BAUHAUS - UNIVERSITÄT - WEIMAR
FAKULTÄT MEDIEN
MEDIENSYSTEME

Diplomarbeit

Intrinsische Plagiaterkennung am Beispiel einer Artikelsammlung

Marion Kulig

August 2006

Betreuer:

Sven Meyer zu Eißen
Prof. Dr. Benno Stein (Erstgutachter)
Prof. Dr. Bernd Fröhlich (Zweitgutachter)

Erklärung:

Hiermit versichere ich, dass ich die vorliegende Diplomarbeit selbständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe. Die Arbeit wurde in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt.

Weimar, den 14. August 2006

Marion Kulig

Danksagung:

Zunächst möchte ich mich bei Herrn Prof. Dr. Benno Stein für die Überlassung des Themas dieser Diplomarbeit bedanken.

Ein besonderer Dank geht an meinen wissenschaftlichen Betreuer Herrn Dipl.-Inf. Sven Meyer zu Eißen dafür, dass er mir während der gesamten Zeit sehr motivierend, kompetent und hilfsbereit zur Seite stand, mir in all meinen Fragen weiterhelfen konnte und auch zu mehrfachem Korrekturlesen dieser Arbeit bereit war.

Zusammenfassung

Durch die Verbreitung des Internets stehen jedem Nutzer frei zugängliche Informationen zur Verfügung, deren Anzahl ständig wächst. Das Entnehmen und Hinzufügen fremder Texte zur eigenen Arbeit ohne Verweis auf die Quelle wird als Plagiieren bezeichnet. Dieses Vergehen kann mit Hilfe des Internets und der „copy & paste“- Funktion sehr leicht begangen werden. So ist das Plagiieren, besonders für die Wissenschaft, zu einem großen Problem geworden.

Gegenstand dieser Diplomarbeit ist die Fragestellung, ob und wie gut Plagiate intrinsisch erkannt werden können. Intrinsisch bedeutet allgemein „von innen kommend“ oder „aus sich heraus“. Dieser Definition folgend, werden unter intrinsischer Plagiaterkennung Verfahren bezeichnet, die Plagiate nur aus dem verdächtigen Dokument heraus entdecken können. Es wird kein Referenzkorpus, bestehend aus Originaldokumenten, benötigt.

Das Prinzip der intrinsischen Plagiaterkennung basiert auf einer Stilanalyse des verdächtigen Dokuments. Für diese Stilanalyse werden quantifizierbare Stilmerkmale für den gesamten Text und für jeden einzelnen Abschnitt berechnet. Jeder Wert des Abschnitts wird zum Wert des gesamten Textes ins Verhältnis gesetzt. Mit dieser Methode werden Abweichungen des Schreibstils einzelner Abschnitte vom durchschnittlichen Schreibstil des gesamten Textes berechnet. Eine signifikante Abweichung eines Abschnitts deutet darauf, einen plagiierten Abschnitt entdeckt zu haben.

Für die Evaluierung der intrinsischen Plagiaterkennung wurde ein Korpus erstellt, der auf 100 wissenschaftlichen Artikeln aus der „ACM Digital Library“ basiert. Die Texte wurden extrahiert und durch einfügen von fremden Abschnitten plagiiert. Der Konstruktionsalgorithmus generierte aus diesen vorverarbeiteten Dokumenten Instanzen mit verschiedenen plagiierten Anteilen, sodass insgesamt 3200 plagiierte Dokumente zur Durchführung der Experimente zur Verfügung standen.

Die Ergebnisse wurden mit den bekannten Klassifizierungsverfahren der Diskriminanzanalyse und einer Support-Vektor-Maschine gewonnen. Es zeigte sich, dass die intrinsische Plagiaterkennung funktioniert. Die Precision lag durchschnittlich bei 70 %, der Recall bei 75 %. Des Weiteren konnte festgestellt werden, dass es Stilmerkmale gibt, die besonders gut für die intrinsische Plagiaterkennung geeignet sind.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung und Motivation

Mit der Etablierung des Internets ist es sehr leicht geworden, an Informationen über die vielfältigsten Themen zu gelangen. Durch die ständige Vergößerung der Publikationsmenge steht ein riesiges Angebot frei verfügbarer Beiträge zur Auswahl. Die einfache „copy & paste“ - Funktion des Mediums stellt eine große Versuchung dar, kopierte Informationen dem eigenen Material hinzuzufügen ohne Angaben zum Ursprung zu machen. Aufgrund der scheinbar unüberschaubaren Masse an Informationen besteht oft die Hoffnung des Nichtentdecktwerdens.

Auch, weil oft das Wissen über die rechtlichen Probleme und über die daraus folgenden Konsequenzen fehlt, ist das Plagiieren zu einem wachsenden Problem geworden. Oft wird dieses Vergehen von den Verursachern fälschlicherweise als Kavaliersdelikt angesehen, das, wenn überhaupt, nur gering zu ahnden sei. Es würde ja der Wissenschaft kein unmittelbarer Schaden entstehen, wie das beispielsweise bei Datenmanipulation oder -fälschung der Fall ist. Das ist ein Trugschluss. Bei Wissenschaftlern ist festgestellt worden, dass das Plagiieren, das unrechtmäßige Aneignen fremden geistigen Eigentums, zu *„Demotivierung, ja empirisch nachgewiesenen Entfremdungsverfahren führen und damit auch der Wissenschaftsgemeinde schaden kann“* [15]. Da sich gute Ideen nicht beliebig ersetzen, wiederholen oder versichern lassen, wiegt das Plagiat schwerer als der Diebstahl materiellen Eigentums.

Schlussendlich werden viele wertvolle Erkenntnisse der Öffentlichkeit nicht mehr oder nur eingeschränkt zugänglich gemacht, weil für Wissenschaftler und Organisationen, die in Besitz dieser Informationen sind, die Risiken plagiiert zu werden, inakzeptabel sind [21]. Als Beispiel soll hier die „ACM Digital Library“ stehen, auf deren Publikationen nur kostenpflichtig zugegriffen werden kann.

Um Plagiate finden zu können, kann das Internet eine große Unterstützung sein. Beispielsweise wird der kostenpflichtige „Einreichdienst“ Turnitin[1] bereits an vielen Universitäten der USA, Kanadas und Großbritanniens eingesetzt, um die Arbeiten von Studenten auf Online-Plagiarismus zu kontrollieren. Jede eingereichte Arbeit wird dabei in einer Datenbank gespeichert, um diese Ar-

[1] **Turnitin:** www.turnitin.com

beiten auch untereinander zu vergleichen. Ohne die Einverständnis des Studenten findet aber ein Verstoß gegen datenschutzrechtliche und urheberrechtliche Bestimmungen statt.

Des Weiteren ist der Erfolg solcher Software immer vom Aufwand des Studenten abhängig. Fremdsprachige Texte, die mittels Übersetzungsprogramm und Fehlerkorrektur angepasst wurden, werden nach Erfahrungsberichten [49] nicht gefunden. Der größte Nachteil dieser Programme aber ist, dass der Nachweis des Plagiats nur bei den Texten erreicht wird, deren Ursprungsquellen sich auch online finden lassen. Plagiierte Stellen, die aus anderen Quellen, wie zum Beispiel aus Büchern stammen, bleiben weiter unentdeckt.

In dieser Diplomarbeit wird untersucht, inwieweit es möglich ist, plagiierte Textabschnitte anhand Änderungen des Schreibstils im verdächtigen Dokument zu entdecken und somit unabhängig von jeglichen vergleichenden Quellen zu sein. Die Idee der intrinsischen Plagiaterkennung beruht auf der Tatsache, dass jeder Mensch einen eigenen Schreibstil hat, den er nur bedingt bewusst verändern kann. Plagiierte Abschnitte, selbst wenn diese etwas umgeschrieben wurden, sollten somit nicht dem Schreibstil des plagiierenden Autors entsprechen.

Den ersten Teil dieses Kapitels bilden Informationen über rechtliche Grundlagen des Plagiats, im Besonderen der Zusammenhang des Plagiats zum Urheberrecht. Es folgt eine Taxonomie der Plagiate und die Indizien, die für ein solches sprechen können. Den Abschluss bildet eine Taxonomie der Plagiatvergehen und deren Nachweis, was die Grundlage des zweiten Kapitels dieser Diplomarbeit ist, in dem eine Auswahl aktueller Algorithmen zur Erkennung von Plagiaten vorgestellt wird.

Im dritten Kapitel werden Methoden zur Quantifizierung von Stilmerkmalen erläutert, eine Auswahl derer dann zur Plagiaterkennung verwendet wird. Zur Evaluierung dieser Algorithmen entstand ein Korpus, dessen Aufbau und zugrunde liegende korpuslinguistische Kriterien im vierten Kapitel beschrieben sind. Die im fünften Kapitel dokumentierte Evaluierung wurde unter Verwendung bekannter statistischer Klassifizierungsverfahren durchgeführt. Den Abschluss bildet das sechste Kapitel, in dem alle gewonnenen Erkenntnisse dieser Arbeit zusammengefasst sind.

1.1 Plagiat und Urheberrecht

Allgemein wird unter einem Plagiat der Diebstahl fremden geistigen Eigentums verstanden. Das ist der Fall, wenn eine Idee, eine Hypothese, eine Theorie oder ähnliches ohne Nennung des ursprünglichen Autors übernommen wird. Dabei gibt es keine festen Angaben, ab welchem Umfang von einem Plagiat gesprochen werden kann. Die Rechtssprechung dazu ist einzelfallbezogen und ein Satz kann bereits ausreichen.

Wenn Teile eines Werks identisch in ein anderes übernommen werden, ist das Urheberrecht betroffen. Das Vergehen stellt einen Eingriff in das „Recht auf Anerkennung der Urheberschaft“ dar:

> **§ 13 UrhG**
> *Der Urheber hat das Recht auf Anerkennung seiner Urheberschaft am Werk. Er kann bestimmen, ob das Werk mit einer Urheberbezeichnung zu versehen und welche Bezeichnung zu verwenden ist.*

Dabei ist es urheberrechtlich unerheblich, ob der Plagiierende sich ein fremdes Werk gesamt oder nur in Teilen anmaßt. Auch die absolute Länge spielt keine Rolle.

Wenn der Text eines anderen umgeändert und unter eigenem Namen veröffentlicht wird, ist es ein Plagiat und verstößt unter Umständen gegen das Bearbeitungsrecht:

> **§ 23 S.1 UrhG**
> *Bearbeitungen oder andere Umgestaltungen des Werkes dürfen nur mit Einwilligung des Urhebers des bearbeiteten oder umgestalteten Werkes veröffentlicht oder verwertet werden.*

Die „freie Benutzung“ eines fremden Werks dagegen ist kein Plagiat. Dabei muss aber die schöpferische Leistung des fremden Werks hinter den eigenen individuellen geistigen Leistungen zurücktreten. Es muss ein neues selbständiges Werk geschaffen werden.

> **§ 24 (1) UrhG**
> *Ein selbständiges Werk, das in freier Benutzung des Werkes eines anderen geschaffen worden ist, darf ohne Zustimmung des Urhebers des benutzten Werkes veröffentlicht und verwertet werden.*

Die Abgrenzung von „freier Benutzung“ und „Bearbeitung“ ist schwierig. Das zeigt sich besonders in der Wissenschaft, in der das Plagiat spezieller definiert ist. Bereits die Paraphrasierung

eines Textes oder die nicht gekennzeichnete Übernahme einer Argumentation ohne Quellenangabe wird als solches verstanden, ebenso die Übernahme von Gedanken eines anderen ohne Verweis auf diesen. Ideen sind aber urheberrechtlich nicht schützbar. Trotzdem wird in diesem Fall von Plagiat gesprochen und zwar auch dann, wenn das plagiierte Werk eine völlig andere Formulierung hat und damit keine Urheberrechtsverletzung vorliegt [50].

Die Hochschulrektorenkonferenz stuft in ihrer Empfehlung das Plagiat, die *„unbefugte Verwertung unter Anmaßung der Autorenschaft“*, als *„schwerwiegendes Fehlverhalten“* ein. Bei dessen Nachweis können als Konsequenz der Entzug akademischer Grade oder der Entzug der Lehrbefugnis folgen. Es können je nach Sachverhalt außerdem arbeits-, zivil-, straf- oder ordnungsrechtliche Maßnahmen eingeleitet werden [22].

1.2 Taxonomie der Plagiate

Die eindeutige Form des Plagiats ist die Kopie. Das reicht dabei von der vollständigen Kopie, bei der der Plagiierende seinen Namen über ein komplettes fremdes Werk setzt, bis hin zum Kopieren einzelner Sätze. Wenn ein Autor in einer Aussage auf einen anderen verweist, dabei aber den Wortlaut des anderen verwendet, wird auch von einem Plagiat gesprochen, weil das Fassen in eigene Worte fehlt. Dazu ein Beispiel:

> **Original:**
>
> *„Der Begriff **Korpuslinguistik** bezeichnet eine methodische Ausrichtung der Sprachwissenschaft, in welcher linguistische Fragestellungen und Hypothesen an großen Sammlungen authentischer Texte (so genannte Korpora) empirisch untersucht werden.“* [38]
>
> **Plagiat:**
>
> Nach Paprotté [38] bezeichnet der Begriff der Korpuslinguistik eine methodische Ausrichtung der Sprachwissenschaft, in welcher lingusitische Fragestellungen und Hypothesen an großen Sammlungen authentischer Texte empirisch untersucht werden.

Zu den Plagiatarten zählen des Weiteren Modifizierungen von fremden Werken. Dazu gehören Übersetzungen, Paraphrasierungen und Plagiate von Form und Idee. Unter Paraphrasierung wird z.B. die Umstellung der Syntax, die Bearbeitung von Sätzen, Teilsätzen, das Austauschen von

einzelnen Wörtern und das Umstellen von Aufzählungen verstanden. Bei dem Plagiat einer Form, ist die Struktur, also der gedankliche Aufbau eines Werks übernommen worden. Trotz anderer Wortwahl ist dies ein Plagiat. Im Gegensatz dazu wird bei der Ideeplagiierung ein eigener Wortlaut für eine fremde Idee verwendet.

Die hier beschriebenen Plagiatarten lassen sich in einer Taxonomie zusammenfassen, die in Abbildung 1.1 dargestellt ist.

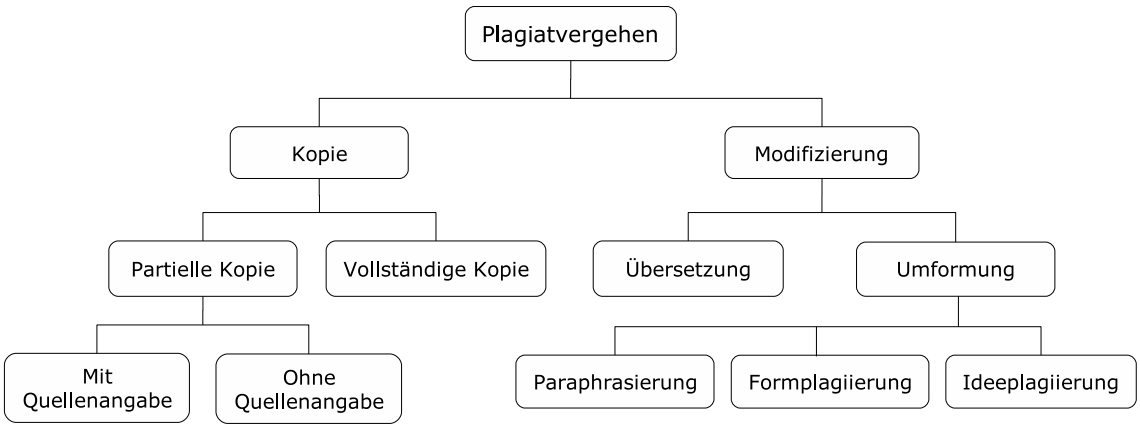


Abbildung 1.1: *Taxonomie der Plagiate*

1.3 Indizien für ein Plagiat

Für einen geübten Leser ist es oft möglich, offensichtlich plagiierte Stellen innerhalb eines Dokuments zu erkennen. Die hier in Anlehnung an Weber-Wulff [58] beschriebenen drei Indizienarten können dafür sprechen, ein plagiiertes Dokument vorliegen zu haben.

Im Dokument gibt es Stilwechsel. Es wechseln sich hervorragende Formulierungen und schlechte ab, subjektiv und objektiv geschriebene oder es treten plötzliche Wechsel zwischen den Zeitformen auf. Zu den Stilwechseln zählen außerdem Formatänderungen (Zeilenabstände, Schriftgrößen, Nummerierungen), uneinheitliche Überschriften (Kapitel-, Zwischenüberschriften, Tabellen- und Bildunterschriften, ...), angehäuftes Auftreten von Fremdwörtern oder wechselnde Zitierstile.

Das zweite Indiz für ein Plagiat sind Fehlerhäufigkeiten. So können orthographische, grammatikalische und Fehler in der Zeichensetzung dafür sprechen, von einer fremden Quelle übernommen worden zu sein. Auffällige und wiederkehrende Tippfehler in Eigennamen und Fachbegriffen können ebenso für ein Plagiat sprechen, wie auch die unmotivierte, übertriebene und fehlerhafte Verwendung dieser.

Literaturverzeichnisse, die identisch sind oder aus Einträgen bestehen, die als Verweise nicht im Text auftauchen, sind Indizien dafür, dass sowohl die Literatur, als auch Teile des betreffenden Dokuments plagiiert sein könnten.

1.4 Plagiatvergehen und deren Nachweis

Wenn der Verdacht geschöpft wird, eine plagiierte Arbeit vorliegen zu haben, muss dieser Verdacht nachgewiesen werden, was bis jetzt darauf hinausläuft die Quelle des Plagiats zu finden. Eine einfache Möglichkeit ist, verdächtige Stellen mittels einer Literaturrecherche zu überprüfen. Suchmaschinen bieten eine erste Möglichkeit für das Finden von Internetquellen.

Die in Kapitel 1.3 genannten Indizien werden häufiger in Schulhausarbeiten oder Seminararbeiten gefunden, als zum Beispiel in wissenschaftlichen Veröffentlichungen. Es stellt sich die Frage, wie überhaupt verdächtige Stellen und demzufolge Plagiate erkannt werden können, wenn keine solch offensichtlichen Fehler bezüglich Stilwechsel, Fehlerhäufigkeiten und Literaturverzeichnissen vorhanden sind, oder der Aufwand für die Suche mittels aufmerksamen Lesens zu groß ist.

Für diese Zwecke gibt es automatische Erkennungsmethoden, die auf verschiedene Art und Weise Plagiatvergehen entdecken können. In Abbildung 1.2 von Stein und Meyer zu Eißen [36] werden die Plagiatvergehen den Erkennungsmethoden gegenübergestellt. Eine Beschreibung dieser Methoden folgt im anschließenden Kapitel 2.

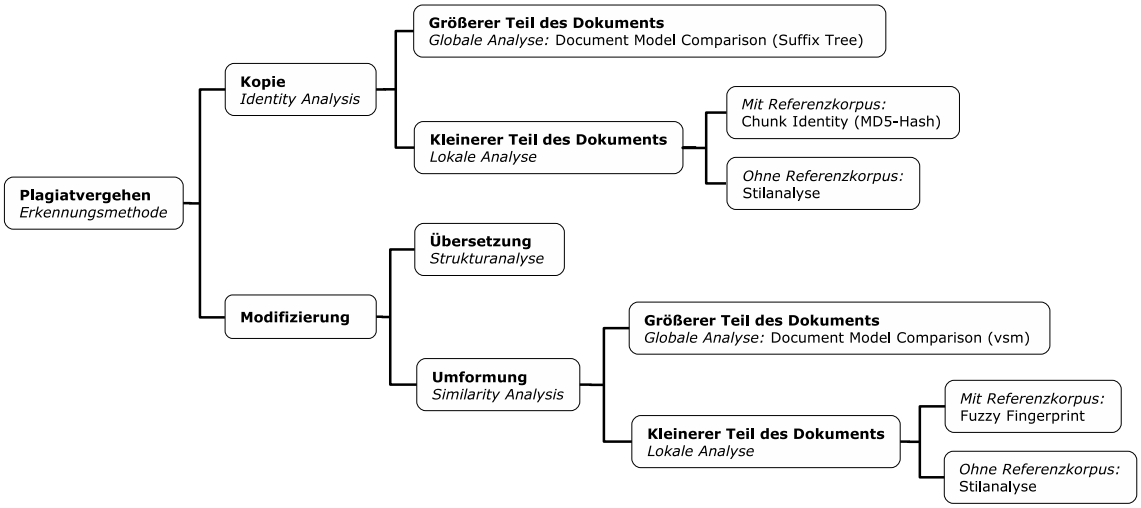


Abbildung 1.2: *Taxonomie der Plagiatvergehen und ihre Erkennungsmethoden [36]*

2 Algorithmen zur Erkennung von Plagiatvergehen

Die in Abbildung 1.2 zusammengestellten Erkennungsmethoden für Plagiatvergehen werden in diesem Kapitel näher erklärt. Dabei erfolgt eine Einteilung nach globaler und lokaler Dokumentanalyse. Bei einer globalen Analyse wird das gesamte Dokument für den Vergleich mit einer Kandidatendokumentmenge herangezogen. Im Gegensatz dazu werden bei der lokalen Analyse Textabschnitte miteinander verglichen.

2.1 Globale Dokumentanalyse

Bei der Plagiaterkennung mittels globaler Dokumentanalyse werden Modelle als Repräsentation von Dokumenten verwendet. So werden beim Vektorraummodell Dokumente als Vektoren innerhalb eines Vektorraums repräsentiert. Die Realisierung dieses Modells geschieht in drei Schritten. Zuerst werden Indexterme[1] anhand der Datenbasis, bzw. Kandidatendokumentmenge festgelegt. Dafür werden zum Beispiel alle Funktionswörter[2] entfernt, da diese nicht signifikant für den Inhalt eines Textes sind. Für jeden dieser Indexterme jedes Dokuments wird ein Gewicht berechnet und im Dokumentvektor abgespeichert. Das Gewicht wird aus der Häufigkeit des Auftretens des Terms im Dokument, in allen Dokumenten und/oder einem Normalisierungsfaktor bezüglich der Dokumentlänge berechnet. So kann schließlich eine Rangfolge bezüglich der Ähnlichkeit eines Dokuments zur Datenbasis festgelegt werden. Dies erfolgt mittels Berechnung der Skalarprodukte oder der Winkel zwischen den entsprechenden Vektoren. Es wird also festgelegt, dass sich zwei Dokumente „ähnlich“ sind, wenn der Anteil gemeinsamer Worte groß ist.

Um „ähnliche“, also plagiierte Abschnitte erkennen zu können, werden die Dokumente in Abschnitte, sog. Chunks aufgeteilt. Unter Chunking wird allgemein das *„Bündeln oder Gruppieren von Informationseinheiten“* [42] verstanden. Die theoretisch minimale Chunklänge ist die Länge

[1] **Term:** Der Begriff des Terms wird hier als mathematisch korrektes Wort oder Wortgruppe einer Sprache verstanden.
[2] **Funktionswörter:** Artikel, Pronomen, Negationswörter, Adverbialpronomen, Präpositionen, Konjunktionen

eines Zeichens, die maximale entspräche dem gesamten Dokument. So wird im Vektorraummodell nicht jedes Dokument, sondern jeder Chunk als Vektor repräsentiert.

Für das Finden eines plagiierten Abschnitts im Referenzkorpus, bestehend aus n Dokumenten, muss jeder Chunk des verdächtigen Dokuments mit jedem Chunk jedes Dokuments des Korpus verglichen werden, was einer Laufzeitkomplexität von $O(n \cdot c^2)$ entspricht. c ist die durchschnittliche Chunkanzahl eines Dokuments.

Je kleiner die Chunks gewählt sind, desto kleinere plagiierte Abschnitte können gefunden werden. Da Worthäufigkeiten und keine grammatikalische Struktur verglichen werden, können auch veränderte Abschnitte entdeckt werden. Bei großen Korpora ist dieser Algorithmus aufgrund der quadratischen Laufzeitkomplexität nicht praktikabel. Eine Möglichkeit plagiierte Abschnitte in linearer Zeit zu finden, bietet das Modell der Suffixbäume.

Ein Suffixbaum T für eine Zeichenkette S bestehend aus w Wörtern ist ein gerichteter Baum mit w Blättern. Jeder Knoten von T besitzt mindestens zwei Kinder, deren Kantenbeschriftungen nie mit dem gleichen Wort beginnen. Jeder Suffix eines Korpus wird als Kante im Suffixbaum gespeichert. Der i-te Suffix einer Zeichenkette beginnt beim i-ten Wort und reicht bis zu ihrem Ende. Anstatt Wörter können auch Buchstaben, Sätze oder andere Teilstrings verwendet werden. In einer Zeichenkette mehrfach auftretende Teilstrings sind nur einmal in T enthalten. Das Suchen eines Strings der Länge n wird mittels Suffixbaums in einer Laufzeit von $O(n)$ ermöglicht.

Für tiefergehende Informationen bezüglich Suffixbäumen wird an dieser Stelle auf Gusfield [20] verwiesen.

2.2 Lokale Dokumentanalyse

Bei Methoden der lokalen Dokumentanalyse werden Textabschnitte des verdächtigen Dokuments auf Plagiatvergehen untersucht. Dafür gibt es mehrere Ansätze, die entweder auf einen Referenzkorpus angewiesen sind, oder nicht.

2.2.1 Hashing

Jedes Dokument eines Korpus D, bestehend aus n Dokumenten wird in Chunks eingeteilt. Mittels einer Hashfunktion h werden Hash-Codes der Chunks berechnet. Zusammen stellen alle Hash-Codes eines Dokuments den Fingerprint dieses dar. Alle Hash-Codes werden in einer Hashtabelle gespeichert. Wenn ein Hash-Code $h(d)$ eines Chunks $d \in D$ gleich dem Hash-Code $h(d')$ eines

Chunks $d' \in D$ ist, dann ist mit großer Wahrscheinlichkeit eine Gleichheit der beiden Chunks d und d' gegeben:

$$h(d) = h(d') \Rightarrow d = d'$$

Diese Überschneidungen in den Fingerprints zweier Dokumente deuten auf plagiierte Textabschnitte. Das Finden dieser hat eine Laufzeitkomplexität von $O(n \cdot c)$ bei n Dokumenten im Korpus und einer durchschnittlichen Anzahl von c Chunks pro Dokument.

Da eine Hashfunktion den selben Wert für verschiedene Chunks berechnen kann, sind falsche Treffer möglich. Die meisten Anwendungen verwenden für das Hashing den MD5-Algorithmus.

Für die Plagiaterkennung ist es wichtig zu wissen, dass ein einzelnes verändertes Zeichen im Chunk ausreicht, um einen völlig anderen Fingerprint zu erzeugen. Somit können nur identisch plagiierte Textabschnitte gefunden werden. Aus diesem Grund müssen die Chunks ausreichend klein gewählt sein, was den Aufwand bezüglich Fingerprintberechnung, -vergleich und -speicherung erhöht. Zu kleine Chunks können dagegen zu falschen Übereinstimmungen führen. Als minimale Länge werden in vielen Anwendung 50 bis 60 Zeichen verwendet. Beispiele dafür sind die „Shingling - Methode“ [4] und das SCAM[3] System [44].

2.2.2 Fuzzy-Fingerprinting

Das von Stein [52] eingeführte und mit Meyer zu Eißen [53] für die Plagiaterkennung weiterentwickelte Konzept der Fuzzy-Fingerprints basiert auf dem Prinzip des „Similarity-Hashing“, bei dem Hashkollisionen als Indikator für Ähnlichkeit interpretiert werden.

Beim Hashing, wie in Abschnitt 2.2.1 beschrieben, werden die Fingerprints der Chunks gleichmäßig in einer Hashtabelle, unabhängig von ihren Beziehungen untereinander, verteilt. So können keine Ähnlichkeiten zwischen Chunks festgestellt werden. Beim Similarity-Hashing wird im Gegensatz dazu von Hashfunktionen gefordert, dass die Verteilung der Chunks in der Hashtabelle von ihren Ähnlichkeiten zueinander abhängen. So bedeuten gleiche Fingerprints exakt gleiche Chunks, während gleiche Fuzzy-Fingerprints eine bestimmte Ähnlichkeit zwischen Chunks anzeigen.

Wenn der mittels Hashfunktion h_φ berechnete Fuzzy-Fingerprint $h_\varphi(d)$ eines Dokuments $d \in D$ gleich dem Fuzzy-Fingerprint $h_\varphi(d')$ eines Dokuments $d' \in D$ ist, dann ist das mit großer Wahrscheinlichkeit eine Bedingung für die Ähnlichkeit der Dokumente d und d' bezüglich eines

[3] **SCAM:** Stanford Copy Analysis Method

Schwellwerts ϵ. Die Ähnlichkeit wird mittels einer Ähnlichkeitsfunktion $\varphi(\mathbf{d}, \mathbf{d'})$ gemessen. Die Feature-Vektoren **d** und **d'** stellen die Modelle der Dokumente d, bzw. d' dar[4].

$$h_\varphi(d) = h_\varphi(d') \Rightarrow \varphi(\mathbf{d}, \mathbf{d}') \geq 1 - \epsilon, \quad \text{mit } d, d' \in D, d \neq d', 0 < \epsilon \ll 1$$

Um Dokumente effizient mit wenigen Dimensionen im Gegensatz zum Vektorraummodell zu beschreiben, stellt Stein [52] Präfixklassen vor. Eine Präfixklasse P_w für eine Buchstabenfolge w ist die Menge aller Worte $W_w \subseteq W$, die mit w beginnen. Für das Fuzzy-Fingerprinting werden zwischen 10 und 40 Präfixklassen verwendet, deren Präfixe kurz sind. Meist bestehen sie nur aus einem Buchstaben. Die Präfixklassen können dann wie folgt zu einem Fuzzy-Fingerprint $h_\varphi(d)$ für ein Dokument $d \in D$ verrechnet werden.

Der erste Schritt ist die Berechnung des Vektorraummodells **d** aus d. Im zweiten Schritt folgt die Berechnung von **pf**, dem Vektor der relativen Häufigkeiten von k Präfixklassen für die Indexterme in **d**. Der dritte Schritt beginnt mit der Normalisierung von **pf** bezüglich der Präfixklassen eines Referenzkorpus[5]. Es wird Δ_{pf} als Vektor der Abweichungen zur erwarteten Verteilung berechnet.

Im vierten Schritt wird Δ_{pf} fuzzifiziert, d.h. die exakten Abweichungen in Δ_{pf} werden mit Hilfe einer Zugehörigkeitsfunktion auf linguistische Variablen (z.B. „keine“, „kleine“, „große“ Abweichung) abgebildet. Das Fuzzifizierungsschema kann bis zu r Intervalle haben. Stein [52] schlägt vor zwei bis drei Intervalle zu verwenden.

Aus dem normalisierten Vektor der Abweichungen eines Dokuments Δ_{pf} der Länge k kann dann ein Fuzzy-Fingerprint $h_\varphi^{(\rho)}$, wie in Formel 2.1 dargestellt, kodiert werden. Dafür sei ein Fuzzifizierungsschema ρ mit r Intervallen gegeben.

$$h_\varphi^{(\rho)}(d) = \sum_{i=0}^{k-1} \delta_i^{(\rho)} \cdot r^i \ , \quad \text{mit} \quad \delta_i^{(\rho)} \in \{0, \ldots, r-1\} \tag{2.1}$$

$\delta_i^{(\rho)}$ ist ein dokumentspezifischer Wert, der die fuzzifizierte Abweichung der Häufigkeit der Präfixklasse i beim Hinzufügen des Fuzzifizierungsschemas ρ kodiert. Für detailliertere Erklärungen sei an dieser Stelle auf die Veröffentlichung von Stein [52] und die Diplomarbeit von Potthast [39] verwiesen.

Das Verfahren des Fuzzy-Fingerprinting erlaubt eine größere Chunkgröße, um ähnliche Chunks zu entdecken. So wird eine kleinere Datenbasis als beim Hashing benötigt. Ein Experiment von Stein

[4] Fuzzy-Fingerprints können als eine Abstrahierung des klassischen Vektorraummodells gesehen werden.

[5] Als Referenzkorpus verwendet Stein [52] den „British National Corpus“, einer Sammlung von 100 Mio. Wörter, geschriebener und gesprochener Sprache als repräsentativer Querschnitt der gegenwärtigen englischen Sprache.

und Meyer zu Eißen [53] zeigt, dass Fuzzy-Fingerprints im Durchschnitt eine Chunkgröße von 100 Wörtern erlauben, während MD5-Fingerprinting nur akzeptabel bei Chunkgrößen zwischen drei und zehn Wörtern arbeitet. Fuzzy-Fingerprinting identifiziert in kürzerer Zeit verdächtige Chunks, die aufgrund der vielfach größeren Chunklänge hochsignifikant sind.

2.2.3 Stilanalyse

Die Stilanalyse ist ein Verfahren, das intrinsische Entdeckung von Plagiaten ermöglicht. Das heißt, Plagiate können aus der Analyse des verdächtigen Dokuments heraus gefunden werden. Es wird kein Referenzkorpus benötigt.

Der individuelle Schreibstil eines Menschen kann mit Hilfe quantifizierbarer Stilmerkmale berechnet werden. Als Voraussetzung für das Erkennen von Plagiaten müssen zwei beliebige Autoren immer anhand ihrer berechneten Schreibstile voneinander unterschieden werden können. Den Nachweis dieser Tatsache erbrachten Baayen et. al [2]. In ihrer Studie lieferten acht Versuchspersonen jeweils neun Texte aus verschiedenen Genres und Themen ab. Auf diese Texte wandten Wissenschaftler bekannte Verfahren zur Berechnung von Stileigenschaften an und stellten Unterschiede in den Schreibstilen der Autoren innerhalb eines Genres fest, sofern der Autor seinen Stil nicht bewusst in einem Text änderte.

Wird innerhalb eines Dokuments eines Autors ein Abschnitt gefunden, dessen Schreibstil signifikant vom Schreibstil des gesamten Dokuments abweicht, so ist dieser Abschnitt mit hoher Wahrscheinlichkeit plagiiert.

Die intrinsische Plagiaterkennung mittels Stilanalyse basiert auf folgendem Prinzip. Zunächst wird der Schreibstil A eines Dokuments D auf Basis einer Menge von Stilmerkmalswerten $\{a_1, \ldots, a_x\}$ berechnet. Ein einfaches Beispiel für einen Stilmerkmalswert ist „Durchschnittliche Wortanzahl pro Satz“. Das Dokument wird dann in n Textabschnitte[6] eingeteilt. Für jeden Abschnitt $c \in D$ wird der Schreibstil B aus den Stilmerkmalswerten $\{b_1, \ldots, b_x\}$ berechnet.

Anschließend wird das Verhältnis der Werte $b \in \{b_1, \ldots, b_x\}$ jedes Abschnitts $c \in \{c_1, \ldots, c_n\}$ zum jeweiligen Stilwert $a \in \{a_1, \ldots, a_x\}$ von D nach Formel 2.2 berechnet.

$$f(a,b) = 2 \cdot \left(\frac{a}{a+b}\right) - 1 \;\text{ mit }\; a, b \geq 0\ ,\ a + b > 0 \tag{2.2}$$

$f(a,b)$ liegt im Intervall $[-1,+1]$. Bei 0 sind der jeweilige Stilmerkmalswert des Abschnitts und der Stilmerkmalswert des gesamten Textes identisch. Maximale Unterschiede können aus Wer-

[6] Textabschnitte sind hier durch Leerzeilen getrennt.

ten nahe -1 und +1 geschlossen werden. Da zur Schreibstilberechung x Stilmerkmale berechnet werden, ergeben sich x Werte für jeden Abschnitt $c \in \{c_1, \ldots, c_n\}$. Befindet sich ein Großteil dieser Werte eines Abschnitts nahe den maximalen Abweichungswerten von -1 und +1, so ist der betreffende Abschnitt mit großer Wahrscheinlichkeit plagiiert.

Da dieses Verfahren der intrinsischen Plagiaterkennung auf der Berechnung von Stilunterschieden innerhalb eines Textes basiert, können vollständige Plagiate, bei denen ein Autor seinen Namen auf ein fremdes Werk setzt, nicht erkannt werden.

3 Methoden zur Quantifizierung von Stil

Bei extrinsischen Methoden zur Plagiaterkennung wird ein verdächtiges Dokument mit einer elektronisch verfübaren Sammlung von Referenzdokumenten verglichen. Ein plagiierter Text, dessen Ursprung elektronisch nicht vorhanden ist, wie zum Beispiel Text aus einem Buch, kann so nicht entdeckt werden. Eine mögliche Lösung bietet hier der Einsatz von intrinsischen Methoden, die plagiierte Stellen eines Dokuments anhand von Schreibstiländerungen erkennen können.

Der Schreibstil eines Autors ist zum einen durch messbare, stilistische Eigenschaften, wie Satzlängen, Wortlängen oder Wortvielfalt gekennzeichnet. In der Literaturwissenschaft gehören zum anderen auch Sprachfiguren, wie Ironie oder Metaphern zum Stilbegriff. Diese semantischen Merkmale ergeben sich aus der Bedeutung und dem Sinn des geschriebenen und sind momentan noch nicht zuverlässig quantifizierbar. Sie nehmen deshalb zur Stilberechnung in dieser Arbeit keinen Platz ein.

Nach Sorensen [48] beschäftigte sich schon im 19. Jh. der Mathematiker DeMorgan mit der Frage, wie Schreibstileigenschaften von Autoren berechnet werden könnten. Eine Idee von ihm war, die durchschnittliche Buchstabenanzahl eines Wortes dafür zu verwenden. 1901 veröffentlichte Mendenhall eine Studie, in der er anhand von Wortlängen-Häufigkeiten schloss, dass es unwahrscheinlich ist, dass Bacon Werke von Shakespeare geschrieben hat. Diese Annahme könnte falsch sein, wie später gezeigt wurde, da beim damaligen Vergleich verschiedene Textsorten untersucht worden sind (ältere Prosa und jüngere Verse[1]) [5].

Die Wissenschaft, die sich mit der Quantifizierung von Stilmerkmalen[2] beschäftigt, heißt Stilometrie. Es wird der Stil von Texten untersucht, um Aussagen zur Urheberschaft eines Werks zu treffen. Es können so beispielsweise Texte einem bestimmten Autor zugeordnet oder Ergänzungen anderer Autoren erkannt werden. Dafür müssen diejenigen Stilmerkmale gefunden werden, die einen Texabschnitt eindeutig von einem anderen unterscheiden. Zur stilometrischen Analyse sollten Stilmerkmale idealerweise quantifizierbar und andere Textgruppen relevant sein. Sie soll-

[1] Als Prosa werden alle sprachlichen Darstellungs- und Mitteilungsformen bezeichnet, die nicht an Verse gebunden sind. [29]

[2] Hier wird der Begriff „Stilmerkmale“ als Übersetzung zu „Style marker“ verwendet .

ten des Weiteren möglichst nicht der Kontrolle des Autors unterliegen, was besonders im Hinblick auf die Plagiaterkennung sehr wichtig ist.

Stilmerkmale basieren auf den in Abbildung 3.1 dargestellten quantifizierbaren Stileigenschaften. Diese können in wort-/zeichenbasiert und syntaktisch eingeteilt werden. Nach Baayen [1] können erstere allein, den Schreibstil nicht repräsentieren. Dafür wird eine Kombination mehrerer Stilmerkmale aus beiden Gruppen benötigt.

Stilmerkmale sind die am häufigsten gebrauchten Merkmale, die in Studien verwendet werden [62]. Bis zum heutigen Tage sind mehr als 1000 in entsprechenden Stilometrie-Studien identifiziert worden.

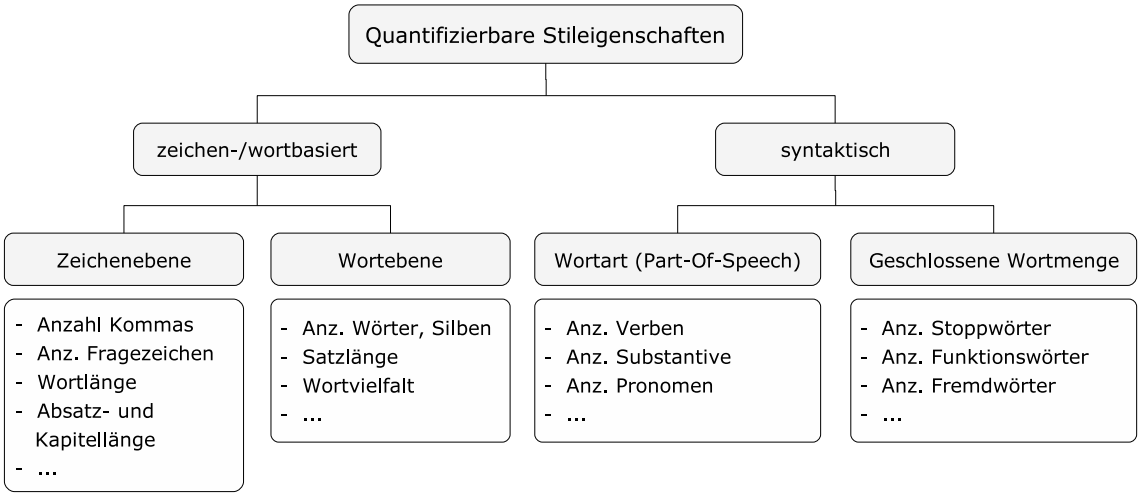


Abbildung 3.1: *Taxonomie der quantifizierbaren Stileigenschaften. Stileigenschaften die Längen von Zeichenketten berechnen, können sowohl zur Zeichenebene, als auch zur Wortebene gezählt werden. So kann die Länge eines Satzes auch in Anzahl Zeichen, und die Absatzlänge in Anzahl Wörter angegeben werden.*

Für dieses Kapitel wurde eine, in Abbildung 3.2 von Meyer zu Eißen et. al [34] dargestellte, repräsentative Auswahl an häufig verwendeten Stilmerkmalen für die englische Schriftsprache getroffen. Diese werden bezüglich ihrer Vorteile und Nachteile zur Plagiaterkennung beschrieben und miteinander verglichen.

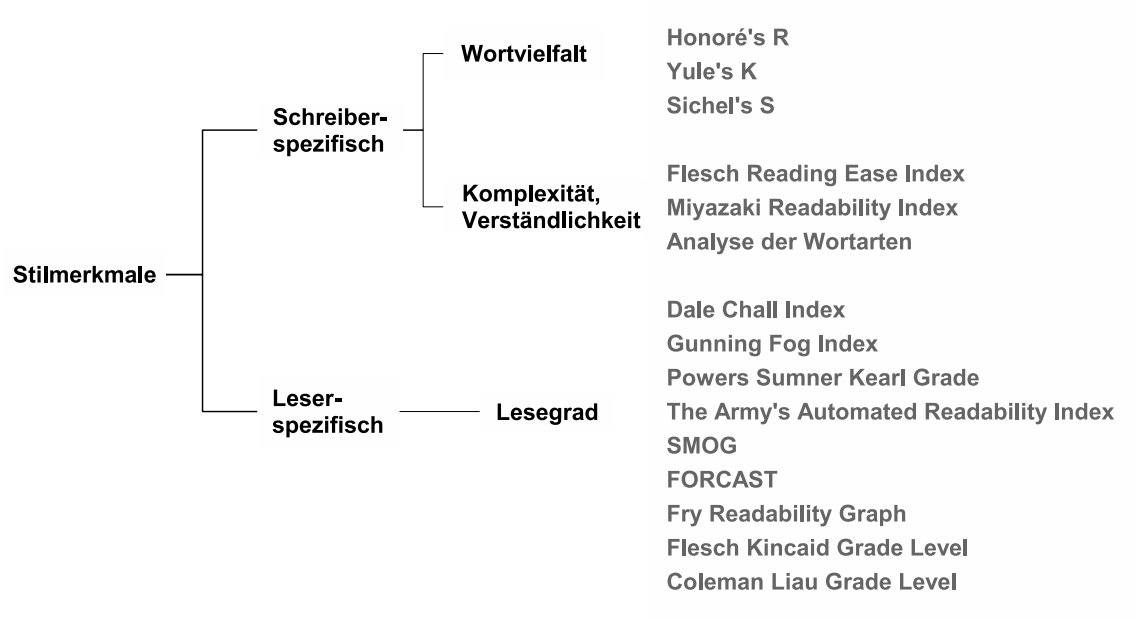


Abbildung 3.2: *Taxonomie der Stilmerkmale. Mit schreiberspezifischen Formeln kann die Fähigkeit des Schreibers beurteilt werden. Lesespezifische Formeln bewerten den Leser.*

3.1 Berechnung der Wortvielfalt

Die Wortvielfalt, bzw. lexikalische Reichhaltigkeit eines Autors kann mit Hilfe sogenannter Vocabulary-Richness-Measures (VRM) berechnet werden. Das meist verwendete VRM ist das Verhältnis der verschiedenen Wörter zur Gesamtanzahl aller Wörter eines Textes (Type/Token ratio TTR). Der Nachteil dieses Maßes ist die Abhängigkeit von der Textlänge. Die Anzahl aller Wörter ist theoretisch unbegrenzt, während die Zahl verschiedener hingegen endlich ist und mit zunehmender Textmenge langsamer zunimmt [55]. Deshalb ist das TTR nur für Vergleiche von Texten mit identischer oder zumindest ähnlicher Länge geeignet [23]. Aus diesem Grund sind viele erweiterte Methoden entwickelt worden, die sich nur auf einen Auschnitt der gesamten Wortmenge, wie zum Beispiel nur auf Funktionswörter oder ausgewählte Uni-, Bi-, oder Trigramme [3], beziehen.

Es gibt drei weitere VRMs die sich auf spezielle Worthäufigkeiten stützen. Für diese sei V die Menge aller Wörter eines Dokuments $D = w_1, ..., w_n$, N die Anzahl aller Zeichen von D und V_i die Anzahl aller Wörter die genau i-mal in D auftreten.

Honoré's R verwendet für die Berechnung der Worthäufigkeit Wörter, die nur ein einziges Mal

[3] zusammenhängende Wortgruppen, bestehend aus einem, zwei oder drei Wörtern

in einem Text auftreten. Diese Wörter werden „hapax legomena“ genannt. Je mehr von ihnen in einem Text enthalten sind, desto differenzierter gilt der Wortschatz eines Autors. Honoré geht von der Annahme aus, dass das Verhältnis der hapax legomena zur Anzahl aller Wörter eines Textes bezüglich des Logarithmus der Textgröße konstant ist.

Honoré's R (1979) [24]:

$$R = \frac{100 \cdot \log N}{1 - \frac{V_1}{V}} \tag{3.1}$$

Sichel's S verwendet Wörter, die genau zweimal innerhalb eines Textes auftreten (hapax dislegomena). Sichels Untersuchungen ergaben, dass das Verhältnis dieser zur Gesamtanzahl der Wörter für Texte zwischen 1000 und 400000 Wörtern nahezu konstant ist. Zur Unterscheidung verschiedener Autoren ist dieses Maß nach Tweedie und Baayen [55] nicht geeignet, da die Variabilität zwischen verschiedenen Autoren genauso groß war, wie jene innerhalb von Texten eines Autors.

Sichel's S (1975) [45]:

$$S = \frac{V_2}{V} \tag{3.2}$$

Yule's K-Charakteristik berücksichtigt für die Analyse der Wortvielfalt das gesamte Spektrum aller Wörter. Je öfter sich Wörter wiederholen, desto größer ist der Wert für K.

Yule's K (1944) [61]:

$$K = 10^4 \cdot \frac{(\sum_{i=1}^{V} i^2 V_i) - V}{V^2} \tag{3.3}$$

Nach Studien von Tweedie und Baayen [55] sind all diese VRMs abhängig von der Textlänge. Sie sollen sogar nach Stamatatos et. al [51] sehr instabil bei Textlängen mit weniger als 1000 Wörtern sein. Ob und inwieweit diese VRMs trotzdem für Algorithmen der intrinsischen Plagiaterkennung, die oft auf noch kleinere Abschnitte angewendet werden, funktionieren, wird auch im Rahmen dieser Diplomarbeit untersucht.

3.2 Analyse der Wortarten

Jedes Wort hat in einem Satz eine bestimmte Funktion, die abhängig vom Kontext ist und dementsprechend einer Wortart zuordenbar ist. Mit einer Wortartanalyse, auch POS-Analyse[4] genannt, kann die globale Verteilung aller Wortarten eines Dokuments berechnet und somit die relative Frequenz jeder einzelnen repräsentiert werden. Die in dieser Arbeit verwendeten Attribute und Wortarten sind in Tabelle 3.1 zusammengestellt.

Attribut	Wortart	Attribut	Wortart
w_1	Adjektive	w_{10}	Interjektionen
w_2	Adverben	w_{11}	Modalverben
w_3	alphanumerische Zeichen	w_{12}	Substantive
w_4	Artikel	w_{13}	Präpositionen
w_5	das Wort „be“	w_{14}	Pronomen
w_6	Kopula	w_{15}	Relativpronomen
w_7	das Wort „do“	w_{16}	Symbole
w_8	Fremdwörter	w_{17}	das Wort „to“
w_9	das Wort „have“	w_{18}	Verben

Tabelle 3.1: *Attributmenge W. Die Attribute stellen die relative Frequenz der entsprechenden Wortart dar [28]*

3.3 Berechnung der Lesbarkeit

In der ersten Hälfte des 20. Jh. entstanden die ersten Lesbarkeitsformeln um Vorhersagen über die Lesbarkeit und daraus folgend über die Qualität von Schulbüchern treffen zu können. Nach DuBay [12] ist die Definition des Begriffs der Lesbarkeit von Edgar Dale und Jeanne Chall aus dem Jahre 1949 die umfassendste:

> *„The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting.“* [10]

[4] **POS:** Part of Speech

Demnach sind für das Verstehen eines Textes viele Faktoren ausschlaggebend. So tragen zu großen Teilen die jeweilige Lesekompetenz, die Lesemotivation, das Vorwissen über den Inhalt und über das Textgenre zum Verständnis bei. Die ersten drei Punkte sind leserspezifisch und können damit nicht zur Berechnung der Lesbarkeit anhand eines Textes verwendet werden. Die Lesbarkeitsformeln müssen demnach auf objektiven und quantifizierbaren Stileigenschaften basieren (siehe Abbildung 3.1).

Seit dem Beginn der Entwicklung der Lesbarkeitsformeln sind diese für die Anwendung für jegliche Art von Texten weiterentwickelt worden. Der Schwerpunkt lag dabei auf technischen Dokumenten.

Die bekanntesten Formeln zur Berechnung der Lesbarkeit sind hier in zwei Gruppen eingeteilt worden. Zum einen sind da diejenigen, deren Ergebnisse sich auf einen willkürlich festgelegten Index beziehen, zum anderen geben sie einen Lesegrad an. Der Lesegrad gibt die Anzahl von Schuljahren entsprechend des US-amerikanischen Schulsystems an, die ein Leser absolviert haben muss, um einen Text zu verstehen.

Für die Darstellung der Formeln werden folgende mathematische Definitionen vereinbahrt: Es sei ein Dokument $D = w_1, ..., w_n$ definiert, bestehend aus $w_1, ..., w_n$ Wörtern. Desweiteren sei S die Menge aller Sätze, W die Menge aller Wörter, T die Anzahl aller Silben und N die Anzahl aller Zeichen. $W_{T\geq 3}$ ist die Menge aller Wörter, die aus mehr als zwei Silben bestehen. $W_{T=1}$ ist die Menge der einsilbigen Wörter.

3.3.1 Indexgebundene Formeln

Bei diesen Formeln wird die Lesbarkeit R eines Dokuments auf einer Skala bis 100 Punkten eingestuft. Je höher dabei der Wert ist, desto verständlicher ist das Dokument.

Flesch's-Reading-Ease-Index FREI (Flesch, 1948 [14])

Als Basis zur Aufstellung dieser Formel diente Flesch „McCall-Crabbs Standard Test Lessons in Reading“ aus dem Jahre 1925. Diese Tests bestehen aus standardisierten Texten und daran folgenden Multiple-Choice-Tests, mit denen das Leseverständnis von Schülern geprüft wird. Jeder Text ist anhand von Erfahrungswerten auf einen Schuljahresgrad normiert. Somit stellen diese Tests eine gute Ausgangsposition dar, um die Lesbarkeit von Texten zu berechnen.

Besonders in Hinblick auf die Berechnung der Lesbarkeit für Texte aus dem Alltag, die nicht unbedingt schulspezifisch sind, war Flesch der Ansicht, dass für das Leseverständnis die Länge der

Wörter im allgemeinen, sowie die Satzlängen maßgebend sind. Seine zu findende Lesbarkeitsformel sollte bei den schwierigsten zu verstehenden Texten 0 und bei den leichtesten Texten, bei denen nur wenig Leseverständnis nötig ist, den Wert 100 berechnen. Die verschiedenen Schwierigkeitsstufen der „McCall-Crabbs Standard Test Lessons in Reading“ waren durch die angegebenen Schuljahresgrade bekannt. Indem Flesch die Werte für Wort- und Satzlängen einer großen Anzahl verschieden schwerer Texte ermittelte, konnte er Formel 3.4 aufstellen.

$$R = 206.835 - 1.015 \cdot \frac{W}{S} - 84.6 \cdot \frac{T}{W} \tag{3.4}$$

Zur Darstellung des Prinzips dient Tabelle 3.2, in der hypothetische Werte dreier Texte vom Verfasser zusammengestellt sind. Darauf aufbauend kann das Gleichungssystem 3.5 gebildet und die fehlenden Parameter berechnet werden.

Text	**Index**	**Satzlänge** $\frac{W}{S}$	**Wortlänge** $\frac{T}{W}$
T_1	100	3	1.227
T_2	52	9	1.716
T_3	0	20	2.205

Tabelle 3.2: *Hypothetische Werte dreier unterschiedlich schwieriger Texte.*

$$\begin{aligned} 100 &= a + 3 \cdot b + 1.227 \cdot c \\ 52 &= a + 9 \cdot b + 1.716 \cdot c \\ 0 &= a + 20 \cdot b + 2.205 \cdot c \end{aligned} \tag{3.5}$$

$$\begin{aligned} a &= 210.79 \\ b &= -0.8 \\ c &= -88.34 \end{aligned}$$

Aufgrund der Schlichtheit und einfachen Berechnungsmöglichkeit setzte sich Fleschs Formel in der amerikanischen Lesbarkeitsforschung durch.

Miyazaki-EFL[5]-Readability-Index MRI (Greenfield, 1999 [18])

Formel 3.6 wurde für Dokumente aufgestellt, die von Autoren geschrieben wurden, deren Muttersprache nicht englisch ist. Sie ist an Fleschs Formel angelegt.

$$R = 164.935 - 18.792 \cdot \frac{N}{W} - 1.916 \cdot \frac{W}{S} \tag{3.6}$$

Der Nachteil dieser beiden Formeln ist, dass der Autor die enthaltenen Stilmerkmale durchschnittliche Satzlänge und Silbenanzahl relativ einfach selbst beeinflussen kann.

3.3.2 Berechnung des Lesegrads

Die folgenden Formeln versuchen die Lesbarkeit R eines Textes anhand eines Lesegrads, also der Anzahl besuchter Schuljahre, anzugeben. Da es einen linearen Zusammenhang zwischen Lesealter und Lesegrad gibt (Lesealter = Lesegrad + 5), können alle Formeln, sowohl für die Berechnung des Lesalters, als auch des Lesegrads angewandt werden. Aus Gründen der Übersichtlichkeit berechnen alle hier beschriebenen Formeln den Lesegrad.

Dale-Chall-Index DCI (Dale und Chall, 1948 [7])

Wie Flesch dienten Dale und Chall „McCall-Crabbs Standard Test Lessons in Reading“ der Aufstellung ihrer Formel 3.7. Im Gegenteil zum FREI wurde die Anzahl der durchschnittlichen Silbenanzahl für das Leseverstehen aber als unnötig angesehen. Um eine einfache Handhabung zu gewährleisten basiert der DCI nur auf den zwei Faktoren „durchschnittliche Satzlänge“ und „Anzahl unbekannter Wörter“. Dabei wird eine Liste mit Wörtern benutzt, die Kinder der vierten Klasse kennen sollten. Es wird hierbei angenommen, dass Leser Text einfacher lesen, verstehen und wiedergeben können, wenn die Wörter bekannt sind [60]. Die ursprüngliche Liste, die 1948 mit der Formel veröffentlicht wurde, enthielt 769 einfache Wörter. Unter Verwendung dieser wurden in Untersuchungen Schwächen bezüglich der Berechnung höherer Grade festgestellt. Die Liste wurde deshalb auf 3000 Wörter erweitert und somit ist DCI allgemeiner einsetzbar.

D sei hier der prozentuale Anteil der Wörter, die sich nicht auf der Liste der 3000 Wörter befinden. Das Ergebnis ist ein Wert, der mit Hilfe von Tabelle 3.3 in den entsprechenden Lesegrad übertragen werden kann.

[5] **EFL**: English as a Foreign Language

$$G = 0.1579 \cdot D + 0.0496 \cdot \frac{W}{S} + 3.6365 \tag{3.7}$$

DCI G	**Lesegrad**
<= 4,5	<= 4
5,0 - 5,9	5 - 6
6,0 - 6,9	7 - 8
7,0 - 7,9	9 - 10
8,0 - 8,9	11 - 12
9,0 - 9,9	13 - 15
>= 10,0	>= 16

Tabelle 3.3: *Tabelle zur Übertragung des DCIs G auf den Lesegrad*

Gunning-Fog-Index GFI (Gunning, 1952 [19])

Die Formel 3.8 ist dem FREI ähnlich. Anstatt der Berechnung der durchschnittlichen Silbenanzahl schlug Gunning vor, die Wörter mit drei oder mehr Silben zu verwenden. Er nannte diese „hard words“. Neben dem prozentualen Anteil dieser basiert die Formel außerdem auf der durchschnittlichen Satzlänge.

$$R = \left(\frac{W}{S} + 100 \cdot \frac{W_{T \geq 3}}{W} \right) \cdot 0.4 \tag{3.8}$$

Powers-Sumner-Kearl-Grade PSKG (Powers et al., 1958 [40])

Die Lesbarkeit von Büchern wird hier für Leser zwischen 7 und 10 Jahren gemessen. Deswegen nimmt das Ergebnis bei Anwendung auf allgemeinen Texten auch meist einen kleineren Wert an, als die Ergebnisse der restlichen Formeln. Die eigentliche PSKF verwendet zur Berechnung Textabschnitte bestehend aus 100 Wörtern, sodass hier $W = 100$ ist:

$$R = \frac{W}{S} \cdot 0.0778 + T \cdot 0.0455 - 2.2029$$

Die Formel für beliebige Wortanzahl lautet daraus folgend:

$$R = \frac{W}{S} \cdot 0.0778 + \frac{T}{W} \cdot 4.55 - 2.2029 \tag{3.9}$$

The-Army's-Automated-Readability-Index ARI (Smith und Senter, 1967 [46])

Die Gültigkeit des ARI in Formel 3.10 auf allgemein technische Dokumente wurde 1970 von Smith und Kincaid erfolgreich überprüft.

$$R = 4.71 \cdot \frac{N}{W} + 0.5 \cdot \frac{W}{S} - 21.43 \tag{3.10}$$

Simple Measure of Grabbledygook SMOG (McLaughlin, 1969 [32])

Formel 3.11 wurde auch entwickelt, um Bibliothekaren beim einfacheren Kategorisieren des Bestandes in „Anfänger und Fortgeschrittene“ zu unterstützen [37].

$$R = \sqrt{\frac{W_{T\geq 3}}{S} \cdot 30} + 3 \tag{3.11}$$

FORCAST (Caylor et al., 1973 [6])

Formel 3.12 wurde besonders für die Anwendung auf Dokumente technischen Inhalts aufgestellt. Deren Berechnung der Lesbarkeit erfolgt abschnittsweise für 150 Wörter ($W = 150$):

$$R = 20 - \frac{W_{T=1}}{10} \tag{3.12}$$

Die Formel zur Anwendung auf Texte beliebiger Wortlänge lautet dementsprechend:

$$R = 20 - \frac{W_{T=1}}{W} \cdot 15 \tag{3.13}$$

Fry-Readability-Graph (Fry, 1969 [16])

Nach Berechnung der durchschnittlichen Anzahl von Sätzen und Silben pro 100 Wörter kann der sich daraus ergebene Lesegrad an einem Graphen abgelesen werden. Das Funktionieren des Fry-Graphs wurde anhand Lesematerials für alle Lesegrade bestätigt. Neben einem Graphen für den Grad (siehe Abb. 3.3) gibt es einen entsprechenden zweiten, der das Lesalter angibt.

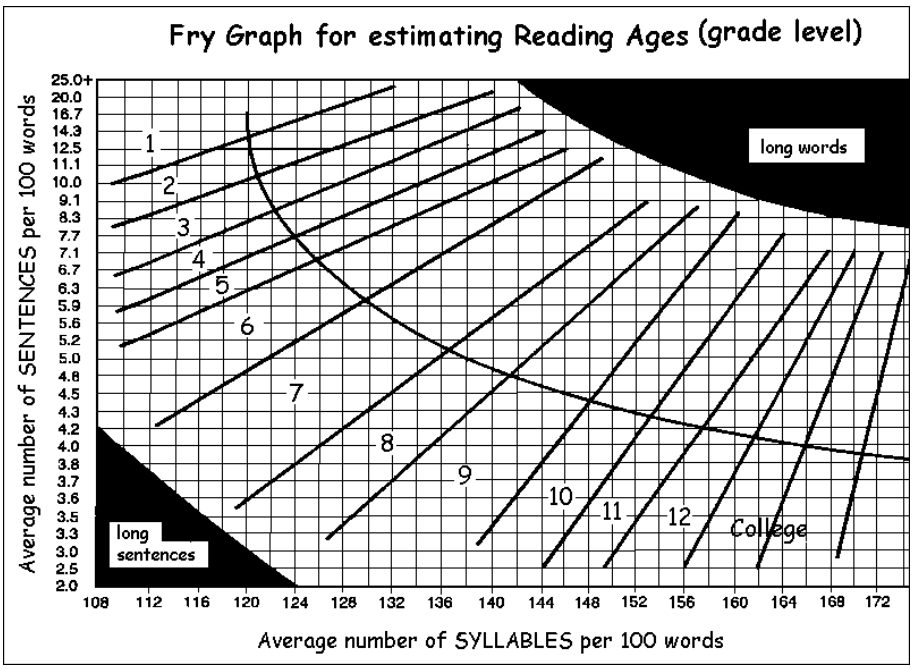


Abbildung 3.3: *Fry Graph für die Bestimmung des Lesegrads [17]*

Flesch-Kincaid-Grade-Level FKGL (Kincaid et al., 1975 [27])

Formel 3.14 ist eine Weiterentwicklung des FREI.

$$R = 0.39 \cdot \frac{W}{S} + 11.8 \cdot \frac{T}{W} - 15.59 \tag{3.14}$$

Coleman-Liau-Grade-Level CLGL (Coleman und Liau, 1975 [8])

$$R = 5.88 \cdot \frac{N}{W} - 29.59 \cdot \frac{S}{W} - 15.8 \tag{3.15}$$

3.3.3 Vergleich der Lesbarkeitsformeln

Zum besseren Vergleich der Formeln wurden anhand folgenden Satzes die jeweiligen Lesbarkeitsgrade der relevanten Formeln, dargestellt in Tabelle 3.4, berechnet:

> *„We will combine a number of pairwise authorship attribution experiments in order to solve the Verification problem.“* [30]

Die aus diesem Satz berechneten Parameter sind: $S = 1$, $W = 17$, $W_{T\geq 3} = 4$, $W_{T=1} = 8$, $T = 34$, $N = 98$, $D = 52,94$

Formel	Grad	Formel	Grad
FREI	20,38	SMOG	13,95
MRI	24,03	FORCAST	12,94
GFI	16,21	FKGL	14,64
PSKG	8,22	CLGL	16,36
ARI	14,22	Dale-Chall-Grade	>=16

Tabelle 3.4: *Berechnung der Lesbarkeitswerte anhand eines Beispielsatzes.*

Ausgenommen FREI und MRI sollen alle für dieses Beispiel verwendeten Lesbarkeitsformeln die allgemeine Anzahl der Schuljahre als Lesegrad angeben. Ohne PSKG, der eigentlich nur für Texte für Grundschüler entwickelt worden ist, variieren diese Ergebnisse bei dem Beispielsatz zwischen 12,94 und 16,36. Durch diese großen Abweichungen untereinander sind die Ergebnisse der Formeln schwierig miteinander vergleichbar. Sie erfüllen nicht das Kriterium der Validität[6]. Für die Verwendung dieser als Methoden zur Plagiaterkennung ist es aber nicht notwendig, dass die Werte der verschiedenen Formeln gleich sind. Viel wichtiger ist die Gleichheit, bzw. „große Ähnlichkeit“ der Werte einer Lesbarkeitsformel angewendet auf einzelne Abschnitte eines Dokuments eines Autors. Das ist in Abhängigkeit des Textes nach Redish [41] nicht der Fall. Für die Anwendung bei der intrinsischen Plagiaterkennung wird mittels Stabilitätstests diese Aussage überprüft.

Die Lesbarkeitsformeln sind nach Schriver [43] hinsichtlich ihrer eigentlichen Bestimmung (Berechnung des menschlichen Leseverständnisses) nicht valide und reliabel[7], da wichtige Eigenschaften, wie Lesekompetenz und -motivation, das Vorwissen über den Inhalt und das Textgenre bei der Berechnung außer acht gelassen werden. Für die Untersuchungen zur Plagiatanalyse sind diese Faktoren nicht relevant.

In Tabelle 3.6 sind alle hier vorgestellten Lesbarkeitsformeln hinsichtlich ihrer Bewertung, ihrer Minimal- und Maximalzustände zusammengefasst. Um vergleichbare Werte zu erhalten, wurden Basiswerte zur Anwendung auf alle Formeln eingeführt (siehe Tab. 3.5). Diese Basiswerte wurden

[6] Validität: „Allgemeingültigkeit“ - Ein Verfahren misst ein Merkmal oder eine Variable so, die es der Erwartung nach, messen soll. Die dahinterstehende Idee, ist die Frage der Generalisierung

[7] Reliabilität: „Zuverlässigkeit“ - Unter den gleichen Rahmenbedingungen muss bei Wiederholung eines Experiments das gleiche Ergebnis erzielt werden.

anhand einer Lesbarkeitsformel (Referenzformel) subjektiv festgelegt. So ist die durchschnittliche Wort- und Silbenanzahl so ausgewählt, dass bei der Berechnung des FREI in Tabelle 3.6 die Minimal- und Maximalwerte 0 (schwer) und 100 (leicht) ergeben.

Basiswert	**Minimum**	**Maximum**	**Referenzformel**
⊘ Wörter pro Satz	3	20	FREI
⊘ Silben pro Wort	1,227	2,205	FREI
⊘ Zeichen pro Wort	4	7	MRI
% Anteil komplexer Wörter	0	50	GFI
$\sum$ komplexer Wörter pro Satz	0	10	SMOG
% Anteil unbekannter Wörter (Dale Score)	0	100	DCI

Tabelle 3.5: *Minimale und maximale Basiswerte in Abhängigkeit einer Referenz-Lesbarkeitsformel*

Bewertung	**Lesbarkeitsformel**	**Minimum (leicht)**	**Maximum (schwer)**
Verständlichkeit/ Komplexität	Flesch-Reading-Ease-Index	100	0
	Miyazaki-Readability-Index	84	-5
	Dale-Chall-Index	4	20
Lesegrad	Flesch-Kincaid-Grade-Level	0	18
	Gunning-Fog-Index	1	28
	SMOG	3	20
	FORCAST	5	20
	Automated-Readability-Index	-1	22
	Coleman-Liau-Grade-Level	-2	24
	Powers-Sumner-Kearl-Grade	4	9
	Fry-Readability-Graph	< 1	> 14

Tabelle 3.6: *Übersicht der Lesbarkeitsformeln unter Verwendung der Basiswerte aus Tab. 3.5*

Auch hier ist zu sehen, dass große Unterschiede, besonders zwischen den Formeln zur Berechnung des Lesegrades bestehen und sogar negative Werte annehmen können. Die Minimalwerte reichen von -2 bis 5 und die Maximalwerte von 18 bis 28. Somit sind diese für den Gebrauch zur genauen Bestimmung der Lesbarkeit eines Textes nur bedingt einsetzbar.

3.4 Zusammenfassung

Zur Berechnung des individuellen Schreibstils eines Autors können Stilmerkmale verwendet werden. Die meisten dieser setzen sich aus quantifizierbaren Stileigenschaften, die zeichen-, wortbasiert oder syntaktisch sind. Um zuverlässige Ergebnisse zu erhalten sollten verschiedenen Stilmerkmale zur Berechnung des Stils eingesetzt werden.

VRMs bestimmen den Umfang und/oder die Differenziertheit des Wortschatzes. Es ist u.A. von Tweedie und Baayen [55] nachgewiesen worden, dass sich mit zunehmender Textlänge die Worthäufigkeiten ändern und dadurch die VRMs verschiedene Werte liefern. Durch diese Abhängigkeit sollten zu vergleichende Texte auch annähernd gleiche Längen haben.

Mittels Analyse der Wortarten kann die relative Frequenz der einzelnen Wortarten eines Textes bestimmt werden. Je länger dieser ist, desto zuverlässiger werden die Ergebnisse. Zu vergleichende Texte müssen aber nicht diesselbe Länge haben.

Lesbarkeitsformeln berechnen die objektive Lesbarkeit eines Textes. Es werden keine semantischen und syntaktischen Eigenschaften gemessen. Somit sollten die zu untersuchenden Texte eine bestimmte Mindestlänge haben. Je länger die Texte, desto zuverlässiger sollten die Ergebnisse sein. Nach Untersuchungen von Redish [41] sind jedoch auch Lesbarkeitsformeln nicht stabil, d.h. sie liefern bei verschieden langen Textabschnitten eines Autors unterschiedliche Werte. Ob dies auch im Rahmen der Plagiaterkennung zutrifft, ist zu untersuchen.

4 Ein Korpus zur Evaluierung von Plagiaterkennungsalgorithmen

Um intrinsische Methoden zur Plagiaterkennung testen zu können, werden große Textmengen mit darin plagiierten Stellen benötigt. Originalabschnitte und plagiierte müssen für diese Aufgabe als solche gekennzeichnet sein. Da solche Datenmengen bislang noch nicht erstellt worden sind, ist der Aufbau eines individuellen Korpus notwendig. Die Basis dafür sind tiefer gehende Kenntnisse über die Korpuslinguistik, die im ersten Unterkapitel zusammengefasst werden. Ausgangspunkt bilden dabei die Definitionen der Begriffe Korpus und Korpuslinguistik. Im Anschluss daran wird der Aufbau des Korpus anhand korpuslinguistischer Kriterien beschrieben.

4.1 Korpuslinguistische Grundlagen

Nach Paprotté [38] ist die Korpuslinguistik eine

> *„...dezidiert datenorientierte, empirische Sprachwissenschaft ‚die sich auf hinreichend große, motiviert strukturierte Mengen geeigneten Textmaterials (Korpora), [sic!] stützt, um besser begründete und vollständigere sprachwissenschaftliche Theorieentwürfe zu machen.“*

Geforscht wird mit Blick auf die Erfassung, Beschreibung und Erklärung von Phänomenen auf der Basis von Korpora.

Korpora sind große Sammlungen authentischer Texte. Diese dürfen dabei nicht in Hinblick auf das zu untersuchende Ziel verändert, bzw. auf ein erwartetes Ergebnis hin, ausgewählt worden sein. Sie sollen sozusagen „Stichproben des Sprachgebrauchs“ sein und die Forderungen nach Repräsentativität und Differenzierung erfüllen.

Des Weiteren sollte das Korpus[1] entsprechend des Untersuchungsziels zusammengestellt werden, sodass alle charakteristischen Eigenschaften des Sprachausschnitts dargestellt sind. Das Korpus sollte elektronisch verfügbar sein, um mittels statistischer Verfahren Analysen zu ermöglichen.

Anfang der 60er Jahre des 20. Jh. wurde das Brown Korpus als erstes allgmeines und maschinenlesbares Korpus an der Brown University in Rhode Island zusammengetragen. Es soll die amerikanische Schriftsprache darstellen und besteht aus ungefähr einer Million Wörtern, die sich in 500 Texte aus den verschiedensten Themengebieten aufteilen. Bis heute ist dieses Korpus prägend für viele andere und deshalb trotz seines Alters eines der wichtigsten überhaupt.

4.2 Aufbau des Korpus

Der gesamte Korpusaufbau lässt sich nach Mehler [33] zusammenfassend auf zwei wesentliche Fragen zurückführen. Zum einen ist da die nach dem Untersuchungsziel und der hiervon abhängigen Repräsentativitätsfrage. Zum anderen ist es die Frage nach Gegenstand und Verfahren der Korpusaufbereitung. Davon ausgehend werden in den folgenden beiden Kapiteln die Auswahl passender Textmengen für das Korpus und die anschließende Generierung dieses beschrieben.

4.2.1 Auswahl und Aufbereitung von Dokumenten

Das zu erstellende Korpus hat das Ziel, das Erforschen von intrinsischen Plagiaterkennungsalgorithmen unter Verwendung von Stilmerkmalen zu ermöglichen. Um dieses Ziel zu erreichen sind große Textmengen notwendig, in denen sich plagiierte Abschnitte befinden. Originale und plagiierte Abschnitte müssen voneinander unterscheidbar sein.

Ein weiterer wichtiger Punkt ist die Authentizität der Textmengen. Darum wurden reale, bereits veröffentlichte Dokumente verwendet und keine nur für den Zweck dieses Korpus automatisch generiert. Die hierfür 100 ausgewählten Dokumente stammen alle aus der ACM-Digital-Library[2] und liegen im PDF-Format vor. Nachfolgend werden diese Dokumente auch „Original-Dokumente“ genannt.

Die ausschließliche Verwendung von Original-Dokumenten aus der ACM-Bibliothek hat zwei Gründe. Der erste ist, dass eine große Anzahl und Themenvielfalt von Dokumenten aus dem Bereich der EDV hier zentral digital vorliegen und nach Begriffen durchsucht werden können. Der

[1] **Korpus:** der Korpus (Körper), das Korpus (Belegsammlung)
[2] **ACM:** Association for Computing Machinery

zweite ist, dass den meisten Dokumenten ein Digitale-Object-Identifier (DOI) zugewiesen ist. Nur Dokumente die solch eine DOI besitzen, wurden für das Korpus verwendet und unter diesem Namen gespeichert. Durch diese DOI ist der Ursprung der Dokumente immer auffindbar, selbst wenn sich Verlinkungen innerhalb ACM verändern, bzw. es den direkten Link zum Dokument nicht mehr gibt.

Der nächste Schritt war die Überführung der Original-Dokumente in ein einheitliches Format, sodass diese sowohl für den Menschen, als auch für die Maschine einfach zu verarbeiten sind. Für solche Zwecke hat sich XML etabliert, welches die hier notwendige Kodierung von Metainformationen ermöglicht. Mit Hilfe solcher Metainformationen können Textabschnitte als original oder plagiiert ausgezeichnet werden.

Zuerst mussten dafür vorerst die Original-Dokumente in TXT-Dokumente umgewandelt werden. Dafür standen die zwei Werkzeuge **pdf2txt** und **pdftotext** zur Auswahl, welche beide Bestandteile von Linuxdistributionen sind. Der große Nachteil von pdftotext ist, dass die Struktur eines zweispaltigen Dokumentaufbaus nicht erkannt wird, sodass die Reihenfolge der Textabschnitte nicht der richtigen entsprach. Somit konnte dieses Werkzeug nicht verwendet werden. Aber auch das schlussendlich verwendete pdf2txt funktioniert nicht fehlerfrei. So musste für alle Texte im Nachhinein aus folgenden Gründen eine Bereinigung stattfinden:

Es wurden meistens keine Textabsätze erkannt und zuviele leere Zeilen eingefügt. Die richtige Einteilung in Absätze ist aber für das korrekte Funktionieren der Algorithmen notwendig. Des Weiteren wurden bei einigen Dokumenten viele Wörter und Buchstaben, unabhängig von den verwendeten Schrifttypen in einem PDF-Dokument, nicht erkannt. Diese Dokumente wurden verworfen. Schlussendlich sind 100 verwendbare Dokumente anhand der Sucheingaben „Information Retrieval“ (40 Dokumente), „CSCW“[3] (30 Dokumente) und „Plagiarism“ (30 Dokumente) zusammengetragen worden. Diese drei Themengebiete werden nachfolgend auch IR, CSCW und PLAG genannt.

Der nächste Schritt war die Überführung der nun 100 bereinigten Text-Dokumente in XML-Dokumente, indem der gesamte Inhalt jedes Dokuments in die Umgebung der Tags <document> und </document>, ergänzt mit dem Attribut der Dokumentquelle (DOI), eingebettet wurde. Für das Testen der Plagiaterkennungmethoden wurden jedem anschließend fünf Textabschnitte mit Längen zwischen einem Satz und mehreren Absätzen aus anderen Dokumenten der ACM-Bibliothek hinzugefügt und mit <inserted>-Tags ausgezeichnet. Diesen Abschnitten wurde jeweils als Metainformation seine DOI und ein Plagiatstyp zugewiesen. Dieser Typ ist „copied“ oder „modi-

[3] **CSCW**: Computer Supported Cooperative Work

fied“, was den Abschnitt als kopiert oder modifiziert plagiiert auszeichnet. Abbildung 4.1 zeigt eine schematische Darstellung eines mit plagiierten Abschnitten versehenen Template-Dokuments.

```
<?xml version="1.0" encoding="UTF-8"?>
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:noNamespaceSchemaLocation="collection_schema.xsd"
 documentSource="http://portal.acm.org/citation.cfm?doid=239041.239045">

 A Probabilistic Relational Algebra for the Integration of
 Information Retrieval and Database Systems NORBERT FUHR and
 THOMAS RO" LLEKE University of Dortmund

 We present
 ...

 <inserted
   source="http://portal.acm.org/citation.cfm?doid=375212.375229"
   type="modified">
   One of our important characteristics is to use domain ontology to
   search for domain terms and its ontological relationship, within
   arious domains at different levels of abstraction. We has to
   capture the ontological model of each mediator and associate
   ...

 </inserted>

 ...
 to be feasible. In order to arrive at a model which can both be
 implemented and which is applicable in practice, one has to take a
 ...

 <inserted source="..." type="...
   ...
 </inserted>

 ...
 ...
 ...

...

</document>
```

Abbildung 4.1: *Beispiel eines Template-Dokuments. Dunkel hinterlegte Abschnitte sind plagiiert, hell hinterlegte sind Originale.*

Die Modifizierung („modified“) eines Abschnitts erfolgte auf eine oder mehrerer der folgenden Arten:

- Hinzufügen, Löschen von Wörtern/Passagen
- Ersetzen von Wörtern/Passagen
- Zusammenfassen/Trennen von Sätzen

Es wurde besonders darauf geachtet, so „real“, wie möglich zu plagiieren. Die einem Template-Dokument hinzugefügten Abschnitte entstammen so zum Beipiel aus dem selben Themengebiet, wie das Original-Dokument selbst. Die Abschnitte wurden des Weiteren auch nicht innerhalb von Sätzen, sondern als eigene Absätze zwischen Sätzen eingefügt. Prinzipiell ist es für die in dieser Arbeit verwendeten Algorithmen egal, an welchen Stellen Absätze eingefügt sind. Trotzdem wurde auch hier weitestgehend auf inhaltliche Übereinstimmung geachtet, d.h. innerhalb

des Kapitels „Introduction“ des Original-Dokuments wurden auch hauptsächlich Abschnitte aus „Introduction“-Kapiteln anderer Dokumente eingefügt.

Die modifizierten Abschnitte sind grammatikalisch korrekt. Es wurde dabei aber nicht auf die Richtigkeit des veränderten Inhalts geachtet. So wurden teilweise zum Beispiel Fachbegriffe oder spezielle Algorithmennamen mit Phrasen wie „the system“ oder „this algorithm“ ersetzt.

Zu Beginn der Datenaufbereitung bestand die Idee eines weiteren Plagiatstyps neben „copied“ und “modified“. Dieser Typ „translated“ war für Abschnitte gedacht, die übersetzt und dann eingefügt wurden. Aus Gründen des Aufbereitungaufwandes taucht dieser nicht im Korpus auf.

Tabelle 4.1 fasst die Attribute und deren Typen zusammen. Das den Template-Dokumenten zugrunde liegende XML-Schema ist in Abbildung 4.2 dargestellt.

Element	**Attribute**	**Attributtypen**
document	documentSource	String
inserted	source	String
	type	copied, modified, translated

Tabelle 4.1: *XML-Schema: Übersicht der Elemente und Attribute*

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="document">
    <xs:complexType mixed="true">
      <xs:choice maxOccurs="unbounded">
        <xs:element ref="inserted" minOccurs="0" maxOccurs="unbounded"/>
      </xs:choice>
      <xs:attribute name="documentSource" type="xs:string" use="required"/>
   </xs:complexType>
  </xs:element>
  <xs:element name="inserted">
    <xs:complexType mixed="true">
      <xs:attribute name="source" type="xs:string" use="required"/>
      <xs:attribute name="type" use="required">
        <xs:simpleType>
          <xs:restriction base="xs:string">
            <xs:enumeration value="copied"/>
            <xs:enumeration value="modified"/>
            <xs:enumeration value="translated"/>
          </xs:restriction>
        </xs:simpleType>
      </xs:attribute>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Abbildung 4.2: *Das den Template-Dokumenten des Korpus zugrunde liegende Schema.*

4.2.2 Generierung des Korpus

Für jedes der 100 Dokumente ist ein Verzeichnis mit dem DOI als Verzeichnisnamen angelegt worden, zum Beipiel „1234.1234“. In diesem Verzeichnis befindet sich wiederum ein Verzeichnis „construction“. In diesem liegen das Original-Dokument „1234.1234.pdf“, das mittels pdf2txt umgewandelte Text-Dokument „1234.1234.txt“ und das für die anschließende Generierung des finalen Korpus notwendige Template-Dokument „1234.1234.xml“. In Abbildung 4.3 ist diese Struktur abgebildet.

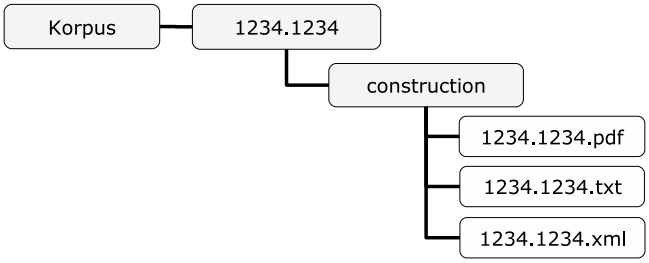


Abbildung 4.3: *Verzeichnisbaum der aufbereiteten Dokumente vor der Korpusgenerierung*

Das Ziel ist jetzt die Erzeugung einer Version aus den Template-Dokumenten, frei von XML-Tags, um das Testen der Algorithmen zur Plagiaterkennung zu ermöglichen. Um einen möglichst breiten Bereich über den plagiierten Anteil eines Textes zu erhalten, werden aus einem Template-Dokument, das k plagiierte Abschnitte enthält, 2^k Korpusdokumente durch Variierung der Anzahl der plagiierten Abschnitte generiert. Aus einem Template-Dokument mit fünf plagiierten Abschnitten wurden so 32 Instance-Dateien erzeugt. In diesen Dateien sind nur Abschnitte gespeichert, die länger als 200 Zeichen sind. Somit wird ausgeschlossen, dass Überschriften und andere kurze Fragmente als Abschnitt gelten, was die Untersuchungsergebnisse beeinflussen könnte.

Eine dieser nun tag-freien Dateien heißt beispielsweise „instance-01110.txt“. An diesem Namen ist die Anzahl der plagiierten Stellen von überhaupt möglichen (drei von fünf) ablesbar, des Weiteren welche dieser plagiierten Abschnitte im Text vorhanden (zweiter, dritter und vierter) und welche nicht vorhanden (erster und fünfter) sind. In diesen Instance-Dateien fehlen aber die Informationen über die Positionen der plagiierten Stellen. Deshalb wird zusätzliche zu jeder Instance-Datei eine Meta-Datei generiert („instance-01110.txt.meta“), in denen diese Informationen gespeichert sind (siehe Abbildung 4.4).

Diese Informationen sind Angaben über die Anzahl und den Typ der plagiierten Abschnitte, die Anzahl der originalen Textabschnitte, Angaben über deren Positionen (von/bis zu welcher Zeile), sowie der daraus resultierende plagiierte Anteil des gesamten Textes. Mit Hilfe dieser Informa-

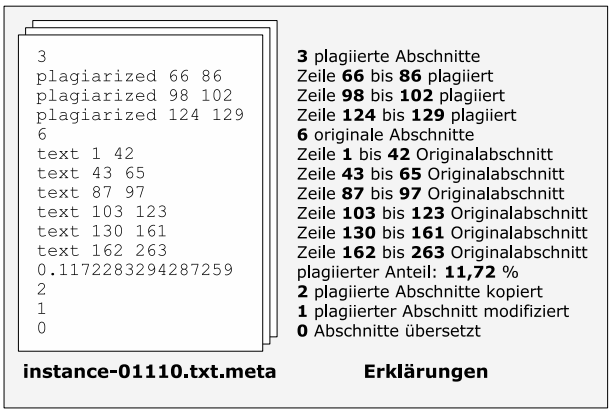


Abbildung 4.4: *Aufbau einer Meta-Datei*

tionen können dann die Ergebnisse der Plagiaterkennungsalgorithmen überprüft und verglichen werden.

Die Berechnung des plagiierten Anteils erfolgte zeichenbasiert, wobei Leerzeichen ausgelassen und Absätze mit weniger als 200 Zeichen nicht berücksichtigt wurden. Die generierten Instance- und Meta-Dateien werden in einem Verzeichnis „generated“ gespeichert, das für jedes Dokument im Dokumentverzeichnis („1234.1234“) angelegt wird. In Abbildung 4.5 ist die Verzeichnisstruktur des nun fertigen Korpus mit den neu eingeführten Begriffen anhand des Beispieldokuments 1234.1234.pdf dargestellt.

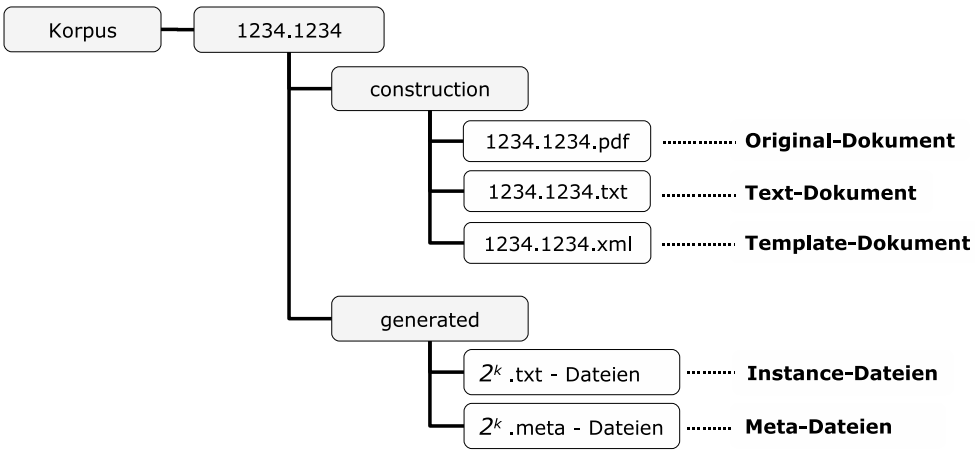


Abbildung 4.5: *Verzeichnisbaum des Korpus anhand des Beispieldokuments „1234.1234“.*

4.3 Zusammenfassung

Aufgrund der Tatsache, dass bisher kein Korpus zur Evaluierung von intrinsischen Plagiaterkennungsalgorithmen existierte, entstand unter der Berücksichtigung allgemeiner korpuslinguistischer Kriterien entsprechender. Die wichtigsten Kriterien und deren Umsetzung im Korpus sind in Tabelle 4.2 zusammengefasst.

Kriterium	**Lösung**
Authentizität	Verwendung wissenschaftlicher Veröffentlichungen, keine nachträgliche Erstellung von Dokumenten zum Zwecke des Korpus
Repräsentativität	Große Anzahl schriftlicher Dokumente aus drei verschiedenen Themenbereichen. Plagiierte Stellen stimmen thematisch mit Ausgangsdokument überein.
Maschinenlesbarkeit	XML-Format, Annotationen
Erweiterbarkeit	Es können sowohl neue Dokumente dem Korpus und weitere Abschnitte den Template-Dokumenten hinzugefügt, als auch neue Plagiatstypen angelegt werden.
Vollständigkeit/ Beobachtungsadäquatheit	Bei der Anwendung der Algorithmen werden sämtliche Daten des Korpus berücksichtigt.
Einfachheit	Kein komplizierter Aufbau des XML-Schemas und daraus folgend der Dokumente, z.B.Beschränkung der Plagiatstypen auf drei.

Tabelle 4.2: *Korpuslinguistische Kriterien nach Mehler [33] und deren Umsetzung*

Der erste Schritt für den Aufbau des Korpus war die Textbeschaffung und Auswahl von PDF-Dokumenten (Original-Dokumente) aus den Themengebieten IR, PLAG und CSCW der ACM-Bibliothek. Diese wurden mittels pdf2txt in TXT-Dokumente umgewandelt (Text-Dokumente). Durch einfügen von ensprechenden Tags wurden diese ins XML-Format überführt. Es erfolgte dann das manuelle Einfügen von Textabschnitten aus anderen Dokumenten, welche als kopiert oder modifiziert ausgezeichnet wurden. Aus den so entstandenen Template-Dokumenten sind Instance- und Meta-Dateien generiert worden. So können die Algorithmen auf den tag-freien Instance-Dateien angewandt und deren Funktionieren mit Hilfe der Meta-Dateien überprüft werden. In Abbildung 4.6 ist dieser Vorgang der Korpusgenerierung bildlich veranschaulicht.

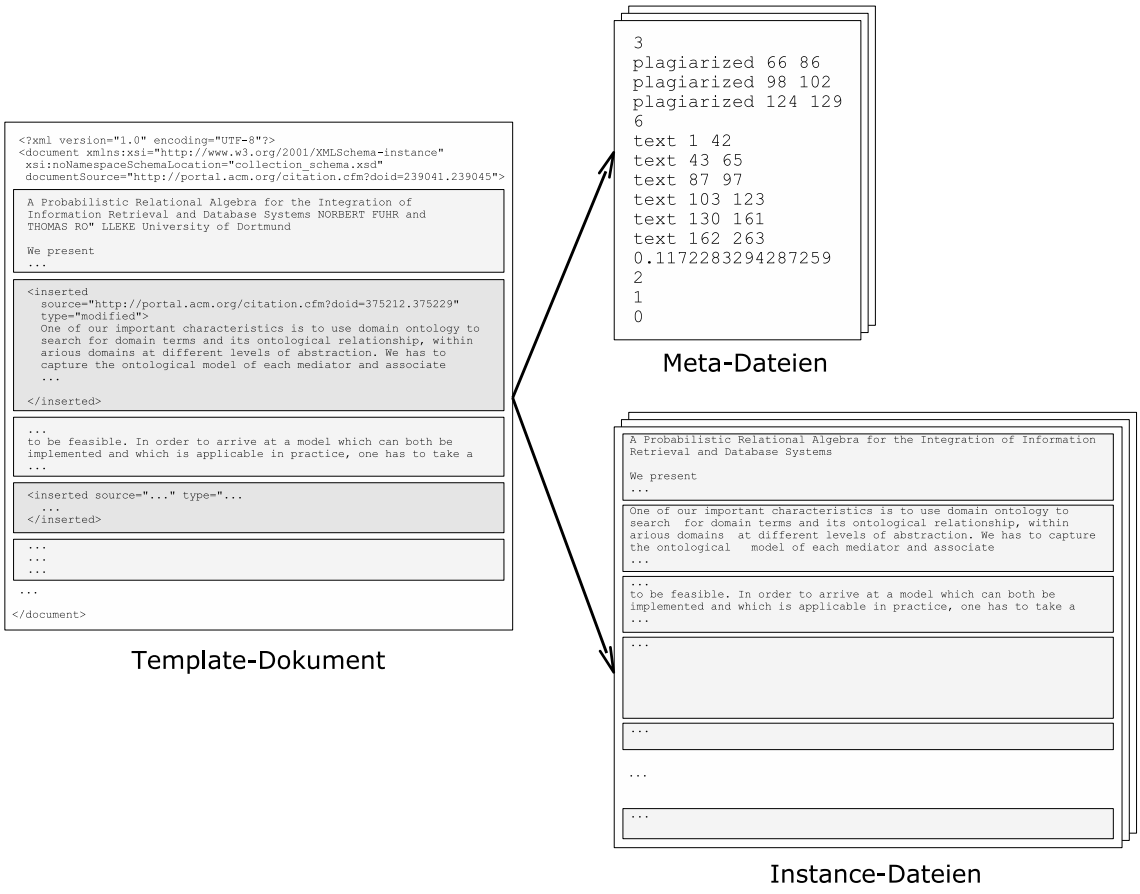


Abbildung 4.6: *Generierung des Korpus. Aus einem Template-Dokument mit k plagiierten Abschnitten werden 2^k Instance- und 2^k Meta-Dateien generiert*

5 Evaluierung

Dieses Kapitel folgt der klassischen Vorgehensweise einer Evaluierung. Nach der Formulierung von Fragen wird der Experimentaufbau zur Beantwortung dieser erklärt. Der Teil des Experimentaufbaus beinhaltet die Zusammenfassung der für die Plagiaterkennung verwendeten Stilmerkmale und die genauere Beschreibung der neu dafür verwendeten Maße „Kullback-Leibler-Divergenz“ und „Average-Word-Frequency-Class“. Des Weiteren wird die Generierung sog. Features beschrieben. Die Datenauswertung des Experiments zur Plagiaterkennung erfolgte mittels Diskriminanzanalyse und Support-Vektor-Maschine. Diese beiden Klassifizierungsverfahren sind in Kapitel 5.3 erklärt. Den Abschluss der Evaluierung bilden Schlussfolgerungen und Zusammenfassung.

5.1 Fragestellungen

1. Können plagiierte Abschnitte intrinsisch durch Einsatz von Stilmerkmalen erkannt werden?
2. Wie gut können plagiierte Abschnitte intrinsisch erkannt werden?
3. Gibt es signifikante Unterschiede in der Erkennung von Plagiaten, wenn diese kopiert oder modifiziert wurden?
4. Gibt es signifikante Unterschiede zwischen verschiedenen Themengebieten? Ist eine Themenabhängigkeit vorhanden?
5. Ab welchem plagiierten Anteil können plagiierte Abschnitte zuverlässig ermittelt werden?
6. Welche Stilmerkmale funktionieren gut, gibt es sozusagen ein bestes und ein schlechtestes?
7. Muss eine Mindestanzahl von Stilmerkmalen verwendet werden für ein zuverlässiges Funktionieren?
8. Wie stabil funktionieren die Stilmerkmale in Abhängigkeit der Textlänge?

5.2 Experimentaufbau

Die Grundlage für den Aufbau des Experiments bildet der Korpus. Für die intrinsische Plagiaterkennung werden die Schreibstile jedes Dokuments und jedes Abschnitts dieses aufgrund ausgewählter Stilmerkmale berechnet. Wenn der Schreibstil eines Abschnitts vom Schreibstil des gesamten Textes signifikant abweicht, ist dieser Abschnitt plagiiert. Für die Berechnung dieser wurde eine Auswahl von Stilmerkmalen aus Kapitel 3 verwendet. Tabelle 5.1 fasst die für die Experimente verwendeten zusammen. All diese Formeln beziehen sich auf syntaktische Eigenschaften von Texten und sind einfach zu berechnen.

Neben diesen Stilmerkmalen werden zwei neue Maße eingesetzt. Zum einen ist das die Kullback-Leibler-Divergenz [28], die in Abschnitt 5.2.1 beschrieben wird. Zum anderen wird die Average-Word-Frequency-Class [36] in Abschnitt 5.2.2 vorgestellt.

Im dritten Abschnitt dieses Kapitels wird die Generierung von Features mittels des Korpus erklärt.

Maß	**Stilmerkmal**	**Verweis**
einfache Stilmerkmale	durchschnittliche Satzlänge	
	durchschnittlicher Stoppwortanteil	
	durchschnittliche Silbenanzahl pro Wort	
Lesbarkeit	Flesch-Reading-Ease-Index FREI	Formel 3.4
	Flesch-Kincaid-Grade-Level FKGL	Formel 3.14
	Gunning-Fog-Index GFI	Formel 3.8
	Dale-Chall-Index DCI	Formel 3.7
Wortvielfalt/ VRMs	Honoré's R	Formel 3.1
	Yule's K	Formel 3.3

Tabelle 5.1: *Übersicht der implementierten und für die Evaluierung verwendeten Stilmerkmale*

5.2.1 Kullback-Leibler-Divergenz

Die Kullback-Leibler-Divergenz KLD, benannt nach Solomon Kullback und Richard Leibler, bezeichnet ein Maß für die Unterschiedlichkeit zweier Wahrscheinlichkeitsverteilungen desselben Ereignishorizonts. Sie berechnet die „Distanz" zweier Verteilungen.

Schreibstiländerungen eines Textes können nicht direkt gemessen werden. Eine indirekte Möglichkeit ist die Untersuchung der Verteilung der Wortarten. Dafür wird der Text in Abschnitte eingeteilt und von jedem dieser Abschnitte die relative Häufigkeitsverteilung der Wortarten berechnet. Diese

lokale Verteilung wird mit der globalen relativen Häufigkeitsverteilung der Wortarten im gesamten Text verglichen. Wenn nun Abweichungen der Wortartverteilung eines Abschnitts von der des Textes festgestellt werden können, deutet dies auf einen anderen Schreibstil.

Mit der Kullback-Leibler-Divergenz können diese Abweichungen berechnet werden. In dieser Arbeit wurden die lokale und die globale Häufigkeitsverteilung der 18, in Tabelle 3.1 dargestellten, Wortarten verwendet.

Unter der Voraussetzung, dass W die Menge der 18 Wortarten und $p(w)$ ($q(w)$) die lokale (globale) Verteilung jedes $w \in W$ bezeichnet, ist die Kullback-Leibler-Divergenz definiert als:

$$KL_W(p,q) = \sum_{w \in W} p(w) \log \frac{p(w)}{q(w)} \tag{5.1}$$

Für die Umsetzung dieser Methode müssen die Wortarten eines Textes bestimmt werden. Das geschieht mit Hilfe sog. Part-Of-Speech-Tagger. Für die Implementierung wurde hier der POS-Tagger QTag von Oliver Mason [31] verwendet. Dieser liefert unter Einsatz statistischer Methoden die Wortart jedes Wortes eines Textes. Die Vorteile dieses Taggers, in Evaluierungen durch Mason untersucht, liegen in seiner Robustheit und guten Annotationsgenauigkeit.

5.2.2 Average-Word-Frequency-Class

Es ist bekannt, dass in natürlichen Sprachen einzelne Wörter in unterschiedlicher Häufigkeit auftreten. Wenige Wörter kommen dabei eher häufig und viele Wörter selten vor. Durch Häufigkeitsklassen kann dieser Zusammenhang dargestellt werden. Die Häufigkeitsklasse eines Wortes (Word-Frequency-Class) entpricht einer dem Zipf'schen Gesetz[1] folgenden Verteilung.

Sei C ein Textkorpus und $|C|$ die Wortanzahl in C. Des Weiteren sei $f(w)$ die Worthäufigkeit eines Wortes $w \in C$ und $r(w)$ der Rang von w in einer Wortliste von C, welche nach absteigender Häufigkeit sortiert ist. Auf die Häufigkeitsverteilung von Wörtern bezogen, sagt das Gesetz in seiner ältesten Form [63] aus, dass das Produkt von $r(w)$ und $f(w)$ konstant sei. Die Häufigkeitsklasse $c(w)$ eines Wortes $w \in C$ ist entsprechend [56] wie folgt definiert:

$$c(w) = \lfloor log_2(f(w^*)/f(w)) \rfloor \tag{5.2}$$

w^* bezeichnet dabei das häufigste Wort in C. In der „Sydney Morning Herald Word Database“[2]

[1] Das Zipf'sche Gesetz wurde nach dem Harvard-Professor für Linguistik George Kingsley Zipf (1902-1950) benannt.

[2] In der kostenfreien „Sydney Morning Herald Word Database“ sind alle Wörter aus allen im Jahre 1994 erschienenen Ausgaben von „The Sydney Morning Herald“ alphabetisch aufgelistet. Die 97031 gefilterten Wörter besitzen zusätzliche Angaben zur Wortfrequenz und -dichte.

von Simon Dennis [11], die für diese Experimente verwendet wird, ist das Wort „the“ das häufigste und entpricht somit der Häufigkeitsklasse 0. Das seltenste hat die Häufigkeitsklasse 19.

Nach Meyer zu Eißen und Stein [35] macht die durchschnittliche Worthäufigkeitsklasse (Average-Word-Frequency-Class AWFC) eines Dokuments Aussagen über die Stilkomplexität und die Größe des Wortschatzes eines Autors, was sehr individuelle Eigenschaften sind.

5.2.3 Generierung der Features

Vom vorliegenden Korpus ausgehend, wird für jeden Abschnitt jeder Instance-Datei eine sog. Feature-Datei generiert, in denen Informationen über die jeweiligen Stilmerkmale gespeichert sind (siehe Abbildung 5.1).

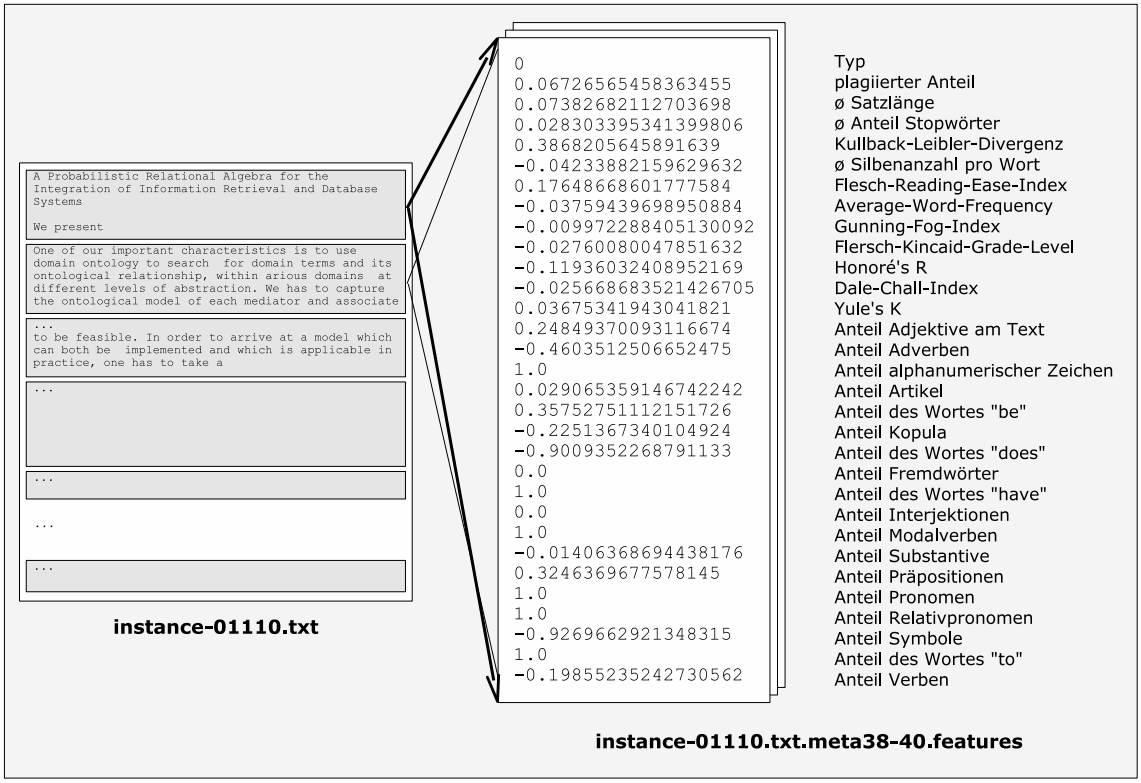


Abbildung 5.1: *Aufbau der Feature-Dateien*

In der ersten Zeile jeder Feature-Datei wird angegeben, ob der entsprechende Abschnitt plagiiert ist oder nicht. Dabei heißt 0 nicht plagiiert und 1 plagiiert. In der zweiten Zeile wird der plagiierte Anteil des gesamten Instance-Dokuments gespeichert. Sowohl die Angabe über den Plagiatstyp,

als auch die Information über den plagiierten Anteil werden aus der dazugehörigen Meta-Datei entnommen.

Die Kullback-Leibler-Divergenz wird nach Formel 5.1 berechnet. Alle anderen Werte der Stilmerkmale werden als Verhältnis des aktuellen Abschnitts zum Gesamttext in den Feature-Dateien gespeichert. Jeder dieser Werte wird folgend als „Feature“ bezeichnet. Sei a der Stilmerkmalswert eines Instance-Dokuments und b der Wert des gleichen Stilmerkmals eines Abschnitts aus demselben Instance-Dokument. Der Wert $f(a, b)$ des Features F wird dann wie folgt normiert:

$$f(a,b) = 2 * \left(\frac{a}{a+b}\right) - 1 \;\text{ mit }\; a, b \geq 0\ ,\ a + b > 0 \tag{5.3}$$

Mit dieser Formel werden die Werte der Features auf den Wertebereich $[-1, +1]$ gelegt. Das Minimum, bzw. Maximum tritt dann auf, wenn die Unterschiede zwischen Abschnitt und Gesamttext maximal sind. Bei Gleichheit der Stilmerkmalswerte beider Texte beträgt der Wert des Features 0.

Insgesamt stehen für die Evaluierung 100 Original-Dokumente zur Verfügung. Von jedem dieser wurden 2^k Instance-Dateien erzeugt, wobei wiederum jede dieser Dateien aus durchschnittlich 70 einzelnen Abschnitten besteht. So ergibt sich die ungefähre Gesamtmenge an 224.000 Feature-Dateien.

Bei geschätzten sechs plagiierten Textabschnitten pro Dokument gibt es insgesamt 600 verschiedene plagiierte. Durch die oben beschriebenen Kombinationen werden diese mehrfach verwendet, sodass schlussendlich ungefähr 13000 plagiierte Abschnitte zur Verfügung stehen. Aus Tabelle 5.2 können die genaue Anzahlen der erzeugten Feature-Dateien über die drei Themengebiete entnommen werden.

Thema	**Anzahl Abschnitte**	**davon plagiiert**
CSCW	68704	4704
Information Retrieval	99552	4449
Plagiarism	54640	3760
Gesamt	**222896**	**12913**

Tabelle 5.2: *Übersicht der Gesamtanzahl der generierten Feature-Dateien*

5.3 Statistische Klassifizierungsverfahren

Für die Evaluierung des Experiments „Plagiaterkennung“ wurde sich für eine klassische Diskriminanzanalyse und das Klassifizieren mittels einer Support-Vector-Machine (SVM) entschieden. Die Funktionsweisen der beiden Klassifizierungsverfahren werden in diesem Kapitel kurz vorgestellt.

5.3.1 Diskriminanzanalyse

Die Diskrimininanzanalyse ist ein multivariates[3] Klassifizierungsverfahren zur Analyse von Gruppenunterschieden. Von einer bestehenden Gruppierung ausgehend, wird im Nachhinein deren Qualität anhand einer Mehrzahl von Variablen untersucht. Dabei können unter anderem folgende Fragen beantwortet werden: Gibt es signifikante Gruppenunterschiede bezüglich der Variablen? Welche Variablen sind für die Unterscheidung besonders geeignet? Des Weiteren lässt sich ermitteln, zu welcher Gruppe neue Objekte aufgrund ihrer Merkmalsausprägungen klassifiziert werden können.

Voraussetzungen für die Durchführung einer Diskriminanzanalyse sind folgende Eigenschaften des Datensatzes:

- metrisch skalierte Merkmalsvariablen
- idealerweise eine Normalverteilung der verwendeten Daten
- ein mindestens doppelt so hoher Stichprobenumfang wie Anzahl an Merkmalsvariablen
- eine größere Anzahl von Merkmalsvariablen als Gruppen

Ziel ist nun die Schätzung einer Diskriminanzfunktion (Trennfunktion), die die Gruppen optimal trennt. Diese Diskriminanzfunktion hat allgemein folgende Form:

$$Y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_j x_j$$

Y ... Diskriminanzvariable
x_j ... Merkmalsvariable j
b_j ... Diskriminanzkoeffizient für Merkmalsvariable j
b_0 ... konstantes Glied

Die Diskriminanzkoeffizienten werden anhand der Werte der Merkmalsvariablen geschätzt. Die Diskriminanzfunktion liefert für jedes Element i einer Gruppe g mit den Merkmalswerten x_{jgi}

[3] Multivariate Verfahren verwenden mehr als eine abhängige Variable gemeinsam und gleichbedeutend.

einen Diskriminanzwert Y_{gi}. Jede dieser Gruppen kann duch einen mittleren Diskriminanzwert (Zentroid) beschrieben werden:

$$\overline{Y}_g = \frac{1}{I_g} \sum_{i=1}^{I_g} Y_{gi}$$

I_g ... Gesamtanzahl der Elemente in Gruppe g

Die Differenz $|\overline{Y}_A - \overline{Y}_B|$ gibt dann die Unterschiedlichkeit zweier Gruppen an. Bildlich veranschaulicht nach Huber [25] können die Werte der Diskriminanzfunktion auf einer sog. Diskriminanzachse abgetragen werden. So lassen sich sowohl einzelne Elemente, als auch die Gruppenzentroide darstellen:

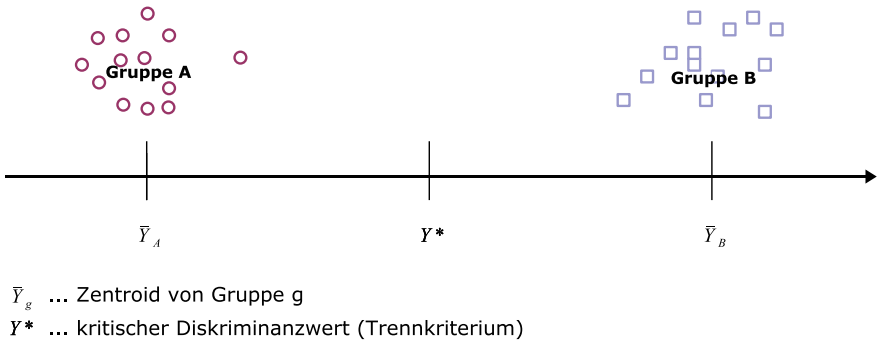


Abbildung 5.2: *Gruppenzentroide auf der Diskriminanzachse*

Y^* ist der kritische Diskriminanzwert, nach dem neue Elemente klassifiziert werden. Wenn der berechnete Diskriminanzwert Y_i eines neuen Elements i kleiner als der kritische Diskriminanzwert Y^* ist, dann wird dieses Element der Gruppe A zugeordnet, ansonsten der Gruppe B.

Die Analyse der Feature-Dateien mittels Diskriminanzanalyse bietet sich sehr gut an, da alle Voraussetzungen, die Erfolg versprechen, vorhanden sind: Die Merkmalsvariablen (29 Stilmerkmale) sind metrisch skaliert, deren Anzahl ist größer, als die Gruppenanzahl (2 Gruppen: plagiiert, nicht plagiiert) und der Stichprobenumfang (Anzahl Feature-Dateien) ist mehr als doppelt so groß, wie die Variablenanzahl.

5.3.2 Support-Vektor-Maschinen

Das Klassifikationsverfahren der Support-Vektor-Maschinen (SVM) wurde 1992 von Boser, Guyon und Vapnik [3] entwickelt. Das Prinzip des Algorithmus ist das Finden einer optimal trennenden Hyperebene in einem hochdimensionalen Merkmalsraum. Das bedeutet, dass der Abstand der zu trennenden Gruppen zu dieser Ebene maximiert werden soll. Es wird davon ausgegangen, dass solch eine Ebene immer gefunden werden kann. Abbildung 5.3 veranschaulicht das Prinzip einer SVM zweidimensional. Die dargestellten Support-Vektoren definieren dabei die Hyperebene. Die anderen Datenpunkte haben keinen Einfluss auf deren Lage.

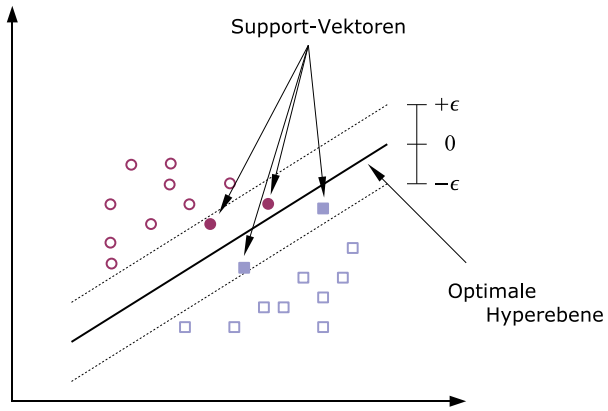


Abbildung 5.3: *Schematische Darstellung der Funktionsweise einer SVM.*

Folgende mathematischen Erklärungen und Formelschreibweisen basieren auf „A Tutorial on Support Vector Regression“ von Smola und Schölkopf [47]:

Es seien Trainingsdaten $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset X \times \mathbb{R}$ gegeben. X bezeichnet den Raum der Eingabedaten, zum Beispiel $\mathbb{R}^d$. Das Ziel ist nun eine Funktion $f(x)$ zu finden, die höchstens eine Abweichung ϵ von aktuell erhaltener Ziele y_i zu den gesamten Trainingsdaten hat. Vereinfachend lässt sich der Fall anhand linearer Funktionen f beschreiben, die folgende Form haben:

$$f(x) = \langle w, x \rangle + b \text{ mit } w \in X,\ b \in \mathbb{R} \tag{5.4}$$

$\langle \cdot, \cdot \rangle$ bezeichnen das Skalarprodukt in X. Zu beachten ist, das $f(x)$ so flach, wie möglich sein soll. Im Falle von Gleichung 5.4 heißt das ein w zu suchen, was so klein wie möglich sein muss. Schölkopf schlägt vor dafür die Euklidische Norm, beispielsweise $\|w\|^2$ zu minimieren. Dieses Problem des Findens der Hyperebene kann als konvexes Optimierungsproblem folgender Form

beschrieben werden:

$$\begin{aligned} &\text{Minimiere} \quad \tfrac{1}{2}\|w\|^2 \ , \\ &\text{sodass} \quad \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon \\ \langle w, x_i \rangle + b - y_i \leq \epsilon \end{cases} \end{aligned} \tag{5.5}$$

Die Annahme in 5.5 setzt voraus, dass es eine Funktion f gibt, die alle Paare (x_i, y_i) mit ϵ Genauigkeit approximiert. Es ist jedoch möglich, dass das nicht der Fall ist. Um dieses Optimierungsproblem für nicht linear-trennbare Gruppen trotzdem zu lösen, führen Cortes und Vapnik [9] sog. „Schlupfvariablen“ ein, die als Korrekturwerte verstanden werden können. Die daraus resultierende Formulierung des Problems kann bei Vapnik [57] gefunden werden. Für weitere detailliertere Beschreibungen sei hier auf das Tutorial von Schölkopf [47] verwiesen.

Ein großer Vorteil von SVMs ist die Verarbeitungsmöglichkeit von sehr großen Datenmengen. Dabei wird eine Überanpassung (Overfitting) an Trainigsdaten verhindert, sodass neue Daten gut klassifiziert werden können und das trotz der großen Flexibilität der Variablenanzahl. Die Nachteile liegen in der aufwändigen Parametersuche und in dem unklaren Variableneinfluss [13].

5.4 Vorexperiment - Stabilität der Stilmerkmale

Zur Plagiaterkennung ist es notwendig Stilmerkmale zu verwenden, die zuverlässig auf Textabschnittsebene funkionieren. Um dies zu gewährleisten müssen zwei Bedingungen erfüllt werden. Erstens sollte das Stilmerkmal stabil sein, d.h. im Idealfall konstante Werte unabhängig des verwendeten Textumfangs innerhalb des Textes eines Autors liefern. Zweitens sollten die Unterschiede der Werte zwischen verschiedenen Texten möglichst groß sein, um eine Unterscheidbarkeit zwischen fremden Texten zu erreichen.

Ziel dieses Experiments ist es nun herauszufinden, wie gut die verwendeten Stilmerkmale bezüglich kleiner und unterschiedlich großer Textabschnitte funktionieren. Je stabiler ein Feature über unterschiedliche Textlängen ist und je größer der Abstand der Werte zwischen verschiedenen Texten ist, desto besser sollte es für die Plagiaterkennung funktionieren.

5.4.1 Durchführung des Experiments

Die Stabilität wurde von folgenden zehn Stilmerkmalen berechnet: Durchschnittliche Satzlänge, durchschnittlicher Stoppwortanteil[4], durchschnittliche Silbenanzahl pro Wort, DCI, FKGL, FREI, GFI, AWFC, Honoré's R und Yule's K (siehe Kapitel 3).

Für die Berechnung der Stilmerkmale wurden zehn Instance-Dateien unterschiedlicher Länge aus dem Korpus verwendet. Diese enthalten keine plagiierten Abschnitte und sind bereits von Textfragmenten bereinigt, die weniger als 200 Zeichen besitzen. Die dazugehörigen Original-Dokumente tragen nur einen Autor im Titel, um zu gewährleisten, dass nur ein Schreibstil im Dokument vorhanden ist. Zu jeder dieser zehn Instance-Dateien wurden die Werte der 10 Stilmerkmale jeweils in zehnprozentigen Abstufungen von 0 bis 100 % des Textanteil und in gleichen Anteilen von je 10 % berechnet.

Da bei der intrinsischen Plagiaterkennung Textabschnitte, für die die einzelnen Stilmerkmale berechnet werden, meist kleineren Umfangs (< 10 % eines Textes) sind, müssen diese Stilmerkmale insbesondere auf kleinen Textlängen stabil sein.

Es wird erwartet, dass das neue Maß AWFC das stabilste aller Stilmerkmale ist, also die Werte über verschiedene Textlängen sehr ähnlich sein sollten. Des Weiteren sollten die Lesbarkeitsmaße besser abschneiden, als die VRMs Honoré's R und Yule's K.

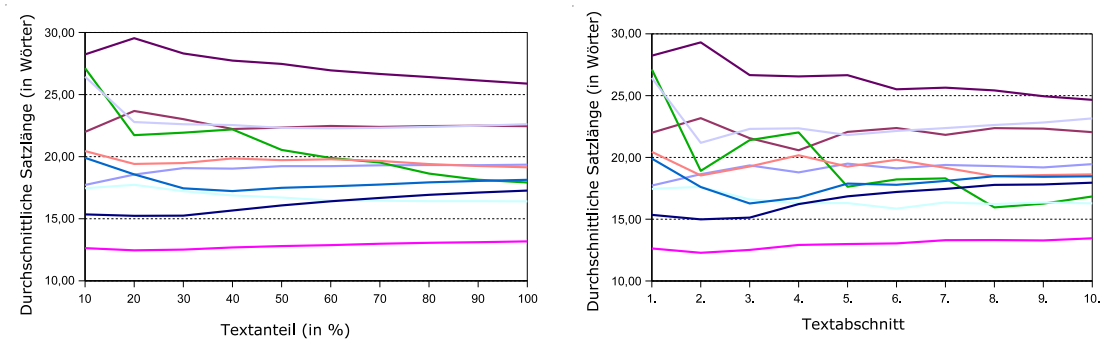
5.4.2 Auswertung der Daten

Die Auswertung der Stabilität der Stilmerkmale erfolgt nach der Gruppierung in Tabelle 5.1. Es werden einfache Stilmerkmale, Lesbarkeitsformeln und VRMs getrennt voneinander betrachtet.

a) Auswertung „einfacher Stilmerkmale“

In Diagramm (a) der Abbildung 5.4 ist zu sehen, dass das Stilmerkmal „Durchschnittliche Satzlänge“ ab einem Textanteil von 50 % stabil ist. Bei der intrinsischen Plagiaterkennung werden aber nur Textabschnitte miteinander verglichen, die 10 % des Gesamtanteils einnehmen. In (a) ist im ersten Abschnitt, der den zehnprozentigen Anteil darstellt, zu erkennen, dass bei diesen kleinen Abschnitten dieses Stilmerkmal sehr instabil ist. Diese Instabilität wird in (b) deutlich. Hier sind die Werte gleich großer Textabschnitte dargestellt.

[4] Stoppwörter sind inhaltsleere Wörter wie Artikel, Präpositionen oder Pronomina, die nicht zur inhaltlichen Beschreibung von Dokumenten verwendet werden. [54]



(a) (b)

Abbildung 5.4: *Stabilität des Stilmerkmals „Durchschnittliche Satzlänge“ von 10 Texten, 10 verschiedener Autoren. (a) Textabschnitte werden in 10 %-Schritten größer. (b) Textabschnitte entsprechen gleich großen Abschnitten von jeweils 10 % des entsprechenden Textes.*

Etwas stabiler hingegen über kleine Textabschnitte von 10 % des Gesamtanteils sind die Stilmerkmale „Durchschnittlicher Stoppwortanteil“ und „Durchschnittliche Silbenanzahl pro Wort“, dargestellt in Abbildung 5.5. In den Diagrammen zeigen sich keine großen Unterschiede zwischen den größer werdenden Abschnitten. Ähnliche Ergebnisse wurden bei den Tests mit gleicher Textlänge ermittelt.

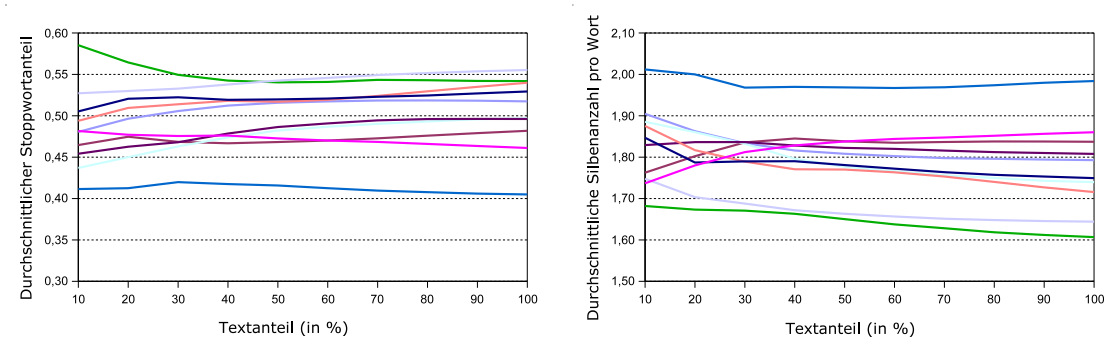
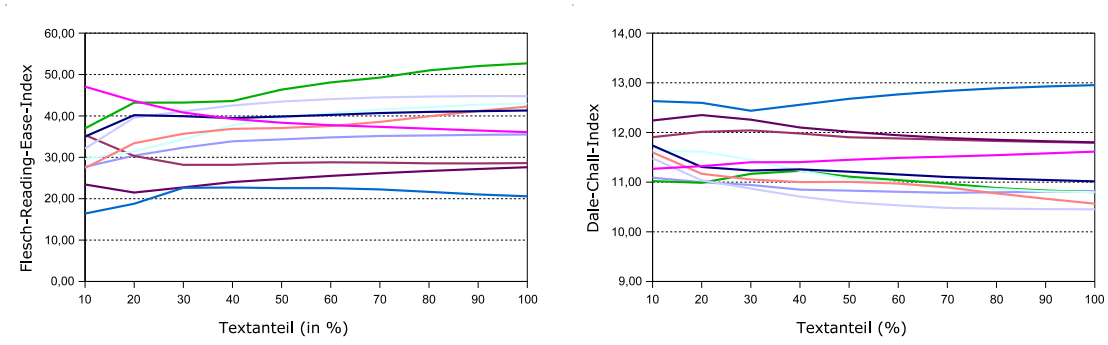


Abbildung 5.5: *Stabilität der Stilmerkmale „Durchschnittlicher Stoppwortanteil“ und „Durchschnittliche Silbenanzahl pro Wort“ von 10 Texten, 10 verschiedener Autoren bei ansteigendem Textumfang.*

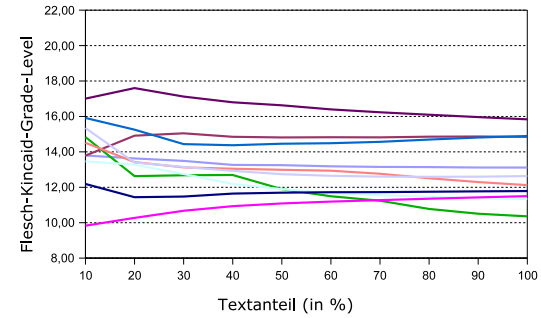
b) Auswertung Lesbarkeitsformeln

Wie in Abbildung 5.6 dargestellt, sind sowohl FREI (a), als auch FKGL (c) und GFI (d) stabil über große Abschnitte ab einem Textanteil von ungefähr 40 %. Kleine Abschnitte von 10 % zeigen große Instabilität. Die besten Ergebnisse bezüglich der Stabilität ergab DCI (b). In diesem Diagramm ist des Weiteren zu sehen, dass die Unterschiede zwischen den Werten von verschiedenen Texten klein sind. Deswegen kann nicht gesagt werden, dass dieses Stilmerkmal besser zur intrinsischen Plagiaterkennung einsetzbar wäre, als die anderen drei.

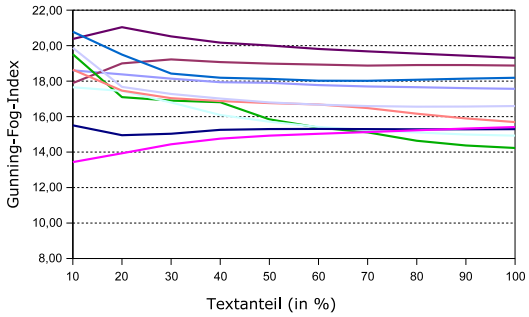


(a) Flesch-Reading-Ease-Index FREI

(b) Dale-Chall-Index DCI



(c) Flesch-Kincaid-Grade-Level FKGL

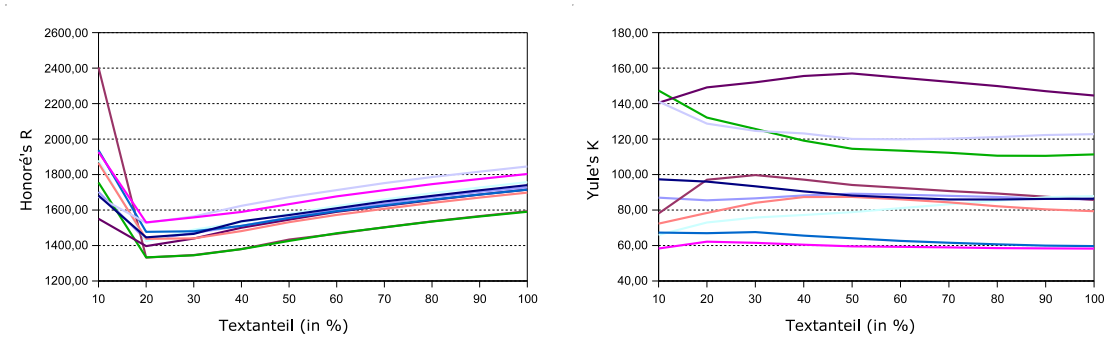


(d) Gunning-Fog-Index GFI

Abbildung 5.6: *Vergleich der Stabilität der verwendeten Lesbarkeitsformeln nach aufsteigendem Textumfang 10 verschiedener Dokumente.*

c) Auswertung VRMs

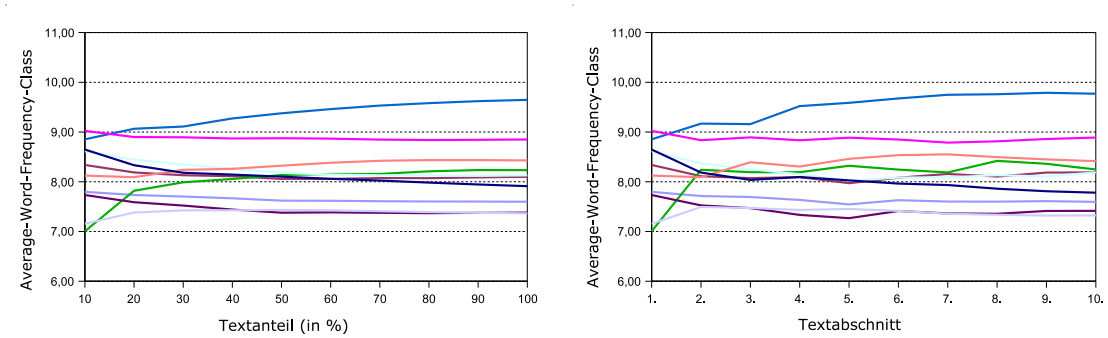
Es wurde neben den bekannten VRMs Honoré's R und Yule's K das neue Maß AWFC untersucht. In den Diagrammen der Abbildung 5.7 ist zu sehen, dass R und K am instabilsten von allen Stilmerkmalen bezüglich kleiner Textabschnitte von 10 %, als auch größerer Textanteile ab 30 % sind.



(a) Honoré's R (b) Yule's K

Abbildung 5.7: *Stabilität bekannter Vocabulary-Richness-Measures.*

Im Gegensatz dazu ist das neue Maß AWFC sehr stabil und das unabhängig vom Textumfang. So zeigen sich weniger Differenzen im Ergebnis bei der Anwendung auf größer werdende Textabschnitte und gleichbleibende. Diesen Zusammanhang zeigt Abbildung 5.8. Dieses Stilmerkmal sollte deswegen besonders gut zur intrinsischen Plagiaterkennung einsetzbar sein.

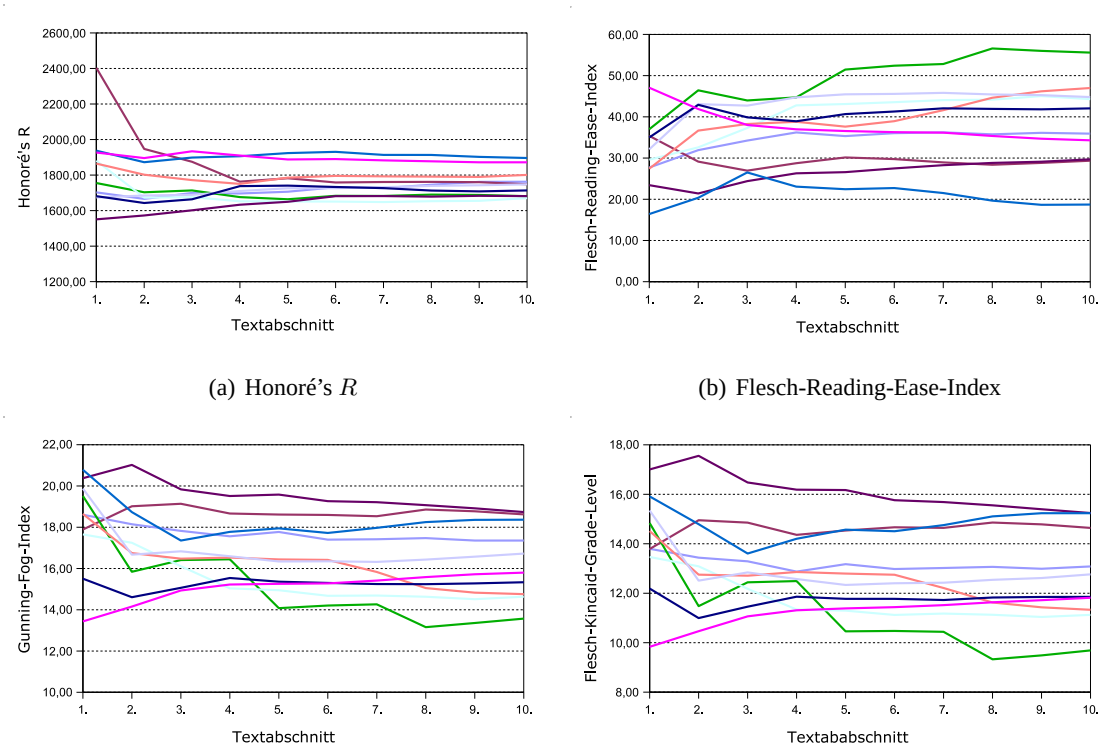


(a) (b)

Abbildung 5.8: *Stabilität des neuen VRMs AWFC. (a) zeigt diese bei steigendem Textumfang, (b) bei gleich großen Abschnitten.*

Weitere Ergebnisse, dargestellt in Abbildung 5.9 zeigen, dass das VRM Honoré's R stabiler, als

die Lesbarkeitsformeln FREI (b), GFI (c) und FKGL (d), über gleichgroße Textabschnitte ist. Die Länge dieser unter Umständen kleinen Abschnitte ist dabei weniger wichtig.



(a) Honoré's R

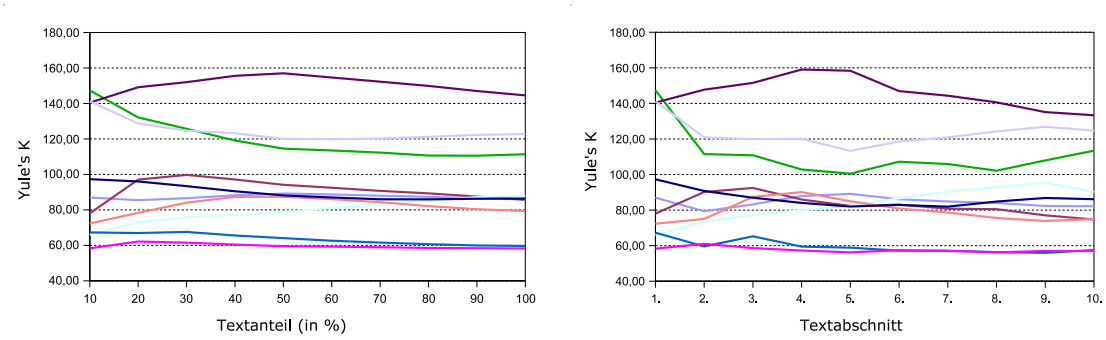
(b) Flesch-Reading-Ease-Index

(c) Gunning-Fog-Index

(d) Flesch-Kincaid-Grade-Level

Abbildung 5.9: *Vergleich der Stabilität von Honoré's R und Lesbarkeitsformeln über Textabschnitte gleicher Längen.*

Yule's K hingegen ist sowohl bei unterschiedlichen Textlängen, als auch bei kleinen Abschnitten nicht stabil, wie Abbildung 5.10 zeigt.



(a) (b)

Abbildung 5.10: *Stabilität des VRM Yule's K. Dieses Maß ist instabiler, als alle anderen, was sowohl gleiche (b), als auch unterschiedlich lange Textanteile (a) betrifft.*

5.4.3 Schlussfolgerungen

Zur intrinsischen Plagiaterkennung ist es notwendig, dass Stilmerkmale stabil funktionieren, das heißt möglichst konstante Werte bei Verwendung auf unterschiedlich langen Texten eines Autors liefern. Des Weiteren sollten Werte verschiedener Autoren möglichst verschieden sein.

Das mit großem Abstand zu allen anderen ausgewerteten, stabilste Stilmerkmal ist das das neu eingeführte VRM Average-Word-Frequency-Class. Das betrifft sowohl sehr kleine Textabschnitte, als auch unterschiedlich große.

Das stabilste Stilmerkmal unter den Lesbarkeitsformeln ist der Dale-Chall-Index. Hier liegen die Werte verschiedener Texte aber nah beieinander. Daraus kann geschlossen werden, dass dieses zur Plagiaterkennung nicht besser als die anderen funktionieren muss. Die einfachen Stilmerkmale „Durchschnittlicher Stoppwortanteil“ und „Durchschnittliche Silbenanzahl pro Wort“ sind in ihrer Stabilität mit den Lesbarkeitsformeln vergleichbar. Instabiler ist das Stilmerkmal „Durchschnittliche Satzlänge. Schlechte Stabilitätsresultate bei den getesteten Dokumenten ergaben die VRMs Honoré's R und Yule's K. Dies entspricht den Erwartungen.

Die Stilmerkmale, die stabil funktionieren, sollten auch bei der Plagiaterkennung diejenigen sein, die am besten zur Unterscheidung in plagiiert und nicht plagiiert beitragen.

5.5 Experiment - Plagiaterkennung

Dieses Experiment soll Aufschluss darüber geben, inwieweit Plagiate intrinsisch mittels Stilanalyse erkannt werden können. Für die Auswertung der vorhandenen Daten wurden die statistischen Maße Precision und Recall verwendet, welche zu Beginn dieses Kapitels erklärt werden. Diese Werte wurden mit Hilfe von den in Kapitel 5.3 vorgestellten Klassifizierungsverfahren gewonnen. Für die Diskriminanzanalyse wurde die Statistik-Software SPSS 13.0 verwendet, SVM^{light} von Joachims [26] als Implementation einer SVM. Im Anschluss wurden die ermittelten Daten, getrennt nach diesen Verfahren, ausgewertet und miteinander verglichen.

5.5.1 Precision und Recall

Precision und Recall sind die am häufigsten verwendeten Maße bei der Evaluierung von Systemen aus dem Gebiet der Information-Retrieval. Dabei stellt Precision (Genauigkeit) den Anteil der relevanten unter den gefundenen Dokumenten dar. Recall (Vollständigkeit) ist der Anteil der relevanten Dokumente, die gefunden wurden. Zur Beurteilung eines Systems müssen immer beide Werte berücksichtigt werden. Abbildung 5.11 zeigt, in welcher Beziehung beide zueinander stehen.

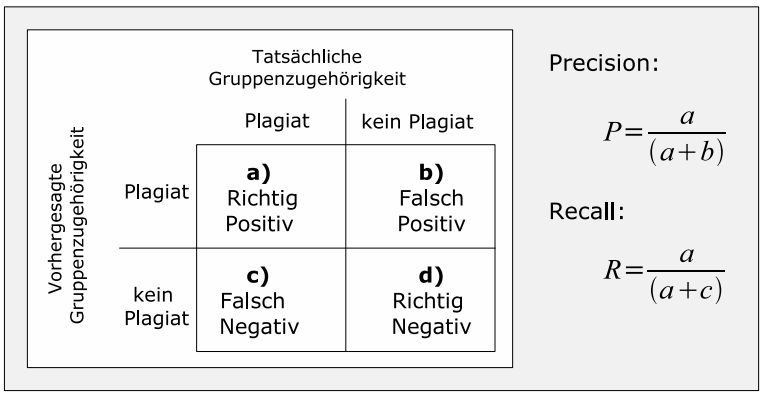


Abbildung 5.11: *Darstellung der Berechnung von Precision und Recall.*

Normalerweise verhalten sich Precision und Recall zueinander invers (siehe Abbildung 5.12). Das heißt bei einer Verkleinerung der Antwortmenge durch eine spezifischere Anfrage wird die Precision größer und der Recall kleiner. Dasselbe gilt umgekehrt.

Im Falle der Plagiaterkennung gibt der Recall-Wert den Anteil der plagiierten Abschnitte an, die tatsächlich gefunden wurden. Der Precision-Wert stellt den Anteil der tatsächlich plagiierten von den gesamten als plagiiert eingestuften Dokumenten dar.

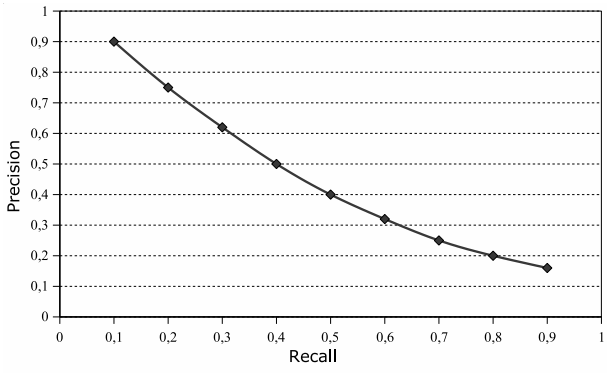


Abbildung 5.12: *Inverses Verhalten von Precision und Recall.*

5.5.2 Durchführung des Experiments

Um mit den Klassifizierungsverfahren Diskriminanzanalyse und SVM die Ergebnisse für Precision und Recall zu erhalten, mussten Eingabedaten geeigneten Formats für SPSS und SVM^{light} generiert werden. Für differenzierte Untersuchungen wurde sich für eine Einteilung nach plagiiertem Anteil eines Textes in den Abstufungen 0.03, 0.06, 0.09, 0.12, 0.15, 0.18 und 0.21 entschieden. Für jeden dieser Anteile sind jeweils eine SPSS-Datei namens „plag-style-*anteil*.spss“ und eine SVM-Dateien namens „plag-style-*anteil*-*training**Anzahl Dateien*.svm“ erzeugt worden.

In diesen wurden maximal 400 zufällig ausgewählte Feature-Dateien mit entsprechendem plagiierten Anteil in Matrixform gespeichert. 0.06 beinhaltet beispielsweise Feature-Dateien, in denen der gepeicherte plagiierte Anteil größer als 0.03 und kleiner gleich 0.06 ist. Jede Zeile entspricht einer Feature-Datei. In den Spalten sind die einzelnen Features abgetragen. In jeder Datei sind die Hälfte der Feature-Dateien plagiiert und die andere Hälfte nicht. Diese Angaben sind in der ersten Spalte gespeichert. In den SPSS-Dateien sind plagiierte Abschnitte mit 1 gekennzeichnet, nicht plagiierte mit 0. In den SVM-Dateien erfolgte dies mit 1 und -1. Der zweite Unterschied ist der, dass in den SVM-Dateien die Features aufsteigend nummeriert sind.

In Tabelle 5.3 sind die Anzahlen eingetragen, die pro plagiertem Anteil und Themengebiet verwendet wurden. Es ist zu sehen, dass es in jeder Gruppe plagiierte Anteile gibt, die die Menge von 400 Feature-Dateien nicht erreicht. Das ist der Grund weshalb „nur“ so wenige Dateien aus der eigentlich sehr großen Gesamtanzahl verwendet wurden. Die Spanne zwischen den Gruppen sollte nicht zu groß sein, da das zu Verfälschungen des Ergebnisses hätte führen können.

Die Klassifizierungen, sowohl mit Diskriminanzanalyse als auch SVM, erfolgten kreuzvalidiert, d.h. dass jede Feature-Datei durch die entsprechenden Funktionen klassifiziert wurde, die von allen

Thema	0.03	0.06	0.09	0.12	0.15	0.18	0.21
CSCW	200	400	400	400	400	400	400
Information Retrieval	400	400	400	400	400	282	170
Plagiarism	164	400	400	400	400	400	388
Gesamt	974	1200	1200	1200	1200	1200	810

Tabelle 5.3: *Gesamtanzahl der zur Auswertung verwendeten Feature-Dateien. Gesamt entspricht dabei nicht der Summe über die drei Themen, sondern der separat verwendeten Anzahl.*

anderen Dateien außer dieser abgeleitet wurden. Anhand der Klassifizierungsergebnisse die SPSS lieferte, Abb. 5.13 zeigt ein Beispiel, wurden die entsprechenden Precision- und Recall-Werte, wie im vorherigen Kapitel beschrieben, berechnet[5]. Die SVM liefert die Ergebnisse für Precision und Recall direkt.

			Vorhergesagte Gruppenzugehörigkeit		
		plag	0	1	Gesamt
Original	Anzahl	0	70	30	100
		1	17	83	100
	%	0	70,0	30,0	100,0
		1	17,0	83,0	100,0
Kreuzvalidiert	Anzahl	0	63	37	100
		1	22	78	100
	%	0	63,0	37,0	100,0
		1	22,0	78,0	100,0

Abbildung 5.13: *Klassifizierung mittels SPSS anhand einer Beispieldatei. Mit 1 markierte Anteile stellen plagiierte, mit 0 gekennzeichnete originale Textabschnitte dar.*

5.5.3 Auswertung der Daten - Diskriminanzanalyse

Der Reihenfolge der Fragestellungen in Abschnitt 5.1 folgend, werden die mittels Diskrimanzanalyse gewonnenen Daten ausgewertet.

a) Wie gut können plagiierte Stellen erkannt werden?

Plagiierte Textabschnitte können gut erkannt werden, wenn Precision und Recall beidermaßen hoch sind. Im Idealfall sollte vor Allem der Precision-Wert hoch sein, da dieser auch Auskunft

[5] Achtung, die Zeilen- und Spalteneinteilung in beiden Tabellen sind verschieden.

darüber gibt, wie hoch der Anteil der zu Unrecht als plagiiert eingestuften Abschnitte ist.

Abbildung 5.14 zeigt die Ergebnisse der Diskriminanzanalyse von Precision und Recall aller zur Verfügung stehenden Datensätze aus den drei Themengebieten. Pro plagiiertem Anteil wurden hier insgesamt 1200 Feature-Dateien verwendet. Davon ist eine Hälfte plagiiert, die andere nicht.

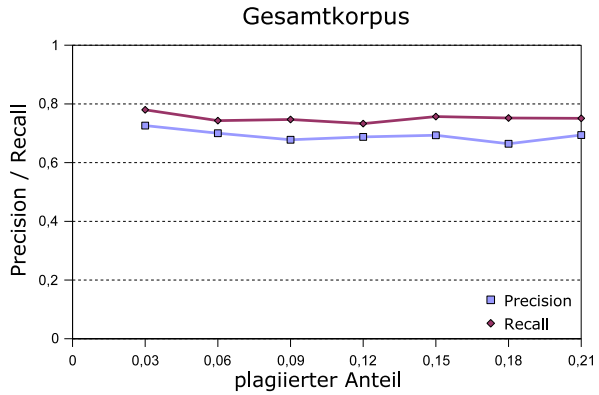


Abbildung 5.14: *Ergebnisse unter Verwendung aller Datensätze (anteilig)*

Zunächst ist zu erkennen, dass es keine signifikanten Unterschiede bezüglich der plagiierten Anteile gibt. Die Werte für Precision pendeln um 70 %, die Werte für Recall um 75 %. Plagiierte Abschnitte, die aus Texten mit nur 3 % plagiiertem Anteil stammen, werden genauso gut erkannt, wie solche aus einem Anteil von 21 %.

Ein weiterer Test sollte zeigen, ob es mögliche Unterschiede in der Erkennung zwischen modifizierten und kopierten Abschnitten gibt. Das war nicht der Fall und ist darauf zurückzuführen, dass auch die modifizierten Abschnitte einen anderen Schreibstil haben, als die kopierten und nicht auf den Schreibstil des umgegebenden Dokuments hin angepasst werden können.

b) Ist eine Themenabhängigkeit vorhanden?

Wenn eine Themenabhängigkeit vorhanden ist, wäre das sehr schlecht. Das hieße, dass je nach verwendetem Thema oder auch Genre Plagiate unterschiedlich gut erkannt werden können und somit die Ergebnisse nicht verallgemeinerbar wären.

Um dies zu testen, haben die verwendeten Original-Dokumente ihren Ursprung aus drei verschiedenen Themengebieten und wurden unter Verwendung dieser plagiiert. Die Ergebnisse der getrennten Analyse sind in Abbildung 5.15 zu sehen. Die Themengebiete CSCW und PLAG zeigen dabei vergleichbare Ergebnisse in Precision und Recall mit denen aus der Untersuchung aller

Dokumente. Das Thema IR weicht etwas davon ab. Da es nicht nachvollziehbar ist, inwieweit Plagiate aus Texten mit 3 % plagiiertem Gesamtanteil besser erkannt werden sollten, als Abschnitte mit dreifachem Anteil von 9 %, sind Unterschiede auf die Anzahl und Auswahl der Dokumente zurückzuführen.

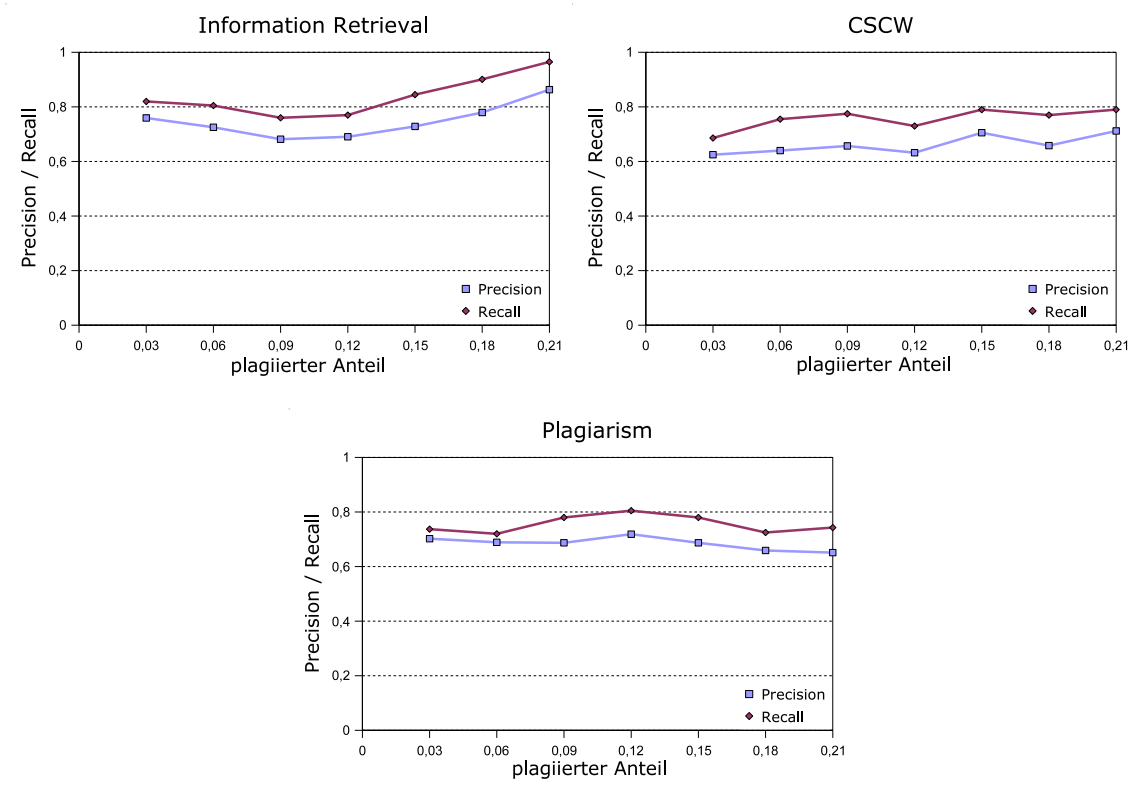


Abbildung 5.15: *Unterteilung in drei Gruppen*

Es bestand die Vermutung, dass bei Dokumenten, die mehr als einen Autor angeben, auch mehrere am Dokument gearbeitet haben. Somit träten schon von vornherein mehrere Schreibstile auf, was dann im Nachhinein zu verfälschten Ergebnissen führen kann. Deswegen wurde ein weiterer Test durchgeführt, der nur diejenigen Dokumente verwendet, die einen Autor im Titel tragen (26 Dokumente). Es kann da natürlich nicht ausgeschlossen werden, dass trotzdem mehrere Autoren das Dokument verfasst haben. Die Auswertung dieser Daten ergab aber keine signifikanten Unterschiede zur Auswertung des Gesamtkorpus, wie in Abbildung 5.16 zu sehen ist.

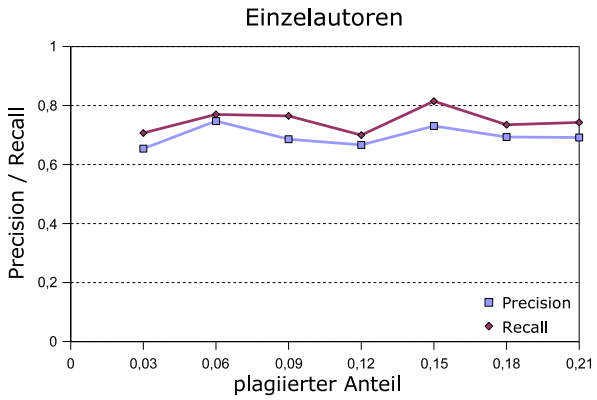


Abbildung 5.16: *Ergebnisse unter Verwendung aller 26 Dokumente, die nur einen Autor im Titel tragen*

3. Gibt es gute und schlechte Features? Wie ist deren Auswirkung auf das Ergebnis?

Eine weitere Möglichkeit, die zu Verschlechterungen der Ergbenisse führen könnte, wäre die, wenn es Features gibt, die so schlecht funktionieren, dass sie negativ beeinflussend wären. Aus diesem Gedanken heraus wurde ein sog. Feature-Ranking erstellt und mit diesem verschiedenene Tests ausgeführt.

SPSS liefert bei der Diskriminanzanalyse eine Tabelle der Struktur-Matrix. In dieser sind die gemeinsamen Korrelationen innerhalb der Gruppen zwischen den Diskriminanzvariablen und der standardisierten kanonischen Diskriminanzfunktion angegeben. Die Variablen, hier Features, sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet. Das Feature mit dem größten absoluten Wert befindet sich an erster Position und hat den größten Anteil an der Unterscheidung in die zwei zu klassifizierenden Gruppen „plagiiert“ und „nicht plagiiert“. Dieses Feature ist das am besten trennende. Anhand dieser Rangfolge wurde allen Features eine natürliche Zahl zugeordnet. Insgesamt wurden 29 Features verwendet. Das beste und an Position 1 stehende bekam die 29 zugeteilt, das schlechteste die 1.

Das Diagramm zur Abbildung 5.17 wurde aus den sieben Struktur-Matrizen (pro plagiiertem Anteil eine) der Dikriminanzanalysen des gesamten Korpus aufgestellt. Aus den sieben Werten pro Feature wurde der Mittelwert gebildet. Bei Betrachtung der Ergebnisse kann festgestellt werden, dass die Average-Word-Frequency-Class das beste Feature ist. Die Stilmerkmale schneiden insgesamt besser ab, als die einzelnen Wortarten. Auffällig schlecht im Vergleich zu den anderen sind Yule’s K und der „Dale-Chall-Index“.

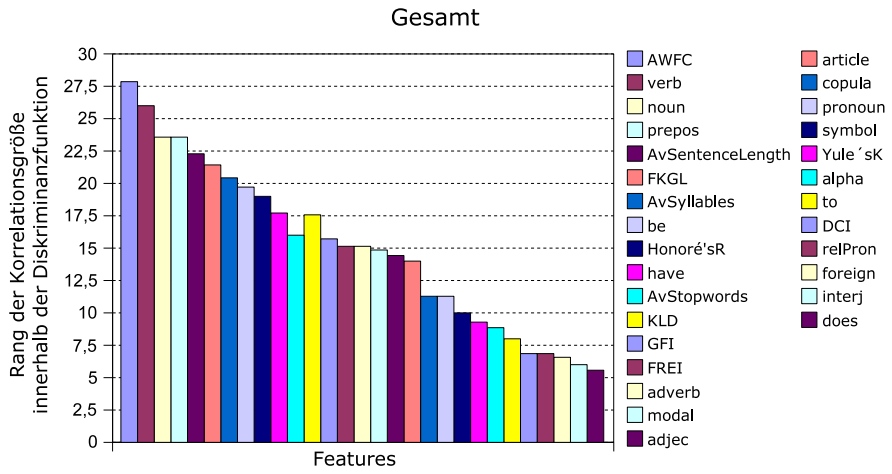


Abbildung 5.17: *Feature-Ranking Gesamt. Das beste Feature „Average-Word-Frequency-Class“ ist durch den höchsten Wert ausgezeichnet.*

Um eventuelle Textabhängigkeiten feststellen zu können, wurden diesselben Diagramme für die drei Themengebiete erstellt. In der dazugehörigen Abbildung 5.18 ist sichtbar, dass die Average-Word-Frequency-Class durchweg mit zu den besten Features gehört. Des Weiteren sind auch die einfachen Stilmerkmale „Durchschnittliche Satzlänge“ und „Durchschnittlicher Stoppwortanteil“ bei allen Themengebiete gut trennende Features. Die besten Wortarten sind Substantive und Verben. Schwankende Werte sind bei den Lesbarkeitsformeln zu erkennen. Bei allen Themengebieten schlechte Features sind der Dale-Chall-Index, Yule’s K und einzelne Wortarten.

Anhand dieser Ergebnisse kann geschlossen werden, dass bei den meisten Features eine Textabhängigkeit vorhanden ist. Bei einer Analyse völlig anderer Texte könnten wieder andere Features besser abschneiden.

Für diesen Zusammanhang sprechen auch die ermittelten Werte von Precision und Recall unter Verwendung der in Abbildung 5.19 dargestellten besten und schlechtesten zehn Features. Die Ergebnisse der besten zehn sind deutlich besser, als die der schlechtesten zehn. Dennoch sind die der besten zehn schlechter, als die Ergebnisse bei Verwendung aller Features (siehe Abbildung 5.14).

Die „schlechten“ Features verschlechtern sozusagen das Gesamtergebnis nicht. Ein Test, bei dem alle Features außer dieser zehn verwendet wurde, zeigte so keine Verbesserung, aber auch keine Verschlechterung der Ergebnisse. Das ist so zu erklären, als dass die Trennkraft dieser schlechten Features einfach zu klein ist, als dass sie für die Ergebnisse verantwortlich sein können. Diese können also weiterhin für die Diskriminanzanalyse verwendet werden.

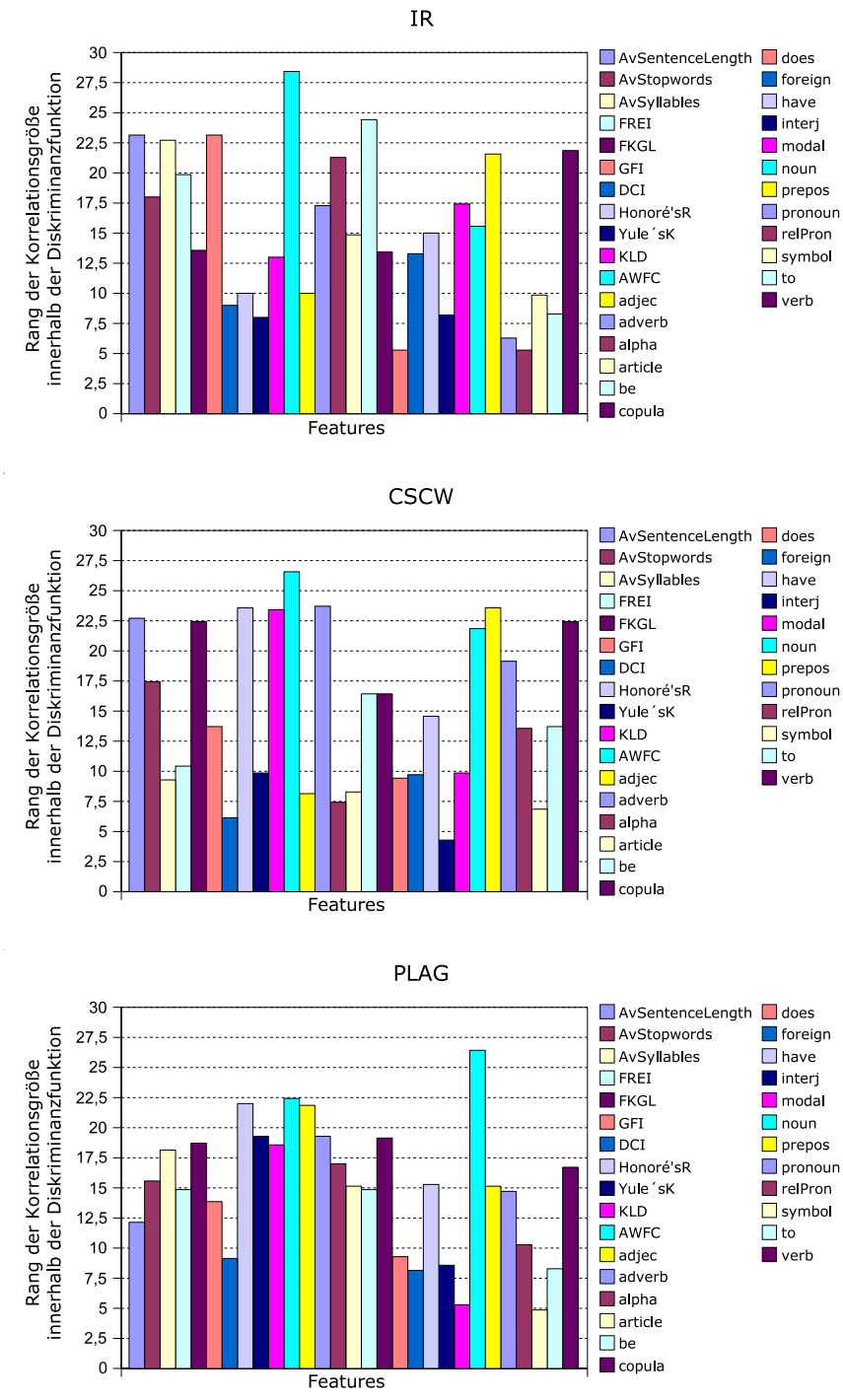
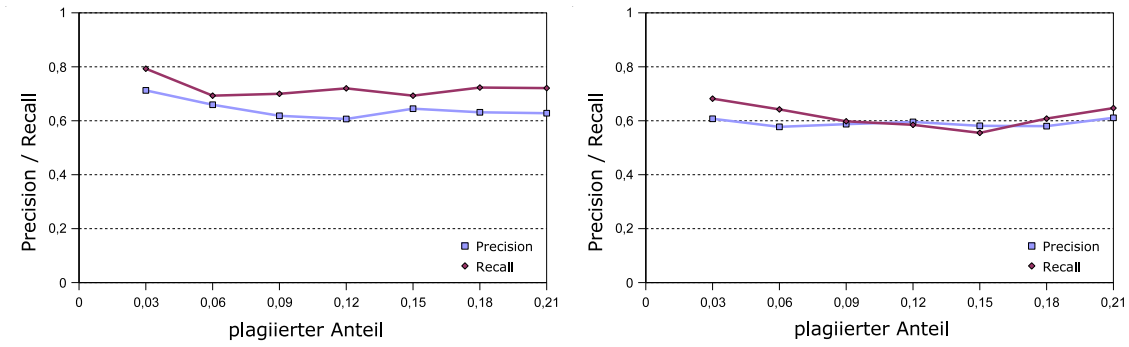


Abbildung 5.18: *Feature-Ranking der drei verwendeten Textgruppen zur Darstellung der Abhängigkeit vom Themengebiet.*



(a) 10 besten Features

(b) 10 schlechtesten Features

Abbildung 5.19: *Vergleich von 10 besten und 10 schlechtesten Features*

Abbildung 5.20 zeigt, dass nicht nur zwischen den drei Textgruppen, sondern auch zwischen den plagiierten Anteilen Unterschiede bezüglich des Feature-Rankings bestehen. Aus Gründen der Anschaulichkeit wurden hier nur die Stilmerkmale zusammengestellt. Die Unterschiede zwischen den Wortarten sind noch größer. Allein die Average-Word-Frequency-Class liegt bei allen Anteilen, außer bei 0.15 und 0.18, an erster Position. Diese Unterschiede in den plagiierten Anteilen bestehen auch für jede Textgruppe separat.

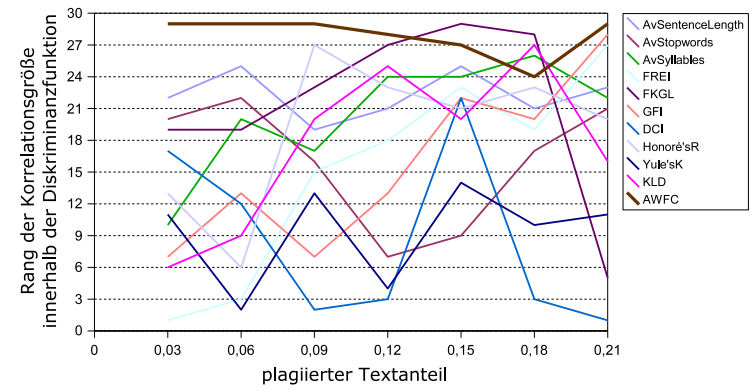
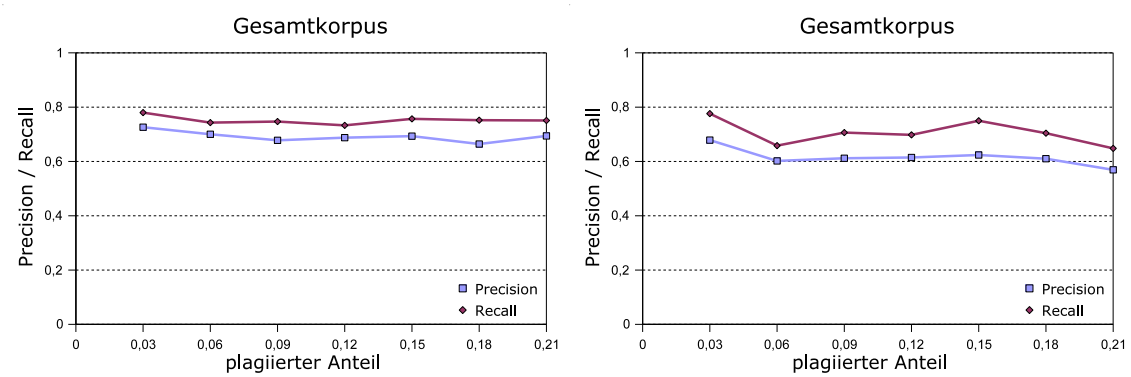


Abbildung 5.20: *Feature-Ranking der Stilmerkmale berechnet über die plagiierten Anteile aller Texte des Korpus. Die Average-Word-Frequency-Class erzielte über den gesamten Bereich die besten Ergebnisse.*

5.5.4 Auswertung der Daten - SVM

Um die Ergebnisse von Diskriminanzanalyse und SVM vergleichen zu können, enthalten die SVM-Dateien diesselben Feature-Dateien, wie die SPSS-Dateien. Da bei der SVM eine Überanpassung bei Verwendung zu kleiner Datensätze verhindert wird, sollten die Ergebnisse besser ausfallen, als bei der Diskriminanzanalyse. Das war nicht der Fall und ist auf die Feature-Normierung (siehe Formel 5.3) zurückzuführen. In Abbildung 5.21 werden die Ergebnisse von Diskriminanzanlyse und SVM, die für die gesamte Dokumentmenge durchgeführt wurden, gegenübergestellt. Es ist zu sehen, dass alle Werte für Precision und Recall von der SVM schlechter sind. Diese Unterschiede zeigten sich auch beim Vergleich der einzelnen Themengruppen IR, CSCW und PLAG.



(a) Diskriminanzanalyse (b) SVM

Abbildung 5.21: *Vergleich der Ergebnisse Diskriminanzanalyse und SVM. Bei der SVM ist die Precision um etwa 10 % , der Recall um etwa 5 % niedriger. Der Grund liegt dafür in der Feature-Normierung.*

Hier zeigt sich der Nachteil der SVM. Da der Einfluss der Variablen nicht nachvollziehbar ist, kann so nicht herausgefunden werden, wodurch die Ergebnisse verbessert hätten werden können. So war zum Beispiel nicht feststellbar, was gute oder schlechte Instance-Dateien sind, oder ob es „schlechte“ Features gibt, die das Ergebnis verschlechtert haben.

5.5.5 Schlussfolgerungen

Grundlage für dieses Experiment waren die Evaluierungsmaße Precision und Recall, die mittels Diskriminanzanalyse und SVM gewonnen wurden. Bei der Verwendung aller Dokumente lag die Precision bei ungefähr 70 % und der Recall bei 75 % und das unabhängig vom plagiierten Anteil.

Bei getrennter Untersuchung der drei Themengebiete IR, CSCW und PLAG zeigten sich kleine Unterschiede. Diese können auf die Korpusdokumente zurückgeführt werden. Es kann somit gesagt werden, dass die intrinsische Plagiaterkennung funktioniert.

Weitere Tests ergaben, dass eine Abhängigkeit vom jeweiligen Text besteht, ob ein bestimmtes Feature gut oder schlecht zur Unterscheidung in „plagiiert“ und „nicht plagiiert“ beiträgt. Es konnten Features gefunden werden, die bei den bis jetzt verwendeten Texten immer schlecht abschnitten. Für die Diskriminanzanalyse ist jedoch der Ausschluss dieser nicht notwendig, da sie nicht zu einer möglichen Verschlechterung beitragen. Genauso gibt es Features, die immer gut bis sehr gut waren. Das beste ist das neu zur Plagiaterkennung eingeführte Maß der „Average-Word-Frequency-Class“. Es funktionierte hervorragend und das mit vereinzelten Ausnahmen über alle Textgruppen und plagiierte Anteile hinweg. Es entsprach des Weiteren den Erwartungen, dass das VRM Yule's K viel schlechter, als Honore's R funktioniert.

Die Auswertung der Ergebnisse, die mit der SVM erzeugt wurden, ergeben keine neuen Erkenntnisse. Im Gegensatz zu den Erwartungen sind die Werte für Precision und Recall nicht besser geworden. Eine genaue Ursache konnte nicht eindeutig bestimmt werden, da nicht festgestellt werden konnte, welchen Einfluss die Features auf die Klassifizierung haben und wie deren Normierung aussehen muss.

5.6 Zusammenfassung

Für das Testen der intrinsischen Plagiaterkennung wurde eine Evaluierung durchgeführt. Dafür wurde zunächst für jeden Abschnitt jeder Instance-Datei eine Feature-Datei generiert, in denen die Feature-Werte der Stilmerkmale gespeichert sind. Vor dem eigentlichen Experiment zur Plagiaterkennung wurde ein Vorexperiment durchgeführt, das zehn Stilmerkmale, inklusive des neuen Maß Average-Word-Frequency-Class, auf ihre Stabilität über verschiedene Textlängen getestet hat.

Als mit großem Abstand stabilste zeigt sich bei den Tests das neue Stilmerkmal Average-Word-Frequency-Class. Auch sehr stabile Ergebnisse liefert der Dale-Chall-Index. Der Nachteil bei diesem ist, dass die jeweiligen Unterschiede zwischen verschiedenen Texten sehr klein sind. Die Lesbarkeitsformeln Flesch-Kincaid-Grade-Level, Gunning-Fog-Index und Flesch-Reading-Ease-Index liefern gute Resultate. Sehr schlecht schneiden allerdings die VRMs Honoré's R und Yule's K ab. Aber im Gegensatz zu der Studie von Tweedie und Baayen [55], ist hier R nicht instabil für kleine, gleichbleibende Textlängen.

Für die Evaluierung des Experiments der Plagiaterkennung wurden die beiden Maße Precision und Recall verwendet, die mittels der Klassifizierungsverfahren Diskriminanzanalyse und SVM

gewonnen wurden.

Die Ergebnisse der Diskriminanzanalyse zeigen, dass die intrinsische Plagiaterkennung unabhängig vom plagiiertem Anteil funktioniert. Unterschiede zwischen Themengebieten werden dabei auf die verwendeten Dokumente und nicht auf die Methode zurückgeführt. Bezüglich des Findens von „guten“ und „schlechten“ Features kann der erwartete Zusammenhang zwischen Stabilität und Trennkraft des Features bestätigt werden. Das beste Feature ist die Average-Word-Frequency-Class, was auch anhand der Stabilitätsresultate zu erwarten war. Ebenso wird mittels der Stabilität über kleine Textabschnitte begründet, dass das VRM Honoré's R deutlich bessere Resultate als Yule's K liefert. Die einfachen Stilmerkmale „Durchschnittliche Satzlänge“, „Durchschnittliche Silbernanzahl pro Wort“ und „Durchschnittlicher Stoppwortanteil“ zeigen zu den Lesbarkeitsformeln vergleichbare Werte.

Die Klassifizierung mit der SVM liefert keine Verbesserungen der Ergebnisse von Precision und Recall. Da hier nicht nachvollziehbar ist, inwieweit die verwendeten Stilmerkmale in die Berechnung einfließen und wie die entsprechende Normierung der Features auszusehen hat, können keine eindeutigen Ursachen für die erzeugten Ergebnisse festgestellt werden.

6 Zusammenfassung und Ausblick

Das Ziel dieser Arbeit bestand darin eine Möglichkeit zu finden, die es ermöglicht Plagiate intrinsisch zu erkennen, um somit völlig unabhängig von jeglichen Originalquellen zu sein. Da bereits nachgewiesen wurde, unter Anderem von Baayen et. al [2], dass der individuelle Schreibstil eines Menschen zuverlässig berechnet werden kann, bot sich das Prinzip der Stilanalyse zur intrinsischen Plagiaterkennung an.

Der erste Schritt dieser Arbeit bestand in der genauen Definition des Plagiats und dessen Zusammenhang zum Urheberrecht. Anhand dieser konnte die Stilanalyse in eine Taxonomie der möglichen Plagiatvergehen eingeordnet werden. Die Berechnung des Schreibstils eines Autors ist mit quantifizierbaren Stilmerkmalen möglich. Die bekanntesten und meist verwendeten Maße zur Berechnung der Wortvielfalt und der Lesbarkeit von Texten sind zusammengestellt worden. Zur intrinsischen Plagiaterkennung wurde sich für eine Auswahl daraus entschieden.

Zur Evaluierung dieses neuen Verfahrens wurde ein Korpus, basierend auf 100 wissenschaftlichen Artikeln, entnommen aus der „ACM-Digital-Library“, erstellt. Nach der Extrahierung der Dokumente folgte zur Simulierung des Plagiarismus das Einfügen fremder Abschnitte. Aus diesen präparierten Texten wurden Plagiate verschiedenen plagiierten Anteils generiert. Insgesamt standen 3200 plagiierte Dokumente für die Experimente zur Verfügung.

Die intrinsische Plagiaterkennung basiert auf der Erkennung von Schreibstilunterschieden zwischen Textabschnitten eines Dokuments. Es wurden die Stilmerkmale jedes Abschnitts und des Textes berechnet. Für jeden Abschnitt wurde eine Feature-Datei erzeugt, in der das Verhältnis jedes Stilmerkmalwerts zum entsprechenden Stilmerkmalswert des Textes gespeichert wurde. Wenn ein Großteil der Werte eines Abschnitts von den Werten des Textes abweicht, so ist das ein Indiz für einen plagiierten Abschnitt. Neben bekannten Stilmerkmalen wurden die zwei neuen Maße Kullback-Leibler-Divergenz und Average-Word-Frequency-Class eingesetzt.

Es galt nun herauszufinden, wie gut dieses Verfahren plagiierte Abschnitte erkennen kann. Dazu wurden die Maße Precision und Recall verwendet, die mittels der beiden bekannten Klassifizierungsverfahren Diskriminanzanalyse und Support-Vektor-Maschine (SVM) gewonnen wurden.

Im Ergebnis kann gesagt werden, dass intrinsische Plagiaterkennung durch Verwendung von Stilmerkmalen funktioniert. Etwa 75 % aller vorher festgelegten plagiierten Abschnitte wurden gefunden (Recall). Von allen ausgegebenen verdächtigen Abschnitten waren etwa 70 % tatsächlich plagiiert (Precision). Das ist ein sehr gutes Ergebnis. Im Vergleich dazu konnte Turnitin in einem Test von Weber-Wulff [59] nur 50 % bis 70 % der plagiierten Stellen richtig zuordnen. Hier zeigt sich aber auch der Nachteil der intrinsischen Plagiaterkennung. Während extrinische Plagiate nur entdecken oder nicht entdecken, werden intrinsisch auch Plagiate angezeigt, die keine sind. Dadurch und da bei Plagiatsverdacht der Verdächtigende in der Beweispflicht steht, müssen anhand der aufgezeigten plagiierten Stellen die Originalquellen gefunden werden.

In Studien wurde herausgefunden, dass sowohl Vocabulary-Richness-Measures nach Tweedie und Baayen [55], als auch Lesbarkeitsformeln nach Redish [41] nicht stabil auf kleinen Textabschnitten funktionieren. Diese Ergebnisse konnten nur zum Teil, mit dem in dieser Arbeit durchgeführten Stabilitätsexperiment, bestätigt werden. Die Lesbarkeitsformeln funktionieren im Allgemeinen stabil, die untersuchten Maße zur Bestimmung der Wortvielfalt, Honoré's R und Yule's K, jedoch nicht. Als stabilstes Stilmerkmal stellte sich die Average-Word-Frequency-Class dar. Dies bestätigte sich auch im durchgeführten Feature-Ranking.

A Literaturverzeichnis

[1] BAAYEN, R. H.: *The effects of lexical specialization on the growth curve of the vocabulary*. Comput. Linguist., 22[4]:455–480, 1996.

[2] BAAYEN, R. H., H. VAN HALTEREN, A. NEIJT und F. TWEEDIE: *An experiment in authorship attribution*. 6^{es} Journées internationales d'Analyse statistique des Données Textuelles, 2002.

[3] BOSER, B. E., I. M. GUYON und V. N. VAPNIK: *A training algorithm for optimal margin classifiers*. In: *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, S. 144–152, New York, NY, USA, 1992. ACM Press.

[4] BRODER, A. Z., S. C. GLASSMAN und M. S. MANASSE: *Syntactic Clustering of the Web*. `http://www.ambuehler.ethz.ch/CDstore/www6/Technical/Paper205/Paper205.html`. [Zugriff: 20.03.2006].

[5] CAN, F. und J. M. PALTON: *Change of Writing Style with Time*. Computers and the Humanities, S. 61–82, 2004.

[6] CAYLOR, J. S., T. G. STICHT, L. C. FOX und J. P. FORD: *Methodologies for determining reading requirements of military occupational specialties: Technical report No. 73-5*. Alexandria, VA: Human Resources, 1973.

[7] CHALL, J. und E. DALE: *A formula for predicting readability*. Educational Research Bulletin, 27:11–20, 1948.

[8] COLEMAN, M. und T. L. LIAU: *A computer readability formula designed for machine scoring*. Journal of Applied Psychology, 60:283–284, 1975.

[9] CORTES, C. und V. VAPNIK: *Support vector networks*. In: *M. Learning, 20:273-297*, 1995.

[10] DALE, E. und J. S. CHALL: *The concept of readability*. Elementary English XXVI, S. 23, 1949.

[11] DENNIS, S.: *The Sydney Morning Herald Word Database*. http://www2.psy.uq.edu.au/CogPsych/Noetica/OpenForumIssue4/SMH.html , 1995. [Zugriff: 07.02.2006].

[12] DUBAY, W. H.: *The Principles of Readability*. S. 3, 2004. http://www.impact-information.com/impactinfo/readability02.pdf . [Zugriff: 08.02.2006].

[13] EITRICH, T.: *Support-Vektor-Maschinen: Künstliche Intelligenz und Statistik im Bunde*. http://www.fz-juelich.de/zam/files/docs/vortraege/ja2003/SVM.pdf , 2003. [Zugriff: 21.07.2006].

[14] FLESCH, R.: *A new readability yardstick*. Journal of Applied Psychology, 32:221–233, 1948.

[15] FRÖHLICH, G.: *Mit fremden Federn*. http://www.falter.at/heureka/archiv/99_4/04.php , 1999. [Zugriff: 29.02.2006].

[16] FRY, E. B.: *The readability graph validated at primary levels*. The reading teacher, 22:534–538, 1969.

[17] FRY, E. B.: *Abbildung Fry Graph*. http://school.discovery.com/schrockguide/fry/fry_grades.gif , 2006. [Zugriff: 28.02.2006].

[18] GREENFIELD, G.: *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?*. Ed.D. dissertation, Temple University. University Microfilms No. 99-38670, 1999.

[19] GUNNING, R.: *The Technique of Clear Writing*. New York, NY: McGraw-Hill International Book Co, 1952.

[20] GUSFIELF, D.: *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[21] HEINTZE, N.: *Scalable Document Fingerprinting*. In: *1996 USENIX Workshop on Electronic Commerce*, November 1996.

[22] HOCHSCHULREKTORENKONFERENZ: *Zum Umgang mit wissenschaftlichem Fehlverhalten in den Hochschulen*. http://www.hrk.de/de/beschluesse/109_422.php , 1998. [Zugriff: 28.02.2006].

[23] HOLMES, D. I.: *Authorship Attribution*. Computers and the Humanities, 28:87–106, 1994.

[24] HONORÉ, A.: *Some simple measures of richness of vocabulary*. Association for Literary and Linguistic Computing Bulletin, 7[2]:172–177, 1979.

[25] HUBER, K.: *Diskriminanzanalyse*. http://www.wiwi.uni-passau.de/lehrstuehle/schweitzer/doc/Diskriminanzanalyse.ppt . [Zugriff: 14.06.2006].

[26] JOACHIMS, T.: SVM^{light} - *Support Vector Machine*. http://svmlight.joachims.org . [Zugriff: 29.06.2006].

[27] KINCAID, J. P., R. FISHBURNE, R. L. ROGERS und B. S. CHISSOM: *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. CNTECHTRA Research Branch Report 8-75, 1975.

[28] KLEPPE, A., D. BRAUNSDORF, C. LOESSNITZ *und* S. MEYER ZU EISSEN: *On Web-based Plagiarism Analysis*. 2nd International Workshop on Text-based Information Retrieval (TIR-05), 2005.

[29] KNOW LIBRARY: *Prosa*. http://prosa.know.library.net , 2004. [Zugriff: 18.07.2006].

[30] KOPPEL, M., J. SCHLER und D. MUGHAZ: *Text Categorization for Authorship Verification*. http://citeseer.ist.psu.edu/697544.html . [Zugriff: 07.02.2006].

[31] MASON, O.: *QTag*. http://www.english.bham.ac.uk/staff/omason/software/qtag.html , 2003. [Zugriff: 07.02.2006].

[32] MCLAUGHLIN, G.: *SMOG grading - a new readability formula*. Journal of reading, 22:639–646, 1969.

[33] MEHLER, A.: *Korpuslinguistik: Kapitel III Aufbereitung und Annotation*. http://gewilab.uni-graz.at/mehler/korpuslinguistik/PDF/Korpuslinguistik_3.pdf , 2003. [Zugriff: 08.02.2006].

[34] MEYER ZU EISSEN, S., B. STEIN, *und* M. KULIG: *Plagiarism Detection without Reference Collections*. Bislang unveröffentlichter Artikel. Erstellt am Institut für Content Management und Web Technologien der Fakultät Mediensysteme an der Bauhaus-Universität Weimar. Eingereicht zur '30th Annual Conference of the German Classification Society (GfKl)', 2006.

[35] MEYER ZU EISSEN, S. *und* B. STEIN: *Genre Classification of Web Pages: User Study and Feasibility Analysis*. In: SUSANNE BIUNDO, T. F. und G. PALM (Hrsg.): *KI 2004: Advances in Artificial Intelligence*, Bd. 3228, S. 256–269. Springer, 2004.

[36] MEYER ZU EISSEN, S. *und* B. STEIN: *Intrinsic Plagiarism Detection*. Proceedings of the European Conference on Information Retrieval (ECIR-06), 2006.

[37] NATIONAL LITERACY TRUST: *SMOGGING - how to test the readibility of text*. `http://www.literacytrust.org.uk/campaign/SMOG.html` , 2006. [Zugriff: 10.07.2006].

[38] PAPROTTÉ, W.: *liOn - Linguistik Online: Korpuslinguistik*. `http://luna.lili.uni-bielefeld.de/lion` . [Zugriff: 07.02.2006].

[39] POTTHAST, M.: *Hashing-basierte Indizierungsverfahren im textbasierten Information-Retrieval*. Bislang unveröffentlichte Diplomarbeit. Erstellt am Institut für Informatik der Fakultät für Elektrotechnik, Informatik und Mathematik an der Universität Paderborn, 2006.

[40] POWERS, R. D., W. A. SUMNER und B. E. KEARL: *A recalculation of four adult readability formulas*. J. Educ. Psychol., 49:99–105, 1958.

[41] REDISH, J.: *Readability formulas have even more limitations than Klare discusses*. ACM J. Comput. Doc., 24[3]:132–137, 2000.

[42] RUSCHMEIER, C.: *Spatial Chunking - Organisation von Routenwissen.* `http://www.math.uni-muenster.de/SoftComputing/lehre/seminar/ws0304/vortraege/ruschmeier/SpatialChunking.ppt`. [Zugriff: 20.03.2006].

[43] SCHRIVER, K. A.: *Readability formulas in the new millennium: what's the use?*. ACM J. Comput. Doc., 24[3]:138–140, 2000.

[44] SHIVAKUMAR, N. und H. GARCIA-MOLINA: *Building a scalable and accurate copy detection mechanism*. In: *DL '96: Proceedings of the first ACM international conference on Digital libraries*, S. 160–168, New York, NY, USA, 1996. ACM Press.

[45] SICHEL, H. S.: *On a Distribution Law for Word Frequencies*. Journal of the American Statistical Association, 70:542–547, 1975.

[46] SMITH, E. A. und R. J. SENTER: *Automated readability index*. AMRL-TR-66-22. Wright-Patterson AFB, OH: Aerospace Medical Division, 1967.

[47] SMOLA, A. und B. SCHOELKOPF: *A tutorial on support vector regression*. In: *NeuroCOLT2 Technical Report NC2-TR-1998-030*, 1998.

[48] SORENSEN, J.: *A Competitive Analysis of Automated Authorship Attribution Techniques*. 2004. [Zugriff: 02.03.2006].

[49] SPIEGEL ONLINE: *Trendsport Copy & Paste - Was heißt hier Eigentum?*. `http://www.spiegel.de/unispiegel/studium/0,1518,422280,00.html`, 2006. [Zugriff: 06.07.2006].

[50] SPIELKAMP, M.: *Wenig ist so, wie es scheint*. `http://www.irights.info/index.php?id=34`. [Zugriff: 29.02.2006].

[51] STAMATATOS, E., G. KOKKINAKIS und N. FAKOTAKIS: *Computer-Based Authorship Attribution Without Lexical Measures*. Computers and the Humanities, 35:193–214, 2001.

[52] STEIN, B.: *Fuzzy-Fingerprints for Text-Based Information Retrieval*. In: TOCHTERMANN, M. (Hrsg.): *Proceedings of the I-KNOW '05*, S. 572–579, Graz, Österreich, 2005. JUCS.

[53] STEIN, B. *und* S. MEYER ZU EISSEN: *Near Similarity Search and Plagiarism Analysis*. In: *Proceedings of the 29th Annual Conference of the German Classification Society (GfKl '05)*. Springer, 2005.

[54] T-REX: *Terminosaurus Rex - Die Informationswissenschaft in Begriffen, Definition: Stoppwort*. `http://server02.is.uni-sb.de/trex/index.php?id=2.2.1.2.2.3.1.1.`, 2005. [Zugriff: 11.08.2006].

[55] TWEEDIE, F. und H. BAAYEN: *How variable may a constant be? Measures of lexical richness in perspective*. Computers and the Humanities, 32[5]:323–352, 1998.

[56] UNIVERSITÄT LEIPZIG: *Wortschatz*. `http://wortschatz.uni-leipzig.de`, 1995. [Zugriff: 10.04.2006].

[57] VAPNIK, V.: *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.

[58] WEBER-WULFF, D.: *Fremde Federn Finden - Eine E-Learning Einheit*. `http://plagiat.fhtw-berlin.de/ff/04auffinden/verdacht.html`, 2004. [Zugriff: 07.03.2006].

[59] WEBER-WULFF, D.: *Gute Programme*. `http://plagiat.fhtw-berlin.de/ff/05hilfen/gut.html`, 2004. [Zugriff: 13.07.2006].

[60] WEITZEL, D.: *Who's reading your writing?*. http://www.ext.colostate.edu/PUBS/octnews/oc030602.html , 2004. [Zugriff: 07.02.2006].

[61] YULE, G.: *The Statistical Study of Literary Vocabulary*. Cambridge University Press, 1944.

[62] ZHENG, R., Y. QUIN, Z. HUANG und H. CHEN: *Authorship Analysis in Cybercrime Investigation*. URL: http://www.ecom.arizona.edu/ISI/2003/resources/presentation/AUTHORSHIP_ANALYSIS_BY_RO.PPT , 2003. [Zugriff: 07.02.2006].

[63] ZIPF, G. K.: *Selected studies of the principle of relative frequency in language*. Boston: Houghton-Mifflin, 1932.