital Libraries 2012, 09 2012.

[Büchler 2012d] Marco Büchler, Sebastian Kruse and Thomas Eckart. *Bringing Modern Spell Checking Approaches to Ancient Texts - Automated Suggestions for Incomplete Words.* In Proceedings of Digital Humanities 2012, Hamburg, Germany, July 2012.

[Büchler 2012e] Marco Büchler, Ute Pietruschka and Norman Wetzig. *Transmission of Greek Gnomologia in Syriac and Arabic: The Corpus of Arabic and Syriac Gnomologia.* In Claudine Moulin and Natalia Filatkina, editeurs, International Workshop "Exploring Formulaic Knowledge through Languages, Cultures and Time", Trier, March 2012.

[Büchler 2013a] Marco Büchler. *TRACER: A Java based software library for detecting historical text re-use.* Rapport technique, Leipzig eHumanities Research Group, University of Leipzig, Germany, April 2013. http://etraces.e-humanities.net/TRACER.

[Büchler 2013b] Marco Büchler and Gregory Crane. *Uncovering Serendipity from Historical Data – About Usage of Network Analysis in Humanities.* In The Connected Past: People, Networks and Complexity in Archaeology and History, Oxford, UK, 04 2013. Oxford University Press.

[Büchler 2013c] Marco Büchler, Annette Geßner and Monica Berti. *Measuring the Influence of a Work by Text Reuse and Knowledge Transfer Approaches.* In Ed. Stuart Dunn and Simon Mahony: Digital Classicist Supplement: Bulletin of the Institute of Classical Studies, London. Wiley-Blackwell, 2013.

[Büchler 2013d] Marco Büchler and Maria Moritz. *Historical Bibliometrics: Historical Relevance Feedback Detection by Text Re-use Mining.* March 2013.

[Büchmann 2007] G. Büchmann and W. Hofmann. *Der Neue Büchmann: Geflügelte Worte: Der klassische Zitatenschatz.* Ullstein-Bücher, Allgemeine Reihe. Ullstein, 2007.

[Buckley 2000] Chris Buckley and Ellen M. Voorhees. *Evaluating evaluation measure stability.* In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '00, pages 33–40, New York, NY, USA, 2000. ACM.

[Buckwalter 2004a] Tim Buckwalter. *Issues in Arabic orthography and morphology analysis.* In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Semitic '04, pages 31–34, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[Buckwalter 2004b] Timothy Buckwalter. *Buckwalter Arabic Morphological Analyzer Version 2.0.* Rapport technique LDC2004L02, Linguistic Data Consortium, 2004.

[Burns 2012] Philip Burns. *MorphAdorner*, 2012. URL: http://morphadorner.northwestern.edu/ last accessed Nov. 1st, 2012.

[Buss 2008] Keno Buss and A. Transliteracy. *Table Of Contents Keno Buss Table Of Contents Appendix 1 A Literature Review on Preprocessing for Text Mining 1*, 2008.

[Büttcher 2007] Stefan Büttcher, Charles L. A. Clarke, Peter C. K. Yeung and Ian Soboroff. *Reliable information retrieval evaluation with incomplete and biased judgements.* In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, pages 63–70, New York, NY, USA, 2007. ACM.

[Cayless 2010] Hugh Cayless. *Ktêma es aiei: Digital Permanence from an Ancient Perspective.* In Gabriel Bodard and Simon Mahony (eds.), Digital Research in the Study of Classical Antiquity, pages 139–150, Prague, Czech Republic, June 2010. Association for Computational Linguistics.

[Chakrabarti 2010] Deepayan Chakrabarti, Christos Faloutsos and Mary McGlohon. *Graph Mining: Laws and Generators.* In Aggarwal & Wang [Aggarwal 2010b], pages 69–123.

[Charikar 2002] Moses Charikar. *Similarity estimation techniques from rounding algorithms.* In John H. Reif, editeur, STOC, pages 380–388. ACM, 2002.

[Cheesman 2012] Tom Cheesman. *Delighted Beauty: Version Variation Visualisation*, 2012. URL: http://www.delightedbeauty.org/ last accessed Aug. 29th, 2012.

[Church 1989] Kenneth Ward Church and Patrick Hanks. *Word Association Norms, Mutual Information and Lexicography.* In Julia Hirschberg, editeur, ACL, pages 76–83. ACL, 1989.

[Church 1990] Kenneth Ward Church and Patrick Hanks. *Word association norms, mutual information, and lexicography.* Comput. Linguist., vol. 16, no. 1, pages 22–29, March 1990.

[Clausi 2002] D. A. Clausi. *An analysis of co-occurrence texture statistics as a function of grey level quantization.* Canadian Journal of Remote Sensing, vol. 28, no. 1, page 45–62, 2002.

[Clough 2002] Paul Clough, Robert Gaizauskas, Scott S. L. Piao and Yorick Wilks. *METER: MEasuring TExt Reuse.* In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 152–159, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[Coffee 2012a] Neil Coffee. *Tesserae - Intertextual Phrase Matching*, 2012. URL: http://tesserae.caset.buffalo.edu/ last accessed Aug. 29th, 2012.

[Coffee 2012b] Neil Coffee, Jean-Pierre Koenig, Shakthi Poornima, Christopher W. Forstall, Roelant Ossewaarde and Sarah L. Jacobson. *The Tesserae Project: intertextual analysis of Latin poetry.* Literary and Linguistic Computing, 2012.

[Cormen 2001] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest and Charles E. Leiserson. *Introduction to Algorithms.* McGraw-Hill Higher Education, 2nd édition, 2001.

[Cowie 1998] A.P. Cowie. *Phraseology: Theory, Analysis, and Applications.* Oxford linguistics. Oxford University Press, USA, 1998.

[Crane 1985] Gregory Crane. *The Perseus Project.* World Wide Web electronic publication, 1985. http://www.perseus.tufts.edu/hopper/.

[Crane 1991] Gregory Crane. *Generating and Parsing Classical Greek.* Literary and Linguistic Computing, vol. 6, no. 4, pages 243–245, 1991.

[Crane 2006] Gregory Crane. *What Do You Do with a Million Books?* D-Lib Magazine, vol. 12, no. 3, March 2006.

[Crane 2012] Gregory Crane, Bridget Almas, Alison Babeu, Lisa Cerrato, Matthew Harrington, David Bamman and Harry Diakoff. *Student researchers, citizen scholars and the trillion word library.* In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '12, pages 213–222, New York, NY, USA, 2012. ACM.

[Crochemore 1997] Maxime Crochemore, Renaud Vérin, Renaud V Erin and F-Noisy le gr. *Direct construction of Compact Directed Acyclic Word Graphs.* In Combinatorial Pattern Matching, pages 116–129. Springer-Verlag, 1997.

[Crochemore 2003] M. Crochemore and W. Rytter. *Jewels of Stringology: Text Algorithms.* World Scientific, 2003.

[Damerau 1964] Fred J. Damerau. *A technique for computer detection and correction of spelling errors.* Commun. ACM, vol. 7, no. 3, pages 171–176, March 1964.

[Davis 2012] Mark Davis and Ken Whistler. *Unicode Standard Annex no. 15 - Unicode Normalization Forms (Unicode 6.2.0)*, 2012. URL: http://unicode.org/reports/tr15/ last accessed Nov. 1st, 2012.

[Dawkins 1976] Richard Dawkins. *The selfish gene / Richard Dawkins.* Oxford University Press, New York :, 1976.

[De Saussure 2001] F. De Saussure, C. Bally and A. Sechehaye. *Grundfragen Der Allgemeinen Sprachwissenschaft.* De Gruyter Studienbuch. Walter de Gruyter, 2001.

[Deufert 2010] Marcus Deufert, Judith Blumenstein, Andreas Trebesius, Stefan Beyer and Marco Büchler. *Objective Detection of Plautus' Rules by Computer Support.* In Digital Humanities 2010 – Conference Abstracts, King's College London, London, UK, 2010.

[Dias 2005] Gaël Dias and Špela Vintar. *Unsupervised learning of multiword units from part-of-speech tagged corpora: does quantity mean quality?* In Proceedings of the 12th Portuguese conference on Progress in Artificial Intelligence, EPIA'05, pages 669–679, Berlin, Heidelberg, 2005. Springer-Verlag.

[Dover 1997] Kenneth Dover. *Einleitung in die griechische Philologie*, chapitre Textkritik. B. G. Teubner, Stuttgard und Leipzig, 1997.

[Dué 2009] Casey Dué and Mary Ebbott. *Digital Criticism: Editorial Standards for the Homer Multitext.* Digital Humanities Quarterly, vol. 3, no. 1, January 2009.

[Dunning 1993] T.E. Dunning. *Accurate Methods for the Statistics of Surprise and Coincidence.* Computational Linguistics, vol. 19, no. 1, pages 61–74, 1993.

[Ebbinghaus 1885] H. Ebbinghaus. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie.* Duncker & Humblot, 1885.

[Eckart 2010] Thomas Eckart, Reinhard Förtsch, Sebastian Kruse and Marco Büchler. *Accessing, Visualizing and Annotating Geographical Information in Archeology.* vol. Proceedings of Computer Applications and Quantitative Methods in Archeology CAA 2010, 2010.

[Eckart 2011] Thomas Eckart, David Pansch and Marco Büchler. *Integration of Distributed Text Resources by Using Schema Matching Techniques.* In Proceedings of Digital Humanities 2011, Stanford, USA, July 2011.

[Efthimiadis 1996] Efthimis N. Efthimiadis. *Query expansion.* Annual Review of Information Systems and Technology (ARIST), vol. 31, pages 121–187, 1996.

[Erdős 1959] P. Erdős and A. Renyi. *On Random Graphs I.* Publ. Math. Debrecen, vol. 6, page 290, 1959.

[Erdős 1960] P. Erdős and A Rényi. *On the Evolution of Random Graphs.* In PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES, pages 17–61, 1960.

[Ernst-Gerlach 2008] Andrea Ernst-Gerlach and Gregory Crane. *Identifying Quotations in Reference Works and Primary Materials.* volume 5173 of *Lecture Notes in Computer Science*, pages 78–87. Springer, 2008.

[Evert 2004] Stefan Evert. *The Statistics of Word Cooccurrences: Word Pairsand Collocations.* PhD thesis, Universität Stuttgart, 2004.

[Fähnrich 2010] Klaus-Peter Fähnrich and Bogdan Franczyk, editeurs. *Informatik 2010: Service Science - Neue Perspektiven für die Informatik, Beiträge der 40. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Band 2, 27.09. - 1.10.2010, Leipzig*, volume 176 of *LNI.* GI, 2010.

[Faloutsos 2007] Christos Faloutsos and Vasileios Megalooikonomou. *On data mining, compression, and Kolmogorov complexity.* Data Min. Knowl. Discov., vol. 15, no. 1, pages 3–20, August 2007.

[Fellbaum 1998] Christiane Fellbaum, editeur. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.

[Finnegan 2011] R.H. Finnegan. *Why Do We Quote?: The Culture and History of Quotation.* Open Book Publishers, 2011.

[Fleiss 2004] Joseph L. Fleiss, Bruce Levin and Myunghee C. Paik. *The Measurement of Interrater Agreement.* pages 598–626, 2004.

[Fortnow 2001] Lance Fortnow. *Kolmogorov complexity*, pages 73–86. Berlin: de Gruyter, 2001.

[Foss 2006] Janae N. Foss and Nilufer Onder. *A hill-climbing approach for planning with temporal uncertainty.* Rapport technique, In FLAIRS 2006 Conference, 2006.

[Fucks 1968] Wilhelm Fucks. *Nach allen Regeln der Kunst.* dva, 1968.

[Förtsch 2010] Reinhard Förtsch. *The Syntax of Contextualization.* Miriam S. Balmuth Lectures Series: Classical Culture as Digital Information, Languages of Materiality. Cabot Intercultural Center, ASEAN Auditorium, Tufts University, Boston, USA, 2010.

[Gaizauskas 2001] Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough and Scott Piao. *The meter corpus: A corpus for analysing journalistic text reuse.* pages 214–223, 2001.

[Geßner 2012] Annette Geßner. *GERTRUDE: Göttingen E-Research Text Re-Use Digital Edition - A tool to support detecting textual congruences and creating a digital edition of a text online*, 2012. Poster demonstration during Digital Humanities Deutschland Unconference at 2012 Digital Humanities Conference, 2012.

[Geßner 2013] Annette Geßner. *The Tool GERTRUDE - World literature, intertextuality and crowdsourcing*, 2013. Slides of a talk given at CeRch-seminar, January 29, 2013, King's College London, London, GB.

[Gionis 1999] Aristides Gionis, Piotr Indyk and Rajeev Motwani. *Similarity Search in High Dimensions via Hashing.* In Proceedings of the 25th International Conference on Very Large Data Bases, VLDB '99, pages 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[Goldhahn 2012] Dirk Goldhahn, Thomas Eckart and Uwe Quasthoff. *Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages.* In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012.

[Gradmann 2012] Stefan Gradmann, Jonathan Gray and Christian Morbidoni. *Beyond infrastructure! Further Modelling the Scholarly Research and Collaboration Domain*, 2012. Talk at 2012 Leipzig eHumanities Seminar, Leipzig, Germany, 2012.

[Granovetter 1983] Mark Granovetter. *The Strength of Weak Ties: A Network Theory Revisited.* Sociological Theory, vol. 1, pages 201–233, 1983.

[Guyon 2003] Isabelle Guyon and André Elisseeff. *An introduction to variable and feature selection.* J. Mach. Learn. Res., vol. 3, pages 1157–1182, March 2003.

[Harman 1995] Donna Harman. *Overview of the second text retrieval conference (TREC-2).* Inf. Process. Manage., vol. 31, no. 3, pages 271–289, May 1995.

[Havemann 2009] Frank Havemann. *Einführung in die Bibliometrie*. Philosophische Fakultät I, 2009.

[Hebb 1949] Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, new edition édition, June 1949.

[Henle 2001] H. Henle. *Das Tonstudio Handbuch: Praktische Einführung in die professionelle Aufnahmetechnik. Grundlagen der Akustik. Analoge und digitale Audiotechnik. Auswahlkriterien und Einsatz von Mikrofonen, Mischpulten, Effektgeräten. Analoge und digitale Bandaufzeichnung. Harddisk-Recording. Lautsprecher und Regieraum-Design.* Factfinder-Serie. GC Carstensen Verlag, 2001.

[Heyer 2006] G. Heyer, U. Quasthoff and T. Wittig. *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. IT lernen. W3L-Verlag, 2006.

[Heyer 2008] Gerhard Heyer and Charlotte Schubert. *eAQUA: Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft*, 2008. URL: http://www.eaqua.net last accessed Aug. 29th, 2012.

[Heyer 2009] Gerhard Heyer. *Analyse von Bedeutungsveränderungen in diachronen Textkorpora.* Rapport technique, Natural Language Processing Group, University of Leipzig, Germany, Februar 2009. Vortrag im Forschungsseminar, Leipzig, Germany, February, 2009.

[Heyer 2010] Gerhard Heyer and Marco Büchler. *Some Challenges Posed to Computer Science by the eHumanities.* In Fähnrich & Franczyk [Fähnrich 2010], pages 524–529.

[Heyer 2011a] Gerhard Heyer. *eTRACES: Recherche und Analyse von Zitationsspuren und Wissenstransfer in sozialwissenschaftlichen Texten und deutschsprachiger Literatur.* Rapport technique, Leipzig eHumanities Research Group, University of Leipzig, Germany, July 2011. http://etraces.e-humanities.net/.

[Heyer 2011b] Gerhard Heyer, Marco Büchler and Volker Boehlke. *Aspects of an Infrastructure for eHumanities.* In Supporting Digital Humanities 2011, Nov 2011.

[Heyer 2011c] Gerhard Heyer, Marco Büchler, Thomas Eckart and Maria Moritz. *eAQUA - Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaften: Technologien und Ansätze zu Infrastruktur, Text Mining und Knowledge Transfer*. Leipzig University, Leipzig, Sept. 2011.

[Higgins 2000] Des Higgins and Willie Taylor, editeurs. *Bioinformatics: Sequence, Structure, and Databanks: A Practical Approach*. Oxford University Press, Inc., New York, NY, USA, 1st édition, 2000.

[Hirsch 2005] J. E. Hirsch. *An index to quantify an individual's scientific research output.* Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 46, pages 16569–16572, 2005.

[Hirschmann 2012] Hagen Hirschmann, Anke Lüdeling and Amir Zeldes. *Measuring and coding language change: An evolving study in a multilayer corpus architecture.* J. Comput. Cult. Herit., vol. 5, no. 1, pages 4:1–4:16, April 2012.

[Hofmann 1999] Thomas Hofmann. *Probabilistic Latent Semantic Analysis.* In Kathryn B. Laskey and Henri Prade, editeurs, UAI, pages 289–296. Morgan Kaufmann, 1999.

[Hose 2004] Ron Hose. *CS490 Final Report: Investigation of Sentence Level Text Reuse Algorithms.* 2004.

[Hsu 2012] Wen-Chin Hsu, Chan-Cheng Liu, Fu Chang and Su-Shing Chen. *Cancer classification: Mutual information, target network and strategies of therapy.* Journal of Clinical Bioinformatics, vol. 2, no. 1, pages 16+, 2012.

[Huffman 1952] David Huffman. *A Method for the Construction of Minimum-Redundancy Codes.* Proceedings of the IRE, vol. 40, no. 9, pages 1098–1101, September 1952.

[Huston 2011] Samuel Huston, Alistair Moffat and W. Bruce Croft. *Efficient indexing of repeated n-grams.* In Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, pages 127–136, New York, NY, USA, 2011. ACM.

[IBM 2012] IBM. *International Components for Unicode*, 2012. URL: `http://www-01.ibm.com/software/globalization/icu/` last accessed Aug. 12th, 2012.

[Internation Standard Organisation ISO/TC 37 SC 4 2010] Internation Standard Organisation ISO/TC 37 SC 4. *Language Resource Management: Word Segmentation of Written Texts for Mono-lingual and Multi-lingual Information processing. Part 1: Basic Concepts and General Principles*. International Standard. International Standard Organization, 2010.

[Jain 2005] Anil K. Jain, Ruud M. Bolle and Sharath Pankanti. *Biometrics: Personal Identification in Networked Society*. Springer, October 2005.

[Jalil 2010] Zunera Jalil, Anwar M. Mirza and Maria Sabir. *Content based Zero-Watermarking Algorithm for Authentication of Text Documents.* CoRR, vol. abs/1003.1796, 2010.

[Jin 2002] Contact Liang Jin, Liang Jin, Liang Jin, Chen Li, Chen Li, Sharad Mehrotra and Sharad Mehrotra. *Efficient Similarity String Joins in Large Data Sets.* Rapport technique, 2002.

[Jolliffe 2002] I. T. Jolliffe. *Principal Component Analysis*. Springer, second édition, October 2002.

[Jordanous 2012] Anna Jordanous, K. Faith Lawrence, Mark Hedges and Charlotte Tupman. *Exploring manuscripts: sharing ancient wisdoms across the semantic web.* In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, WIMS '12, pages 44:1–44:12, New York, NY, USA, 2012. ACM.

[Jurish 2012] Bryan Jurish. *Finite-State Canonicalization Techniques for Historical German.* PhD thesis, Universität Potsdam, 2012.

[Keim 2002] Daniel A. Keim. *Information Visualization and Visual Data Mining.* IEEE Transactions on Visualization and Computer Graphics, vol. 8, pages 1–8, 2002.

[Kelih 2005] Emmerich Kelih and Peter Grzybek. *Satzlänge: Definitionen, Häufigkeiten, Modelle (Am Beispiel slowenischer Prosatexte).* LDV Forum, vol. 20, no. 2, pages 31–51, 2005.

[Kenne 1996] P. E. Kenne and Mary O'Kane. *Topic change and local perplexity in spoken legal dialogue.* In ICSLP. ISCA, 1996.

[Knuth 1997a] Donald E. Knuth. *Art of Computer Programming, Volume 1: Fundamental Algorithms (3rd Edition).* Addison-Wesley Professional, 3 édition, July 1997.

[Knuth 1997b] Donald E. Knuth. *Art of Computer Programming, Volume 2 (3rd ed.): Seminumerical Algorithms*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1997.

[Kolmogorov 1963] A. N. Kolmogorov. *On Tables of Random Numbers.* Sankhya: The Indian Journal of Statistics, vol. 25, no. 2, page 369–375, 1963.

[Kolmogorov 1998] A. N. Kolmogorov. *On Tables of Random Numbers (Reprinted from Sankhya: The Indian Journal of Statistics, Series A, Vol. 25 Part 4, 1963).* Theor. Comput. Sci., vol. 207, no. 2, pages 387–395, 1998.

[Kruse 2009] Sebastian Kruse. *Textvervollständigung auf antiken Texten*, 2009.

[Kukich 1992] Karen Kukich. *Techniques for automatically correcting words in text.* ACM Comput. Surv., vol. 24, no. 4, pages 377–439, December 1992.

[Kullback 1951] S. Kullback and R. A. Leibler. *On information and sufficiency.* Annals of Mathematical Statistics, vol. 22, pages 49–86, 1951.

[Kumar 2004] Sudhir Kumar, Koichiro Tamura and Masatoshi Nei. *MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.* Briefings in Bioinformatics, vol. 5, no. 2, pages 150–163, 2004.

[Küper 2011] Christoph Küper. *Current Trends in Metrical Analysis*. Littera. Studies in Language and Literature/Studien Zur Sprache Und Literatur. Peter Lang, 2011.

[Küpfmüller 1974] Karl Küpfmüller. *Nachrichtenverarbeitung im Menschen*, pages 429–455. Springer, 1974.

[L'Ecuyer 1997] Pierre L'Ecuyer, Aaldert Compagner and Jean-François Cordeau. *Entropy Tests for Random Number Generators*, 1997.

[Lee 2005] Michael D. Lee and Matthew Welsh. *An empirical evaluation of models of text document similarity.* In CogSci2005, pages 1254–1259. Erlbaum, 2005.

[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, pages 472–479, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[Lehmer 1951] D. H. Lehmer. *Mathematical methods in large-scale computing units.* In Proceedings of a Second Symposium on Large-Scale Digital Calculating Machinery 1949, page 141–146, Cambridge, Mass., 1951. Harvard University Press.

[Levenshtein 1966] Vladimir I. Levenshtein. *Binary codes capable of correcting deletions, insertions, and reversals.* Rapport technique 8, 1966.

[Li 1989] M. Li, Netherlands). Centrum voor Wiskunde en Informatica (Amsterdam and P.M.B. Vitányi. *Kolmogorov Complexity and Its Applications*. Report. Centre for Mathematics and Computer Science, 1989.

[Li 2008] Ming Li and Paul M.B. Vitnyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Publishing Company, Incorporated, 3 édition, 2008.

[Lotka 1926] Alfred J. Lotka. *The frequency distribution of scientific productivity.* Journal of the Washington Academy of Sciences, vol. 16, no. 12, pages 317–323, 1926.

[Maas 1960] P. Maas. *Textkritik*. Teubner, 1960.

[Maltoni 2009] Davide Maltoni, Dario Maio, Anil K. Jain and Salil Prabhakar. *Handbook of Fingerprint Recognition*. Springer Publishing Company, Incorporated, 2nd édition, 2009.

[Manning 1999] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Pprocessing*. MIT Press, Cambridge, MA, USA, 1999.

[Manning 2008] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[McCreight 1976] Edward M. McCreight. *A Space-Economical Suffix Tree Construction Algorithm.* J. ACM, vol. 23, no. 2, pages 262–272, April 1976.

[McRae-Spencer 2006] Duncan M. McRae-Spencer and Nigel R. Shadbolt. *Also by the same author: AKTiveAuthor, a citation graph approach to name disambiguation.* In Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, JCDL '06, pages 53–54, New York, NY, USA, 2006. ACM.

[Metzler 2005] Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat and Justin Zobel. *Similarity measures for tracking information flow.* In Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM '05, pages 517–524, New York, NY, USA, 2005. ACM.

[Miller 1956] George A. Miller. *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.* The Psychological Review, vol. 63, pages 81–97, 1956.

[Miller 1995] George A. Miller. *WordNet: a lexical database for English.* Commun. ACM, vol. 38, no. 11, pages 39–41, November 1995.

[Mimno 2012] David Mimno. *Computational historiography: Data mining in a century of classics journals.* J. Comput. Cult. Herit., vol. 5, no. 1, pages 3:1–3:19, April 2012.

[Moretti 2005] Franco Moretti. *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, London, 2005.

[Moritz 2011] Maria Moritz. *Fragmentarische Autoren - Extraktion altgriechischer Eigennamen und Belegstellen auf antiken Texten*, 2011.

[Moritz 2013] Maria Moritz. *Aufdecken und Analysieren von Text Re-use Diversity durch philologisches Crowd Sourcing*, 2013.

[Müller 2011] Daniel Müller. *Local Text Reuse Detection mittels Diskreter Kosinustransformation auf Grafik-Hardware*, 2011.

[Neudecker 2011] Clemens Neudecker, Sven Schlarb, Zeki Mustafa Dogan, Paolo Missier, Shoaib Sufi, Alan Williams and Katy Wolstencroft. *An experimental workflow development platform for historical document digitisation and analysis.* In Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, HIP '11, pages 161–168, New York, NY, USA, 2011. ACM.

[Newman 2005] M. E. J. Newman. *Power laws, Pareto distributions and Zipf's law.* Contemporary Physics, vol. 46, pages 323–351, December 2005.

[Ng 2005] Wilfred Ng and Ho Lam Lau. *Effective Approaches for Watermarking XML Data.* In Lizhu Zhou, Beng Chin Ooi and Xiaofeng Meng, editeurs, DASFAA, volume 3453 of *Lecture Notes in Computer Science*, pages 68–80. Springer, 2005.

[NSTC 2006] NSTC. *Biometrics Glossary.* Rapport technique, National Science and Technology Council, Committee on Technology, Committee on Homeland and National Security, Subcommitte on Biometrics, Sept 2006.

[NYUDL 2012] NYUDL. *papyri.info.* Rapport technique, NYU Digital Library Technology Services & the Institute for the Study of the Ancient World, August 2012.

[Olaru 2004] Vlad Olaru and Walter F. Tichy. *Request Distribution Aware Caching in Cluster-Based Web Servers.* In Proc. of the 3rd IEEE International Symposium on Network Computing and Applications (IEEE NCA04), August 2004.

[Olsen 2011] Mark Olsen, Russell Horton and Glenn Roe. *Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections.* Digital Studies / Le champ numérique, vol. 2, no. 1, 2011.

[Ottmann 1996] Thomas Ottmann and Peter Widmayer. *Algorithmen und Datenstrukturen, 3. Auflage.* Spektrum Lehrbuch. Spektrum, 1996.

[Pansch 2010] David Pansch. *Datenintegration heterogener Quellen im Kontext der eHumanities*, 2010.

[Pansch 2011] David Pansch. *Greek Letter Shaver V.1.1*, 2011. URL: `http://www.eaqua.net/~dpansch/` last accessed Aug. 12th, 2011.

[Paulevé 2010] Loïc Paulevé, Hervé Jégou and Laurent Amsaleg. *Locality sensitive hashing: a comparison of hash function types and querying mechanisms.* Pattern Recognition Letters, vol. 31, no. 11, pages 1348–1358, August 2010. QUAERO.

[Penner 2011] Orion Penner, Peter Grassberger and Maya Paczuski. *Sequence alignment, mutual information, and dissimilarity measures for constructing phylogenies.* PloS one, vol. 6, no. 1, pages e14373+, January 2011.

[Piasecki 2009] M. Piasecki, M.P.), S. Szpakowicz, B. Broda and B. Broda. *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej, 2009.

[Pietruschka 2012] Ute Pietruschka. *Corpus der arabischen und syrischen Gnomologien*, 2012. URL: http://casg.orientphil.uni-halle.de/?lang=en last accessed Aug. 29th, 2012.

[Piotrowski 2012] Michael Piotrowski. *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.

[Porter 1980] M. Porter. *An Algorithm for Suffix Stripping.* Program, vol. 14, no. 3, pages 130–137, 1980.

[Puchalla 2009] Marcus Puchalla. *Termextraktion auf antiken Texten*, 2009.

[Radford 2009] W. Radford, B. Hachey, J.R. Curran and M. Milosavljevic. *Tracking information flow in financial text.* In Proc. ALTA, pages 11–19, 2009.

[Reinhold 2011] Marco Reinhold. *Techniken der Mustererkennung (pattern matching) anhand der Delphischen Freilassungsurkunden.* Rapport technique 17, 2011.

[RFP Evaluation Centers 2010a] RFP Evaluation Centers. *Coleman-Liau Grade Level Readability Score, Reading Scores*, 2010. URL: http://rfptemplates.technologyevaluation.com/readability-scores/coleman-liau-readability-score.html last accessed Jul. 21th, 2010.

[RFP Evaluation Centers 2010b] RFP Evaluation Centers. *Dale-Chall 3000 Simple Word List, Readability Grade Score*, 2010. URL: http://rfptemplates.technologyevaluation.com/dale-chall-list-of-3000-simple-words.html last accessed Jul. 21th, 2010.

[Riffaterre 1994] Michael Riffaterre. *Intertextuality vs. hypertextuality.* New literary history, vol. 25, no. 4, pages 779–788, 1994.

[Romanello 2009] Matteo Romanello, Federico Boschetti and Gregory Crane. *Citations in the Digital Library of Classics: Extracting Canonical References by Using Conditional Random Fields.* In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, pages 80–87, Suntec City, Singapore, August 2009. Association for Computational Linguistics.

[Roueché 2010] Charlotte Roueché. *Sharing Ancient Wisdoms*, 2010.

[Salomon 2002] David Salomon. *Handbook of massive data sets.* chapitre Data compression, pages 245–309. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[Salton 1975] G. Salton, A. Wong and C.S. Yang. *A Vector Space Model for Automatic Indexing.* Communications of the ACM, vol. 18, no. 11, pages 613–620, 1975.

[Sander 2010] Sebastian Sander. *Performanceanalyse von SOAP- und REST-basierten Services in einer linguistic Ressources Umgebung*, 2010.

[Schemann 2000] Hans Schemann. *PONS deutsche Redensarten*. Klett, 2000.

[Schleimer 2003] Saul Schleimer, Daniel S. Wilkerson and Alex Aiken. *Winnowing: local algorithms for document fingerprinting.* In Proceedings of the 2003 ACM SIGMOD international conference on Management of data, SIGMOD '03, pages 76–85, New York, NY, USA, 2003. ACM.

[Sediyono 2008] Agung Sediyono and Ku Ruhana Ku-Mahamud. *Algorithm of the longest commonly consecutive word for Plagiarism detection in text based document.* In ICDIM, pages 253–259. IEEE, 2008.

[Seo 2008] Jangwon Seo and W. Bruce Croft. *Local text reuse detection.* In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 571–578, New York, NY, USA, 2008. ACM.

[Shannon 1948] Claude E. Shannon. *A mathematical theory of communication.* Bell System Technical Journal, vol. 27, pages 379–423, 1948.

[Siefkes 2004] Christian Siefkes, Fidelis Assis, Shalendra Chhabra, William S. Yerazunis and Empresa Brasileira De Telecomunicações—embratel. *Combining Winnow and Orthogonal Sparse Bigrams for Incremental Spam Filtering.* In Proceedings of ECM-L/PKDD 2004, LNCS, pages 410–421. Springer Verlag, 2004.

[Smith 1967] E. A. Smith and R. J. Senter. *Automated Readability Index (ARI).* Wright-Patterson AFB, OH: Aerospace Medical Division. AMRL-TR, 66–22, 1967.

[Smith 2010] Neel Smith. *Digital Infrastructure and the Homer Multitext Project.* In Gabriel Bodard and Simon Mahony, editeurs, Digital Research in the Study of Classical Antiquity, pages 121–137. Ashgate Publishing, Burlington, VT, 2010.

[Smith 2012] D.N. Smith and C.W. Blackwell. *Four URLs, Limitless Apps: separation of concerns in the Homer Multitext architecture*. The Center for Hellenic Studies, L. Muellner, ed. Donum natalicium digitaliter confectum Gregorio Nagy septuagenario a discipulis collegis familiaribus oblatum (A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by his Students, Colleagues, and Friends) édition, 2012.

[Spears 1998] Richard A. Spears. *Phrases and Idioms: A Practical Guide to American English Expressions*. McGraw-Hill Contemporary, November 1998.

[Stange 2011] David Stange. *Mental Maps: Aufbau von orts- und zeitabhängigen Bedeutungsräumen zum automatischen Erkennen politischer und gesellschaftlicher Zäsuren im historischen Kontext*, 2011.

[Stein 2007] Benno Stein. *Principles of hash-based text retrieval.* In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '07, pages 527–534, New York, NY, USA, 2007. ACM.

[Sveds 2008] Konstantin Sveds. *Language Independent Sentence Boundary Detection*, 2008.

[Talavera 2000] Luis Talavera, Campus Nord and Jordi Girona. *Dependency-Based Feature Selection for Clustering Symbolic Data*, 2000.

[TEIC 2012] TEIC. *Text Encoding Initiative Consortium - TEI: Text Encoding Initiative*, 2012. URL: http://www.tei-c.org/index.xml last accessed Jul. 5th, 2012.

[Tellex 2003] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes and Gregory Marton. *Quantitative evaluation of passage retrieval algorithms for question answering.* In PROCEEDINGS OF THE 26TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL (SIGIR, pages 41–47. ACM Press, 2003.

[Thomas 2000] C. Thomas. *Das Anagramm-Geheimnis.: Vom Sinn und Hintersinn im Namen*. Droemer, 2000.

[Tschuggnall 2012] Michael Tschuggnall and Günther Specht. *Plag-Inn: Intrinsic Plagiarism Detection Using Grammar Trees.* In Gosse Bouma, Ashwin Ittoo, Elisabeth Métais and Hans Wortmann, editeurs, NLDB, volume 7337 of *Lecture Notes in Computer Science*, pages 284–289. Springer, 2012.

[Turing 1950] Alan M. Turing. *Computing Machinery and Intelligence.* Mind, vol. LIX, pages 433–460, 1950.

[Turner 1998] Loren L. Turner, Workgroup Leader, C. Davis, C. Plumb, T. Holzer, J. Tinsley, M. Reimer, M. Brown and J. Diehl. *Feature Engineering.* In Proceedings of the 9th International Workshop on Software Specification and Design, pages 162–164. IEEE Computer Society, 1998.

[Tuyls 2007] Pim Tuyls, Boris Skoric and Tom Kevenaar. *Security with Noisy Data: Private Biometrics, Secure Key Storage and Anti-Counterfeiting*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[Varol 2010] Marie-Christine Bornes Varol, Marie-Sol Ortola and Jean-Daniel Gronoff. *Aliento project - Intercultural Analysis of Sapiential statements and Transmission*, 2010. URL: http://www.aliento.eu/en/node/63 last accessed Jul. 21th, 2010.

[Voigtländer 2009] Christine Voigtländer. *Morphologische Analyse von antiken Texten im sprach-evolutionären Wandel*, 2009.

[Voigtländer 2013] Elmar Voigtländer. *Entwurf eines hochperformanten und XML-basierten Protokolls für den Distributed Text Re-use*, 2013.

[Wagner 2010] Gerhard Wagner. *Schwein gehabt! - Redewendungen des Mittelalters*. Regionalia Verlag, 2010.

[Wagner 2011a] Gerhard Wagner. *Das geht auf keine Kuhhaut: Redewendungen aus dem Mittelalter*. WBG, Wissenschaftliche Buchgesellschaft, Darmstadt, Germany, 2011.

[Wagner 2011b] Gerhard Wagner. *Wer's glaubt wird selig! Redewendungen aus der Bibel*. WBG, Wissenschaftliche Buchgesellschaft, Darmstadt, Germany, 2011.

[Watts 1998] Duncan J. Watts and Steven H. Strogatz. *Collective dynamics of 'small-world' networks.* Nature, vol. 393, no. 6684, pages 440–442, June 1998.

[Wikipedia 2011] Wikipedia. *Liste der geflügelten Wörter mit T*. World Wide Web electronic publication, jan 2011. http://de.wikipedia.org/wiki/Liste_geflügelter_Worte/T.

[Witschel 2004] F. Witschel. Text, Wörter, Morpheme - Möglichkeiten einer Terminologie-Extraktion. Master's thesis, Universität Leipzig, 2004.

[Zipf 1949] G. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA, 1949.

[Ziv 1977] Jacob Ziv and Abraham Lempel. *A universal algorithm for sequential data compression*. IEEE TRANSACTIONS ON INFORMATION THEORY, vol. 23, no. 3, pages 337–343, 1977.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 3. Februar 2013

Marco Büchler