Bauhaus-Universität Weimar

Fakultät Medien

Studiengang Mediensysteme

Web Technology & Information Systems

Diplomarbeit

# Sprachübergreifendes Retrieval von ähnlichen Dokumenten aus großen Textkollektionen

Katja Schöllner

Betreuer: Martin Potthast

1. Gutachter: Prof. Dr. Benno Maria Stein

1. Dezember 2008

# Eidesstattliche 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.

Erfurt, den 1. Dezember 2008

Katja Schöllner

# Kurzfassung

Mit Entwicklung des Internet wächst die Menge der digital zur Verfügung stehenden Informationen. Gleichzeitig steigt die Zahl der Plagiate an, da Texte ohne großen Aufwand kopiert werden können. Als Plagiat wird ein Dokument bezeichnet, dem fremder Texte ohne Verweis auf eine Quelle hinzugefügt wurde. Besonders in akademischen Bereichen sind Plagiatvergehen ein Problem. Inzwischen ist die Entdeckung verschiedener Arten von Plagiaten Gegenstand der Forschung. Wenig untersucht sind Verfahren zur Ermittlung von Plagiaten, die durch Übersetzungen entstanden sind.

Ziel der Plagiatanalyse ist die Identifizierung möglichst aller Textpassagen, die ohne Kennzeichnung aus anderen Dokumenten entnommen wurden, sowie das Auffinden der jeweiligen Quellen. Die vorliegende Diplomarbeit beschäftigt sich mit Strategien, wie in großen Datenkollektion Dokumente gefunden werden können, die den möglichen Ausgangspunkt für ein Übersetzungsplagiat darstellen. In diesem Rahmen wird die Plagiatanalyse als Retrieval-Aufgabe definiert. Verfahren des Information-Retrieval werden vorgestellt und auf ihre Anwendbarkeit in der multilingualen Plagiatanalyse untersucht.

Im Fokus stehen dabei drei Ansätze. Zwei basieren auf der Suche mit Schlüsselwörtern, die dem verdächtigen Dokument entnommen werden. Die erste Methode übersetzt die Schlüsselbegriffe nach der Extraktion in die vermutete Originalsprache, die zweite übersetzt zunächst das Dokument, bevor die Schlüsselwörter ermittelt werden. Mit diesen Schlüsselbegriffen findet anschließend klassisches, monolinguales Retrieval statt. Der dritte zu untersuchende Ansatz nutzt Hashing-basiertes Retrieval. Vom verdächtigen, übersetzten Dokument wird ein Fingerprint erstellt, mit dem ein Fingerpint-Index angefragt wird. Eine Kollision stellt ein Indiz für ein Plagiatvergehen dar. Untersucht werden Retrievaleigenschaften, das Laufzeitverhalten sowie

die Abhängigkeit von verschiedenen Parametern wie Textlänge oder Anzahl der extrahierten Begriffe.

Es zeigt sich, daß der Fingerprinting-Ansatz das schnellste Verfahren ist. Jedoch ist die Retrieval-Qualität mit einem maximal errreichten Recall von 10% bzw. 17% nicht akzeptabel. Das auf der Übersetzung von Schlüsselwörtern basierende Verfahren scheitert an der Vielfalt der Übersetzungen und dem Fehlen einer geeigneten Heuristik zur Auswahl einer Übersetzungsvariante. Am besten geeignet ist das schlüsselwortbasierte Retrievalverfahren mit aus dem übersetzten Dokument entnommenen Schlüsselwörtern. Der Recall liegt zwischen 84% und 93%.

# Inhaltsverzeichnis

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Listings

# 1 Motivation

Wir leben in einer Gesellschaft, in der täglich große Mengen an Informationen produziert werden. Laut einer Studie von Lyman und Varian (2003) hat sich allein zwischen 1999 und 2002 die Menge der auf Papier, Film, magnetischen oder optischen Medien gespeicherten Daten von 2-3 Exabytes[1] auf 5 Exabytes verdoppelt. Das entspricht einem jährlichen Wachstum der gespeicherten Informationsmenge um 30 Prozent. 2002 befanden sich 92 Prozent der Informationen auf magnetischen Speichermedien.

Die Autoren der Studie gehen davon aus, daß die Menge der öffentlich verfügbaren Informationen im World Wide Web im Jahr 2000 bei 20 bis 50 Terabytes lag. 2003 war die Menge schon auf 167 Terabytes angewachsen. Der erste Suchindex von Google umfaßte 1998 bereits 26 Millionen Webseiten, 2000 wurde die Grenze von 1 Milliarde überschritten und 2008 ist diese Zahl auf über 1 Billion angewachsen (The Official Google Blog (2008)). Damit ist das WWW die größte verfügbare Dokumentkollektion.

Die wachsende Verfügbarkeit von Informationen in digitaler Form führt dazu, daß Dokumente leichter kopiert werden können und Plagiatvergehen zunehmen. Unter einem Plagiat (lat. plagium: Menschenraub) ist die *„vollständige oder teilweise Übernahme eines fremden literarischen, musikalischen oder bildnerischen Werkes in unveränderter oder nur unwesentlich geänderter Fassung unter Vorgabe eigener Urheberschaft“* zu verstehen (Brockhaus (1992)). Besonders an Schulen und in akademischen Bereichen sind Plagiate ein Problem. Hart und Friesner (2004) geben einen Überblick über verschiedene Untersuchungen zur Verbreitung von Plagiatvergehen.

[1] 1 Exabyte $\approx$ 1 Mrd. Gigabyte

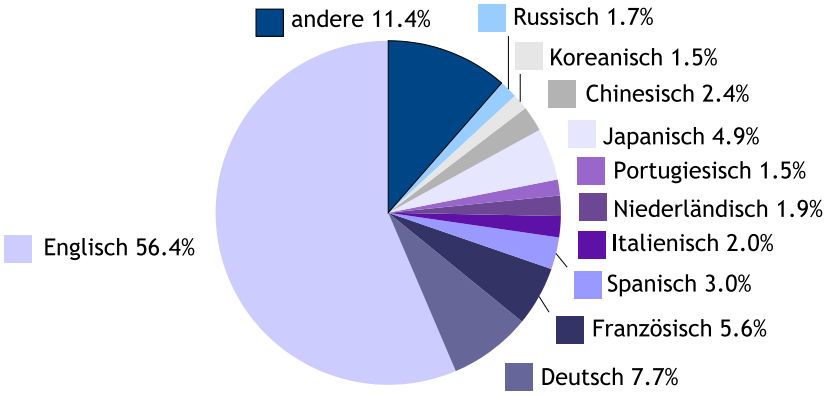


***Abbildung 1.1:*** *Sprachvielfalt von Webseiten im Jahr 2002 (Ebbertz (2002)).*

Vor allem Universitäten in Großbritannien und den USA setzen bei der Suche nach Plagiaten auf maschinelle Plagiaterkennung, wie dies beispielsweise der kommerzielle Dienst *Turnitin*[2] leistet. Viele Systeme können hauptsächlich Eins-zu-eins-Kopien entdecken, scheitern jedoch an Texten, die durch Umformulieren und Umstrukturieren geringfügig verändert wurden (Maurer u. a. (2006)). Eine besondere Form sind Plagiate, die durch Übersetzungen aus einer anderen Sprache entstanden und ohne Quellenangabe übernommen worden sind.

Eine Untersuchung zur Verbreitung von Sprachen im Internet (Ebbertz (2002)) zeigt die Sprachvielfalt, in der Informationen vorliegen (siehe Abbildung 1.1). In der aus dem Jahr 2002 stammenden Studie wird festgestellt, daß die Dominanz englischer Seiten zurückgeht und andere Sprachen an Bedeutung gewinnen. Es ist anzunehmen, daß sich dieser Trend fortgesetzt hat[3]. Daher ist auch davon auszugehen, daß das Plagiieren aus fremdsprachigen Quellen zunimmt. Die Erkennung solcher Vergehen sind bisher kaum Gegenstand der Forschung.

Die vorliegende Diplomarbeit beschäftigt sich mit der sprachübergreifenden Plagiatanalyse. Zu den Zielen der Plagiaterkennung gehört die Identifizierung möglichst aller Textpassagen, die aus anderen Dokumenten entnommen wurden, sowie das Auffinden der jeweiligen Quellen. Im Fokus dieser Arbeit steht dabei die sprachübergreifende Suche. Es werden drei Verfahren untersucht, welche aus großen Dokumentkollektionen wie beispielsweise dem WWW eine Menge von Dokumenten ermitteln,

[2] http://www.turnitin.com

[3] Leider konnte keine aktuellere Studie zur Sprachvielfalt gefunden werden

die den potenziellen Ausgangspunkt für ein Plagiatvergehen darstellen. Insbesondere werden Verfahren des Information-Retrieval auf ihre Anwendbarkeit in der multilingualen Plagiatanalyse untersucht.

Die Arbeit gliedert sich wie folgt:

In Kapitel 2 werden Grundlagen zur sprachübergreifenden Suche erläutert. Die Begriffe Information-Retrieval und Ähnlichkeitssuche werden erklärt sowie Probleme und Lösungsansätze sprachübergreifender Retrieval-Modelle vorgestellt.

Kapitel 3 befaßt sich mit dem Problem der Plagiatanalyse. Es gibt einen Überblick über verschiedene Arten des Plagiierens, stellt Verfahren zu deren Erkennung vor und widmet sich speziell der sprachübergreifenden Analyse. Dabei liegt der Schwerpunkt auf den zu untersuchenden Verfahren des heuristischen Retrievals, welche im Detail beschrieben werden.

Die zu untersuchenden Fragestellungen werden in Kapitel 4 vorgestellt. Außerdem erfolgt ein Überblick über die verwendeten Testkollektionen, die genaue Beschreibung der einzelnen Experimente sowie die Diskussion der Ergebnisse.

Den Abschluss dieser Arbeit bildet Kapitel 5, welches die Resultate dieser Arbeit noch einmal zusammenfaßt.

# 2 Sprachübergreifende Ähnlichkeitssuche

In diesem Kapitel werden die Grundlagen der sprachübergreifenden Suche behandelt. Zunächst erfolgt eine Einführung in das Fachgebiet Information-Retrieval, die Charakterisierung der Retrieval-Aufgabe „Ähnlichkeitssuche“ sowie die Vorstellung ausgewählter Retrieval-Modelle. Anschließend werden Schwierigkeiten bei der sprachübergreifenden Ähnlichkeitssuche erläutert und verschiedene Lösungsstrategien diskutiert.

## 2.1 Information-Retrieval

Der englische Begriff „information retrieval“ bedeutet wörtlich Informationswiedergewinnung oder Informationsbeschaffung. Das Fachgebiet Information-Retrieval beschäftigt sich mit der Repräsentation, der Strukturierung und dem Speichern von Informationen, vor allem aber mit der Suche in großen, unstrukturierten Datenmengen. Die Suche nach Informationen kann dabei die Suche nach Texten, Bildern, Audio- und Videomaterial, u.ä. umfassen und stellt sich als inhaltsorientierte Suche auf Ebene der Semantik (also der Bedeutung) und der Pragmatik (das heißt des Verwendungszwecks) dar. Ein Information-Retrieval-System (IR-System) versucht das Informationsbedürfnis eines Benutzers zu befriedigen, indem es zu einer Anfrage passende Objekte innerhalb einer Datenkollektion findet.

Die Suche ist durch Vagheit und Unsicherheit gekennzeichnet. Anfragen sind vage, da die Antworten nicht im voraus klar definiert sind. Ein Nutzer weiß unter Umständen nicht genau, was er eigentlich sucht, oder ob das Gesuchte existiert.

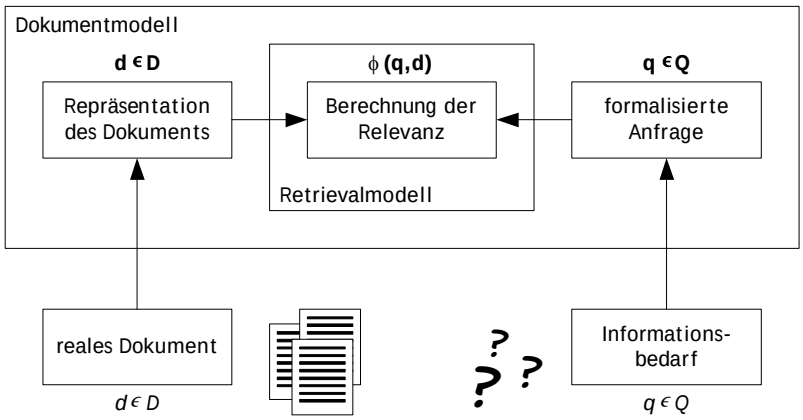


***Abbildung 2.1:*** *Information-Retrieval: Zur Beurteilung der Relevanz eines realen Dokuments $d$ bezüglich des Informationsbedürfnisses $q$ eines Benutzers werden Anfrage und Dokument zu $\mathbf{d}$ und $\mathbf{q}$ abstrahiert. Eine Relevanzfunktion $\varphi$ bestimmt, ob $\mathbf{d}$ für $\mathbf{q}$ relevant ist (vgl. Stein (2007)).*

Im folgenden Abschnitt wird der Begriff „Information-Retrieval“ formalisiert.

### 2.1.1 Formalisierung

Sei $D$ eine Menge von Dokumenten und $d$ ein Dokument daraus. Sei ferner $Q$ die Menge von Informationsbedürfnissen und $q$ ein konkreter Informationsbedarf (engl. query) daraus. Die Aufgabe eines Retrieval-Systems besteht darin, alle Dokumente $D_q \subset D$ zu finden, die bezüglich $q$ relevant sind (siehe Abbildung 2.1). Um das zu ermöglichen, muß ein IR-System sowohl $q$ als auch alle Dokumente in $D$ abstrahieren. $\mathbf{d}$ sei also die Repräsentation von $d$, $\mathbf{q}$ die Abstraktion von $q$ (vgl. Baeza-Yates und Ribeiro-Neto (1999)). Bestandteil der Repräsentation von $d$ können drei verschiedene Sichten auf dessen Inhalt sein. Dazu gehören (1) die Layout-Sicht, die die Darstellung des Dokuments festlegt, (2) die strukturelle bzw. logische Sicht, welche den Aufbau des Dokuments definiert, sowie (3) die semantische Sicht auf ein Dokument; sie ermöglicht die Interpretation des Dokuments. Mit Hilfe einer Funktion $\rho_{\mathcal{R}} : \mathbf{Q} \times \mathbf{D} \rightarrow \mathbf{R}$ kann die Relevanz einer Dokumentrepräsentation $\mathbf{d} \in \mathbf{D}$ bezüglich einer formalisierten Anfrage $\mathbf{q} \in \mathbf{Q}$ ermittelt werden. $\rho$, $\mathbf{D}$ und $\mathbf{Q}$ sind Bestandteil eines Retrievalmodells $\mathcal{R}$. Es liefert die Grundlage, wie aus den Sichten eines Dokuments Rückschlüsse über dessen Relevanz in Bezug auf ein Informationsbedürfnis gezogen werden können. Dabei ist zu beachten, daß die Menge der Dokumente, die

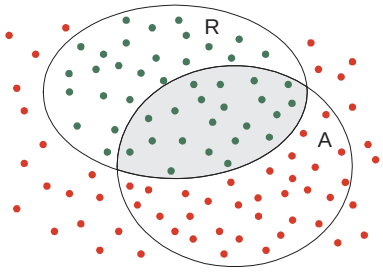


***Abbildung 2.2:*** *Mengen zur Bestimmung von Precision und Recall. Die grünen Punkte bilden die Menge R der relevanten Dokumente. Rote Punkte bedeuten irrelevante Dokumente bezüglich einer Anfrage. Die Menge A enthält alle Dokumente, die vom Retrieval-System gefunden wurden. Die grau schattierte Fläche beinhaltet alle relevanten Dokumente der Antwortmenge.*

das Retrieval-System als relevant bewertet, nicht zwangsläufig mit der Menge der tatsächlich relevanten Dokumente übereinstimmen muß, das heißt $D_q \subset D$ enthält nicht die gleichen Dokumente wie $\mathbf{D_q} \subset \mathbf{D}$. Gesucht sind daher Retrieval-Modelle, die es erlauben, das reale $D_q \subset D$ so gut wie möglich anzunähern.

Die Qualität eines Retrieval-Modells läßt sich über die Maße Precision und Recall bestimmen. Der Recall gibt an, wieviele der tatsächlich relevanten Dokumente von einer Retrievalfunktion als relevant bewertet und damit vom IR-System gefunden wurden. Mit Bezug auf Abbildung 2.2 ist der Recall folgendermaßen definiert:

$$Recall = \frac{R \cap A}{R}$$

Die Precision ermittelt den Anteil der relevanten Dokumente aus der Menge der gefundenen Dokumente.

$$Precision = \frac{R \cap A}{A}$$

Ein ideales Retrieval-System hat sowohl eine Precision als auch einen Recall von 1, das heißt die Mengen A und R enthalten die selben Elemente. Es existieren eine Reihe von weiteren Maßen, die versuchen, Precision und Recall in einem einzigen Wert auszudrücken. Dazu zählen u.a. das F-Measure oder die Mean-Average-Precision, welche im folgenden kurz erklärt werden, da im Verlauf der Arbeit auf sie zurückgegriffen wird.

- F-Measure: Das gewichtete harmonische Mittel aus Precision und Recall wird als F-Measure bezeichnet und ist definiert als

$$F = \frac{2 \cdot Precision \cdot Recall}{(Precision + Recall)}$$

  Hier werden Precision und Recall das gleiche Gewicht zugewiesen. Mit der verallgemeinerten Formel

$$F_\beta = \frac{(1+\beta^2) \cdot (Precision \cdot Recall)}{(\beta^2 \cdot Precision + Recall)} \text{ mit } \beta > 0$$

  sind andere Gewichtungen möglich (van Rijsbergen (1979)). $\beta = 2$ gewichtet den Recall doppelt so hoch wie die Precision, während die Precision mit $\beta = 0.5$ das doppelte Gewicht des Recalls erhält.

- Mean-Average-Precision (MAP): Ist die Ergebnisliste gewichtet, kann für jede Anfrage die Average-Precision bestimmt werden. Dazu wird nach jedem relevanten Dokument in der Ergebnisliste die Precision bestimmt. Diese Werte werden addiert und die Summe anschließend durch die Gesamtzahl aller relevanten Dokumente geteilt. Die Mean-Average-Precision bildet nun noch das Mittel der Average-Precision über alle Anfragen.

Die Wahl eines geeigneten Retrieval-Modells hängt von der zu lösenden Retrieval-Aufgabe und der Art der Anfrage ab. Ein Informationsbedarf kann als Wortanfrage formuliert werden, wie dies bei herkömmlichen Suchmaschinen der Fall ist, oder in Form eines Beispieldokuments. Bei letzterem sollen diejenigen Dokumente gefunden werden, die zu dem Beispieldokument ähnlich sind. Diese Form der Suche wird auch als Ähnlichkeitssuche bezeichnet.

## 2.2 Ähnlichkeitssuche

Die häufigste Aufgabe im Information-Retrieval ist die Suche nach passenden Dokumenten zu gegebenen Schlüsselwörtern. Eher untypisch ist die Ähnlichkeitssuche, also die Suche nach zueinander ähnlichen Dokumenten. Zu einer als Beispieldokument formulierten Anfrage $\mathbf{d_q}$ sollen diejenigen Dokumente $\mathbf{d_x}$ gefunden werden, die thematisch den gleichen Inhalt beschreiben und somit eine hohe Ähnlichkeit aufweisen. Im Folgenden ist mit einer Anfrage immer ein Dokument gemeint.

Für diese Art der Suche wird die Relevanz eines Dokuments anhand einer Ähnlichkeitsfunktion bestimmt. Eine solche Funktion definiert eine Relation $\mathbf{D} \times \mathbf{D} \rightarrow [0, 1]$, so daß je zwei Dokumenten ein Wert zwischen 0 und 1 zugeordnet wird. 0 bedeutet hier, daß keine Ähnlichkeit vorliegt, während 1 größtmögliche Ähnlichkeit[1] anzeigt.

### 2.2.1 Heuristische Ähnlichkeitssuche

Klassische Retrievalmodelle abstrahieren ein Dokument auf eine Menge von Indextermen. Dabei handelt es sich um eine Menge von Begriffen, die direkt aus dem Dokument entnommen werden. Sie werden so ausgewählt, daß sie den Inhalt des Dokuments repräsentieren, jedoch eine klare Abgrenzung einzelner Dokumente voneinander gewährleisten und die Verknüpfung thematisch ähnlicher Dokumente ermöglichen. Im Vektorraummodell (Salton u. a. (1975)) bilden typischerweise die Wortstämme aller Nicht-Stopwörter die Menge der Indexterme $V = f_1, f_2, \ldots, f_n\}$. Je nach Variante werden Strukturinformationen des Dokuments und Stopwörter[2] entfernt oder die Terme auf ihre Stammform reduziert.

Die Terme können als Dimensionen eines Vektorraumes aufgefaßt werden. So lassen sich sowohl die Anfrage $\mathbf{d_q}$ als auch die Dokumente $\mathbf{d_x}$ als Punkte in diesem Raum definieren (Abb. 2.3). Die Ähnlichkeit zweier Dokumente ist umgekehrt proportional zur Distanz zwischen ihren Vektoren und wird durch eine Ähnlichkeitsfunktion $\varphi(d, q)$ bestimmt. Als Ähnlichkeitsmaß im Vektorraummodell wird häufig das Kosinus-Ähnlichkeitsmaß verwendet, das auf dem Winkel $\varphi$ zwischen den Vektoren beruht, und wie folgt definiert ist

$$\varphi_c os(\mathbf{d_q}, \mathbf{d_x}) = \frac{< \mathbf{d_q}, \mathbf{d_x} >}{\|\mathbf{d_q}\| \|\mathbf{d_x}\|}$$

[1] Eine Ähnlichkeit von 1 muß nicht zwangsläufig Identität bedeuten. Im Vektorraummodell, das im nächsten Abschnitt vorgestellt wird, kann dieser Ähnlichkeitswert entstehen, wenn in zwei zu vergleichenden Dokumenten zwar dieselben Wörter verwendet wurden, diese aber in unterschiedlicher Reihenfolge erscheinen.

[2] Stopwörter sind Wörter, die sehr häufig und in nahezu jedem Dokument vorkommen. Aufgrund ihrer Häufigkeit können sie keine Abgrenzung einzelner Dokumente voneinander gewährleisten. Beispiel: Artikel (der, die, ein), Konjunktionen (und, oder), etc.

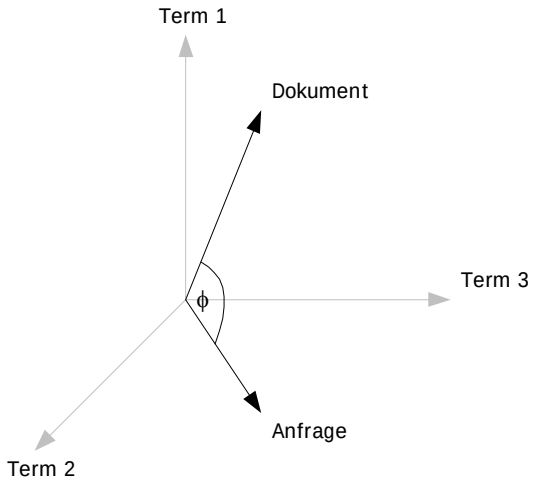


***Abbildung 2.3:*** *Vektorraummodell. Sowohl Dokumente als auch Anfrage werden als Vektoren definiert und die Ähnlichkeit über den zwischen ihnen liegenden Winkel berechnet.*

### 2.2.2 Hashing-basierte Ähnlichkeitssuche

Stein und Potthast (2007) schlagen als weiteren Ansatz zur Ähnlichkeitssuche in großen Dokumentkollektionen Hashing-basierte Indizierung vor. Dem liegt die Idee zugrunde, die Laufzeiteigenschaften von Hashtabellen für die Suche nach ähnlichen Dokumenten auszunutzen, indem eine Funktion gefunden wird, die ähnliche Dokumente mit hoher Wahrscheinlichkeit auf denselben Hashwert abbildet.

Ein Hashing-basierter Index kann direkt aus einer Hashfunktion und einer Hashtabelle konstruiert werden. Dazu wird für die zu indizierenden Dokumente jeweils der Hashwert berechnet und in der Hashtabelle an der entsprechenden Speicherstelle eine Referenz auf das Dokument gespeichert. So verweist jeder Hashwert in der Tabelle auf eine Menge von Dokumenten, die auf diesen Wert abgebildet wurden.

Standard-Hashfunktionen werden so konzipiert, daß sie die Menge von Hashwerten gleichmäßig auf die Speicherstellen der Hashtabelle abbilden. Wird die Funktion jedoch so gewählt, daß ähnliche Dokumente den gleichen Hashwert erhalten, lassen sich durch einmaliges Nachschlagen alle ähnlichen Dokumente finden.

Nach Stein u. a. (2008) wird dafür eine Hashfunktion $h_{\varphi}^{(\rho)} : \mathbf{D} \rightarrow \mathbf{N}$ gewählt, die die Menge der Dokumentrepräsentationen $\mathbf{D}$ auf die natürlichen Zahlen abbildet. $\rho$ ist

abhängig vom gewählten Hashing-Verfahren und ist Element der Menge $\Pi$ der Parameter zur Anpassung der jeweiligen Hashfunktion. Mit einer Schar $h_{\varphi}^{(\Pi)}$ von Hashfunktionen läßt sich dann ein Ähnlichkeitsmaß $\varphi$ nachbilden:

$$h_{\varphi}(d)^{(\Pi)}(\mathbf{d_q}) \cap h_{\varphi}^{(\Pi)}(\mathbf{d_x}) \neq \emptyset \Leftrightarrow \varphi(\mathbf{d_q}, \mathbf{d_x}) \geq 1 - \epsilon$$

$h_{\varphi}(d)^{(\Pi)}(\mathbf{d_q})$ stellt die Menge der Hashcodes von $\mathbf{d_q}$ dar, die alle Hashcodes $h_{\varphi}^{(\rho)}(\mathbf{d_q})$ mit $\rho \in \Pi$ enthält. $\epsilon$ mit $0 < \epsilon \ll 1$ ist ein Schwellwert zur Beurteilung der Ähnlichkeit zweier Dokumente zueinander. Auf diese Weise wird ein Dokument durch eine Menge von Hashcodes charakterisiert, die den sogenannten Fingerprint (Fingerabdruck) darstellen.

Ähnlichkeitssensitive Hashfunktionen erzeugen den Hashcode in 3 Schritten (vgl. Stein u. a. (2008)):

1. *Dimensionsreduktion*: Die hochdimensionale Dokumentrepräsentation $\mathbf{d}$ wird auf ein niedrigdimensionales $\mathbf{d}'$ abgebildet. Dies geschieht durch Einbettung oder Projektion. Bei der Einbettung werden die Dimensionen des niedrigdimensionalen Raums aus denen des hochdimensionalen Raums neu berechnet. Im Gegensatz dazu werden bei der Projektion nur ausgewählte Dimensionen in den niedrigdimensionalen Raum übernommen und die restlichen verworfen.

2. *Quantisierung*: Das niedrigdimensionale $\mathbf{d}'$ wird auf den Vektor $\mathbf{d}''$ abgebildet, indem beispielsweise reellwertige Komponenten diskretisiert werden.

3. *Kodierung*: In diesem letzten Schritt erfolgt die Abbildung von $\mathbf{d}''$ auf eine natürliche Zahl mittels einer Kodierungsvorschrift. Ergebnis ist der Hashcode für das Dokument.

Ein Beispiel für diese Art von Hashing-Verfahren ist das Fuzzy-Fingerprinting.

### Fuzzy-Fingerprinting

Fuzzy-Fingerprinting wurde speziell für textbasiertes Retrieval entwickelt (Stein (2005)). In Abbildung 2.4 werden die Schritte zur Konstruktion eines Fuzzy-Fingerprints dargestellt. Ausgangspunkt ist die Vektorraumdarstellung $\mathbf{d}$ eines Dokuments. Diese wird in ein niedrigdimensionales Modell $\mathbf{d}'$ eingebettet, welches auf Präfixklassen beruht.

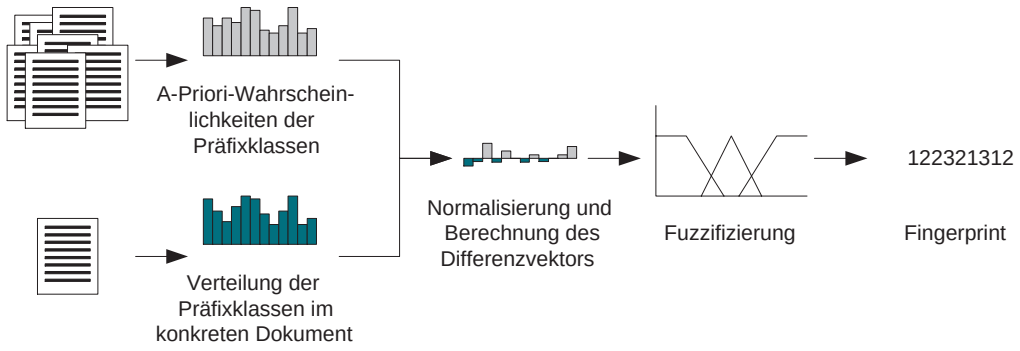


***Abbildung 2.4:*** *Schritte zur Konstruktion eines Fuzzy-Fingerprints (vgl. Stein (2005))*

Präfixklassen sind Äquivalenzklassen, in denen alle Wörter zusammengefaßt werden, die mit dem selben Präfix starten. Im Extremfall beinhaltet eine Präfixklasse diejenigen Wörter, welche mit demselben Buchstaben beginnen. Für jede dieser Klassen läßt sich die erwartete Häufigkeit mit Hilfe eines Referenzkorpus bestimmen. Stein und Potthast (2007) verwenden dafür den British National Corpus (BNC). Zur Repräsentation eines Dokuments wird dann die Differenz aus der erwarteten Häufigkeit und der im Dokument gefundenen Häufigkeit der Präfixklassen ermittelt. $\mathbf{d}'$ beinhaltet also die reellwertigen Abweichungen vom Erwartungswert und bildet die Grundlage für den zu berechnenden Fingerabdruck. Im nächsten Schritt werden die exakten Abweichungen quantifiziert. Dies geschieht, indem die Vektorkomponenten von $\mathbf{d}'$ mit Hilfe einer Fuzzifizierungsfunktion $\sigma$ auf linguistische Variablen, die beispielsweise „kleine Abweichung“ , „mittlere Abweichung“ und „große Abweichung“ lauten, abgebildet, so daß $\mathbf{d}'$ zum ganzzahligen Vektor $\mathbf{d}''$ wird. Anschließend erfolgt die Kodierung von $\mathbf{d}''$ zu einem Hashcode mittels einer Summierungsvorschrift $h_\varphi(\mathbf{d})^{(\rho)} = \sum_{i=0}^{|\mathbf{d}''|} \mathbf{d}''[i] \cdot r^i$, wobei $\mathbf{d}''[i]$ die $i$-te Komponente von $\mathbf{d}''$ darstellt.

Es kann vorkommen, daß sehr ähnliche Dokumente nicht den gleichen Fingerprint zugewiesen bekommen. Um dem entgegen zu wirken, werden zwei bis drei Fuzzifizierungsschemata mit bis zu vier Intervallen verwendet, so daß eine Menge von Hashcodes $h_\varphi(d)^{(\Pi)}(\mathbf{d_q})$ des Dokuments $\mathbf{d_q}$ entsteht, die alle Hashcodes $h_\varphi^{(\rho)}(\mathbf{d_q})$ mit $\rho \in \Pi$ enthält.

Ziel des Fingerprintings ist es, die Laufzeit des Retrieval-Prozesses zu verringern und den Speicherbedarf zu senken.

## 2.3 Sprachübergreifendes Information-Retrieval

Sprachübergreifendes Information-Retrieval (engl. *cross-language information retrieval*, CLIR) stellt eine Erweiterung des klassischen Information-Retrieval dar. Dokumente und Informationsbedarf sind in unterschiedlichen Sprachen formuliert. Ziel ist es, zu einer Anfrage alle relevanten Dokumente zu finden, unabhängig von deren Sprache. Die Schwierigkeit besteht darin, die Relevanz sprachübergreifend zu bestimmen. Es existieren also ein Informationsbedürfnis $q_L$, das in der Sprache $L$ formuliert ist, eine Kollektion von Dokumenten $D_{L'}$ in der Sprache $L'$ sowie eine Relevanzfunktion $\rho_{\mathcal{R}} : \mathbf{Q_L} \times \mathbf{D_{L'}} \rightarrow \mathbf{R}$, welche die Relevanz zwischen Anfrage und Dokument ermittelt.

Prinzipiell müssen Wörter, Konzepte oder Dokumente einer Sprache bzw. deren Repräsentation denen einer anderen Sprache zugeordnet werden, um die Ähnlichkeit zu bestimmen. Bisher existierende Ansätze basieren auf einer der folgenden Ressourcen:

- *Wörterbücher oder Thesauri*: Mit Hilfe von Wörterbüchern und Thesauri können Wörter direkt von einer Sprache in eine andere übersetzt werden. Ein Wörterbuch enthält neben der Übersetzung auch Informationen zur Aussprache und grammatikalischen Eigenschaften wie Wortart, Geschlecht und Flexion. Die Begriffe eines Thesaurus sind durch Relationen miteinander verbunden, das heißt zu jedem Wort lassen sich Informationen über Synonyme, Unter- oder Oberbegriffe finden. Multilinguale Thesauri enthalten Äquivalenzrelationen zwischen Begriffen in unterschiedlicher Sprache.

- *Multilinguale Korpora*: Ein multilingualer Korpus beinhaltet Dokumente in verschiedenen Sprachen, wobei einem Dokument $d$ in der Sprache $L$ jeweils ein Dokument $d'$ in der Sprache $L'$ zugeordnet ist. Man unterscheidet hierbei zwischen vergleichbaren Korpora, bei denen die Dokumentpaare denselben Inhalt beschreiben, und parallelen Korpora, bei denen das Dokument $d'$ eine Übersetzung des Dokuments $d$ darstellt. Die Zuordnung kann auf Dokument-, Paragraph-, Satz- oder Wortebene erfolgen. Je detaillierter die Zuordnungen sind, desto schwieriger sind entsprechende Korpora verfügbar. Werden nur zwei Sprachen

verwendet, spricht man von einem bilingualen Korpus. Eine genaue Klassifikation paralleler und vergleichbarer multilingualer Korpora nehmen McEnery und Xiao (2007) vor.

### 2.3.1 Ansätze

Im sprachübergreifenden Retrieval werden die folgenden drei Vorgehensweisen unterschieden:

#### Übersetzung der Anfragen

Der am weitesten verbreitete Ansatz beschäftigt sich mit der Übersetzung der Anfrage in die Sprache(n) der zu durchsuchenden Kollektion. Anschließend wird monolinguales Retrieval durchgeführt. Grundlage für die Übersetzung stellt in vielen Fällen ein Wörterbuch dar. In der Literatur wird fast ausschließlich von Wortanfragen ausgegangen und nicht von Anfragen mittels eines Beispieldokuments, so daß die Anfrage aus wenigen Wörtern besteht. Daraus ergibt sich die größte Schwierigkeit dieser Strategie: Viele Wörter haben mehrere Übersetzungsmöglichkeiten. In der Anfrage fehlt jedoch Kontextwissen, um die richtige Übersetzung in Bezug auf das Informationsbedürfnis auszuwählen. Pirkola u. a. (2001) und Ballesteros und Croft (1997) identifizieren folgende Probleme bei der Übersetzung einzelner Anfrageterme mittels eines Wörterbuchs und nennen Lösungsvorschläge:

- ***Ambiguität (Mehrdeutigkeiten)***: Zu vielen Begriffen existieren alternative Übersetzungen mit teilweise sehr unterschiedlichen Bedeutungen. Beispielsweise hat der deutsche Begriff „Arbeit“ die englischen Übersetzungen {„activity“, „assignment“, „chore“, „employment“, „job“, „labor“, „occupation“, „paper“, „task“, „work}. Welche Variante soll nun ausgewählt werden?

  *Lösung*: Die naheliegendste Strategie ist Part-of-Speech-Tagging, bei dem zu jedem Wort der Anfrage die Wortart bestimmt wird. Ausgehend davon werden dann nur die Übersetzungsvarianten ausgewählt, die zur Wortart passen (Ballesteros (2001), Hull (1997), Pirkola u. a. (2001)). Dies ist aber nur dann geeignet, wenn die Anfrage nicht nur durch einzelne Wörter sondern wenigstens durch

einen (Teil-)Satz formuliert wird. So wird beispielsweise das Nomen „Arm“ nur zu {„arm“, „base“, „boom“} übersetzt und nicht auch noch zu {„beggarly“, „deplected“, „indigent“, „meager“, „needy“, „poor“}. Dadurch wird das Problem nicht gelöst, jedoch die Anzahl der Varianten eingeschränkt. Anspruchsvollere Verfahren zur Disambiguierung basieren auf parallelen oder monolingualen Korpora. Beispielsweise versuchen viele Ansätze die richtigen Übersetzungen über Kookkurrenzanalyse auszuwählen (Ballesteros (2001), Chen u. a. (1999), Gao u. a. (2002), Jang u. a. (1999), Maeda u. a. (2000), Monz und Dorr (2005), Liu und Jin (2005)). Das gemeinsame Auftreten zweier (oder mehrerer) Terme in einem Satz oder Dokument wird als Kookkurrenz bezeichnet. Treten die Wörter auffällig häufig zusammen auf, dann wird davon ausgegangen, daß sie voneinander abhängig sind. Durch eine Kookkurrenzanalyse wird bestimmt wie wahrscheinlich das gemeinsame Auftreten von Termen ist. Also werden alle möglichen Kombinationen der Übersetzungskandidaten untersucht und diejenigen mit geringer Wahrscheinlichkeit verworfen. Query-Expansion-Techniken finden in den Arbeiten von Adriani und van Rijsbergen (1999), Hull (1997), Ballesteros (2001), Sheridan u. a. (1997) und Sadat (2002) Anwendung. Dabei werden die Anfragen vor oder nach dem Übersetzen um weitere Begriffe erweitert, von denen angenommen wird, daß sie die Anfrage genauer beschreiben. Mit Hilfe dieser Begriffe wird der Kontext quasi erweitert, was die Übersetzung erleichtern oder den Einfluß der bereits in der Anfrage enthaltenen falschen Übersetzungen abschwächen soll. Verbreitet ist auch die Anwendung von sogenanntem Query-Structuring, bei dem alle Übersetzungsvarianten eines Terms als Synonyme aufgefaßt und in der Anfrage durch geeignete Operatoren miteinander verknüpft werden (Pirkola u. a. (2003), Ballesteros (2001)).

- ***Übersetzung von zusammengesetzten Begriffen, Wortgruppen oder Redewendungen***: Beispielsweise ergibt die Wort-für-Wort-Übersetzung des englischen Begriffes „operating system“ ins Deutsche {„Betrieb“, „betrieblich“, „in Betrieb befindlich", „Betriebs-“} und {„Anlage“, „Anordnung“, „Aufbau“, „Methode“, „Ordnung“, „System“, „Systematik“, „Verfahren“, „-wesen"}. Die Schwierigkeit besteht darin, die in diesem Fall gemeinte Variante „Betriebssystem“ auszuwählen.

*Lösung*: Mehrteilige Begriffe können in sogenannten Phrasen-Wörterbüchern nachgeschlagen werden, nachdem sie von speziellen Algorithmen als solche identifiziert wurden. Eine andere Möglichkeit besteht darin, jeden Teil des Wortes einzeln durch Nachschlagen im Wörterbuch zu übersetzen und mit Hilfe von Kookkurrenzstatistiken die richtige Übersetzungsvariante auszuwählen.

- ***Umfang des zugrundeliegenden Wörterbuchs***: Die Qualität der Übersetzung hängt entscheidend von der Domäne und der Größe des Wörterbuchs ab. Allgemeine Wörterbücher enthalten Begriffe aus den unterschiedlichsten Bereichen. Zur Übersetzung von Fachbegriffen sind sie eher weniger geeignet. Daher sind je nach Anwendungsbereich auch zusätzliche Wörterbücher zu empfehlen, die sich auf das entsprechende Thema spezialisieren. Je mehr Terme einer Anfrage mittels des Wörterbuchs korrekt übersetzt werden können, desto besser ist die Qualität der Retrievalergebnisse.

- ***Übersetzung gebeugter Wortformen***: Dieses Problem tritt vor allem in Sprachen mit starker Flexion (z.B. Finnisch oder Deutsch) auf. In Wörterbüchern sind oft nur die Grundformen eines Wortes zu finden. Gebeugte Formen können dann nicht korrekt übersetzt werden.

  *Lösung*: Hier bietet sich sogenanntes „Stemming“ an, bei dem zum Beispiel über statistische Verfahren oder basierend auf Verkürzungsregeln versucht wird, morphologische Varianten eines Wortes auf ihre Stammform zurückzuführen.

#### Übersetzung der Dokumente

Bei diesem Ansatz werden alle Dokumente einer Kollektion in die Sprache(n) der Anfrage übersetzt. Nachdem die Dokumente übersetzt wurden, werden ebenfalls monolinguale Retrievalstrategien angewandt. In der Praxis ist dieser Ansatz jedoch kaum geeignet, da Aufwand, Speicherbedarf und Rechenzeit mit der Korpusgröße stark ansteigen.

Die Übersetzung von Fließtext bringt zusätzlich zu den im vorigen Abschnitt beschriebenen Schwierigkeiten weitere Probleme mit sich. Während bei der Übersetzung der Anfrage die Wortfolge keine Rolle spielt, ist das Ziel bei der Übersetzung kompletter

Dokumente, syntaktisch und semantisch korrekte Sätze in der Zielsprache zu erzeugen. Neben der Auswahl der richtigen Übersetzungsvariante eines Begriffes muß also auch die korrekte Struktur eines Satzes gebildet werden.

Folgende Probleme treten u. a. auf:

- Ambiguität: Nicht nur Wörter sind mehrdeutig, auch ganze Sätze können uneindeutig sein. Sätze beziehen sich mitunter aufeinander und zur korrekten Übersetzung eines Satzes sind Informationen aus anderen Sätzen nötig.
- Kulturelle Unterschiede zwischen Sprachen: Für Wörter einer Sprache kann es in einer anderen keine Entsprechung geben. Sie müssen dort mit Hilfe von Wortgruppen ausgedrückt werden.
- Morphologische und syntaktische Unterschiede zwischen Sprachen: Isolierende Sprachen wie Vietnamesisch besitzen nur Wörter, die aus einem einzigen Morphem[3] bestehen, während in polysynthetischen Sprachen wie Eskimosprachen ein einzelnes Wort sehr viele Morpheme enthalten kann, für die in anderen Sprachen ein kompletter Satz benötigt wird (Jurafsky und Martin (2000)). Dadurch erschwert sich die Übersetzung von einer Sprache in eine andere.

Es existieren verschiedene Ansätze zur Übersetzung von Fließtext, die den zu übersetzenden Text unterschiedlich stark analysieren. Direkte Übersetzungsmethoden basieren auf Wörterbüchern und versuchen den Text Wort für Wort mit einfachen grammatischen Anpassungen, welche die Wortordnung oder die Morphologie (z. B. Überführung eines Verbs in die richtige Form) beeinflussen, zu übersetzen. Sie kommen ohne linguistische Theorien aus. Transfer-Methoden analysieren die Satzstruktur der Ausgangssprache und überführen diese in eine Struktur der Zielsprache, aus welcher der Satz schließlich erzeugt wird. Interlinguale Verfahren versuchen eine sprachunabhängige Repräsentation eines Satzes zu finden, aus der dann der Satz in der Zielsprache erzeugt wird. Andere Methoden sind korpusbasiert. Hier werden bilinguale Korpora verwendet, in denen Dokumentepaare einander auf Teilsatzebene zugeordnet sind. Über statistische Korrelationen können die Konzepte einer Sprache und somit die

[3] Ein Morphem ist die kleinste bedeutungstragende Einheit in einem Sprachsystem.

Übersetzungen gelernt werden. Die Auswahl der wahrscheinlichsten Übersetzung erfolgt auf Basis eines monolingualen Sprachmodells der Zielsprache[4].

### Sprachunabhängige Repräsentation

Diesem Ansatz liegt die Idee zugrunde, sowohl Dokumente als auch Anfragen in einen sprachunabhängigen Konzeptraum zu transferieren und ohne aufwendige Übersetzungsverfahren auszukommen. Einige Modelle basieren auf der Untersuchung der Vektorräume von Dokumenten in zwei verschiedenen Sprachen. Mit Hilfe statistischer oder algebraischer Verfahren wird versucht, Korrelationen zwischen den Dimensionen dieser Vektorräume herauszufinden. Daraus läßt sich eine Übersetzung erlernen ohne manuell übersetzen zu müssen. Solche Modelle benötigen einen parallelen Korpus und die Analysen sind mitunter sehr aufwendig. Zudem sind sie oft auf einzelne Domänen beschränkt.

Carbonell u. a. (1997) entwickeln eine Methode, die auf einem Generalized-Vector-Space-Modell beruht. Dumais u. a. (1997) nutzen „Latent Semantic Indexing“ (LSI) zur Konstruktion eines bilingualen semantischen Raumes. Das sprachübergreifende Relevanzmodell von Lavrenko u. a. (2002) errechnet ausgehend von einer Anfrage in der Ausgangssprache das Modell für relevante Dokumente in der Zielsprache. Die Berechnung basiert auf Wahrscheinlichkeitsmodellen und Query-Expansion-Verfahren. Potthast u. a. (2008) entwickeln ein sprachübergreifendes Konzeptraummodell, das sprachübergreifende Ähnlichkeitsanalyse ermöglicht.

[4] Weitere Informationen zu statistischen Übersetzungsverfahren sind in der Arbeit von Brown u. a. (1990) nachzulesen.

# 3 Multilinguale Plagiaterkennung

Plagiate gibt es nicht erst, seit der Zugang zu Informationen durch die Entwicklung des WWW enorm erleichtert wurde. Bereits Shakespeare soll in seinen Werken von anderen kopiert und deren Ideen übernommen haben (Weber-Wulff (2007)). Jedoch ist dieses Vorgehen erst mit der Entwicklung des Urheberrechts bzw. der Entstehung der Idee des Copyrights zu verurteilen. Das „Gesetz zum Schutz des Eigentums an Werken der Wissenschaft und Kunst“, das 1837 in Preußen erlassen wurde, gilt als das erste moderne Urheberrecht (Lumb (2008)). Auch Thomas Mann, Vladimir Nabokov (Welt Online (2004)) oder Bertolt Brecht (Weber-Wulff (2007)) werden verdächtigt, sich Anregungen bei anderen Autoren geholt zu haben. Die Grenzen zwischen erlaubter Anregung und zu verurteilendem Kopieren sind fließend.

Nicht nur in Texten sondern bei allen Werken, die im Urheberrechtsgesetz genannt werden, sind Plagiatvergehen möglich. Danach gehören laut §2 Absatz 1 UHG

> *„[z]u den geschützten Werken der Literatur, Wissenschaft und Kunst . . . 1. Sprachwerke, wie Schriftwerke, Reden und Computerprogramme; 2. Werke der Musik; 3. pantomimische Werke einschließlich der Werke der Tanzkunst; 4. Werke der bildenden Künste einschließlich der Werke der Baukunst und der angewandten Kunst und Entwürfe solcher Werke; 5. Lichtbildwerke einschließlich der Werke, die ähnlich wie Lichtbildwerke geschaffen werden; 6. Filmwerke einschließlich der Werke, die ähnlich wie Filmwerke geschaffen werden; 7. Darstellungen wissenschaftlicher oder technischer Art, wie Zeichnungen, Pläne, Karten, Skizzen, Tabellen und plastische Darstellungen.“* Urheberrechtsgesetz (1965).

Ebenfalls von Plagiarismus betroffen sein können Handelsmarken, wie sie im „Gesetz

über den Schutz von Marken und sonstigen Kennzeichen“ definiert sind (Markengesetz (1994)).

Die Motivation zur Erkennung von Plagiaten ist daher vielfältig. Im wirtschaftlichen Bereich ist es wichtig, finanziellen Schaden zu vermeiden und zu beweisen, daß beispielsweise der Quellcode der Software einer Firma nicht von dem einer anderen abgeschrieben wurde. Im akademischen Bereich muß überprüft werden, inwiefern Abschlußarbeiten tatsächlich in eigenständiger Leistung erbracht wurden.

Diese Arbeit beschränkt sich auf Plagiate in Form von Texten. In den folgenden Abschnitten werden diese zunächst charakterisiert und Methoden der automatischen Erkennung, die auf den im vorherigen Kapitel genannten Retrieval-Techniken basieren, vorgestellt. Anschließend wird näher auf eine spezielle Form des Plagiats eingegangen, das durch Übersetzungen entsteht.

## 3.1 Charakterisierung von Plagiaten

Maurer u. a. (2006) führen verschiedene Arten des Plagiierens und auch verschiedene Ursachen für das Entstehen von Plagiaten an. Sie unterscheiden zwischen

- *versehentlichem Kopieren*, das durch fehlendes Zitierverständnis oder Unsicherheiten in der Definition eines Plagiats entsteht,
- *unbeabsichtigtem Kopieren*, dessen Ursache in der enormen Menge an Informationen zu einem bestimmten Thema zu suchen sind: Der Verfasser eines Textes wird durch sie derart beeinflußt, daß er unbewußt Ideen oder Formulierungen anderer Autoren übernimmt und der Meinung ist, sie stammten von ihm selbst,
- *Selbstkopien*, das heißt die Verwendung eigener, bereits veröffentlichter Arbeiten, ohne mit einer Quellenangabe auf diese zu verweisen, sowie
- *bewußtem Kopieren*, bei dem teilweise oder vollständig Werke anderer ohne Quellenangabe übernommen werden.

Dabei können verschiedene Typen des Plagiierens unterschieden werden. Bei einer Eins-zu-eins-Kopie wird Wort für Wort aus einer anderen Arbeit übernommen und es fehlen Zitierzeichen oder Quellenangaben. Ebenso gelten Texte als Plagiate, die durch

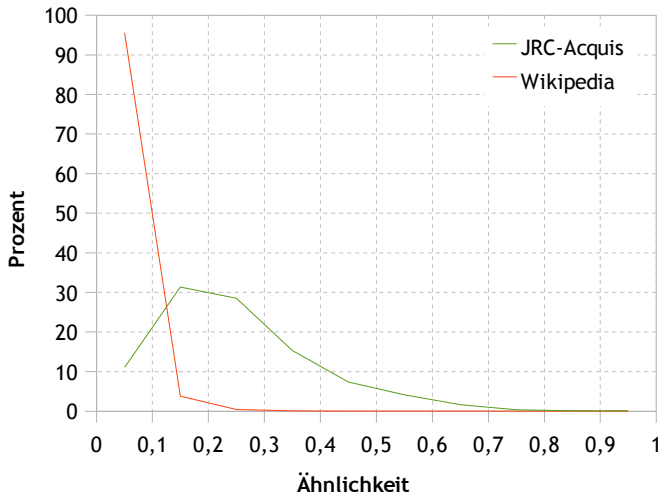


***Abbildung 3.1:*** *Paarweise Ähnlichkeit von Dokumenten. Untersucht wurden 10000 zufällige Wikipedia-Dokumente sowie 10000 Dokumente aus dem JRC-Acquis-Korpus, der juristische Texte enthält.*

Umformulieren anderer Texte entstanden sind. Sie werden lediglich minimal verändert, indem die Struktur eines Satzes umgestellt wird oder einzelne Wörter durch Synonyme ersetzt werden. Auch die Übernahme des strukturellen Aufbaus einer Arbeit, beispielsweise durch thematisch gleiche Kapitel, oder die Übernahme einer Argumentationskette können als Plagiat angesehen werden, da die Ideen eines anderen Autors übernommen werden. Maurer führt auch Texte als Plagiate auf, in denen zwar eine Quelle genannt wird, der übernommene Abschnitt aber keine Zitierzeichen aufweist oder in eigenen Worten wiedergegeben ist. Schließlich zählen auch Übersetzungen ohne Quellenangaben zu den Plagiaten. Hier werden Passagen aus fremdsprachigen Arbeiten übersetzt, in den eigenen Text übernommen und als eigene Formulierungen ausgegeben. Es gibt jedoch keine klare Abgrenzung bzw. Definition eines Plagiats. Die Übergänge sind fließend und es muß von Fall zu Fall entschieden werden, ob tatsächlich ein Plagiatvergehen vorliegt.

## 3.2 Automatische Erkennungsmethoden

Einem menschlichen Leser fällt es oft nicht allzu schwer, ein Plagiat zu entdecken. Indiz für ein Plagiat können Inkonsistenzen im Schreibstil wie beispielsweise Wechsel von sehr guten und schlechten Formulierungen, Wechsel der Zeitform, gehäuftes

Auftreten von Fremdwörtern oder auch uneinheitliche Überschriften sein. Auch das Gefühl, man hätte einen Abschnitt in der Form schon einmal irgendwo gelesen, kann ein erster Verdacht für ein Plagiatvergehen sein. Da es unzählige Möglichkeiten gibt, ein und dieselbe Sache zu beschreiben, und jeder Mensch eine eigene Art hat, Dinge auszudrücken und zu formulieren, steigt die Wahrscheinlichkeit, daß ein Text Vorlage für einen anderen war, je ähnlicher sich beide in Wortwahl, Satzstruktur und Aufbau sind. Abbildung 3.1 zeigt, daß zwei beliebige Dokumente äußerst selten ein hohe Ähnlichkeit zueinander aufweisen. Es wurden 10000 Wikipedia-Dokumente zufällig ausgewählt und jeweils die Ähnlichkeit zwischen zwei Dokumenten bestimmt. Der Graph zeigt die prozentuale Verteilung der Ähnlichkeiten: Die meisten Dokumente weisen eine sehr geringe Ähnlichkeit zueinander auf ($< 0.1$) und nur sehr wenige erreichen einen Wert über 0.7. Selbst beim Vergleich von Dokumenten, die aus dem selben Themengebiet stammen, sind Ähnlichkeiten über 0.7 sehr unwahrscheinlich. Dies wird aus dem Vergleich von 10000 Dokumenten aus dem JRC-Acquis-Korpus, der juristische Texte enthält (siehe Kapitel 4.1), ersichtlich. Große Ähnlichkeit zwischen Dokumenten und Inkonsistenzen im Schreibstil werden auch bei der maschinellen Plagiaterkennung, die die Analyse wesentlich größerer Mengen von Dokumenten ermöglicht, ausgenutzt. Zu ihren Zielen gehört es, in einem verdächtigen Dokument alle Passagen zu identifizieren, die nicht vom angegebenen Autor stammen, und deren Quelle zu bestimmen.

Der Prozeß gliedert sich in zwei grundlegende Schritte, wie in Abbildung 3.2 dargestellt. Zu Beginn werden während der Analysephase die verdächtigen Passagen eines Dokuments ermittelt. Hohe Ähnlichkeit zwischen Dokumenten und Inkonsistenzen im Schreibstil alleine sind jedoch keine ausreichenden Kriterien zur sicheren Erkennung eines Plagiats. Diese Eigenschaften können auch bei korrekten Zitaten auftreten. Entscheidend sind die fehlenden Quellenangaben. Daher folgt nach der Analyse ein Nachverarbeitungsschritt, um auszuschließen, daß die verdächtigen Abschnitte korrekte Zitate sind. Prinzipiell kann zwischen zwei Formen der Analyse unterschieden werden: der externen Analyse mit Hilfe eines Referenzkorpus, die auf Vergleichen zwischen Dokumenten beruht, und der intrinsischen Analyse, bei der das Dokument selbst auf Inkonsistenzen untersucht wird.

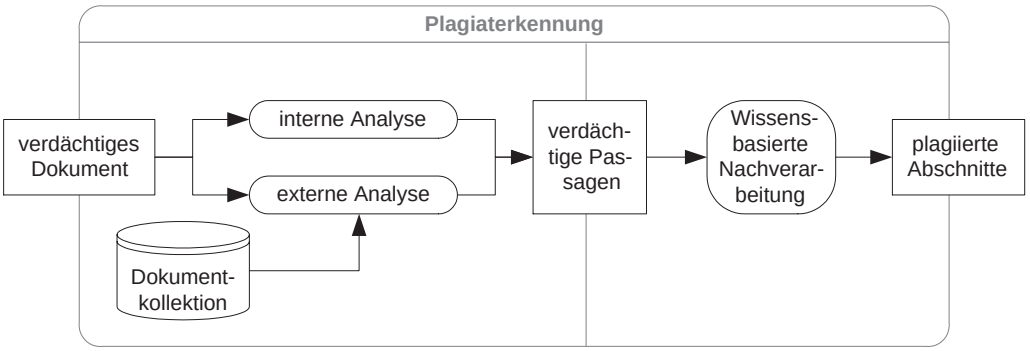


***Abbildung 3.2:*** *Schritte der Plagiaterkennung. Während des Analyseschrittes werden verdächtige Abschnitte eines Dokuments erkannt, die dann in der Nachverarbeitungsphase als Plagiate bestätigt oder als korrekte Zitate verworfen werden.*

### 3.2.1 Externe Analyse

Externe Analyseverfahren vergleichen Abschnitte eines verdächtigen Dokuments mit Abschnitten anderer Dokumente. Da der Referenzkorpus jedoch häufig sehr groß ist (als mögliche Quelle für ein Plagiat können alle Textdokumente des WWW in Frage kommen), muß die Menge $D$ auf eine kleinere Menge Kandidatendokumente $D'$ eingeschränkt werden. Dies geschieht durch einige der in Kapitel 2 beschriebenen heuristischen oder Hashing-basierten Verfahren zur Ähnlichkeitssuche.

Die Frage, ob ein Dokument $d_q$ plagiierte Passagen enthält, kann also als Retrievalaufgabe formuliert werden:

> *„Gegeben ist eine Dokumentkollektion $D$. Enthält $d_q$ einen Abschnitt $p_q$, zu dem ein Dokument $d_x \in D$ gefunden werden kann, das einen Abschnitt $p_x$ enthält, so daß unter einem Retrievalmodell $\mathcal{R}$ die Ähnlichkeit $\varphi_{\mathcal{R}}$ zwischen $p_q$ und $p_x$ nahe 1 ist?“* (Stein u. a. (2007)).

Es existieren zwei grundlegende Ansätze zur Ermittlung der Menge der Kandidatendokumente $D'$ (siehe Abbildung 3.3): heuristisches Retrieval mit anschließender detaillierter Analyse der Kandidatendokumente und Hashing-basiertes Retrieval. Abbildung 3.4 zeigt noch einmal eine Übersicht verschiedener Plagiatvergehen und ordnet ihnen die im folgenden vorgestellten Erkennungsmethoden zu.

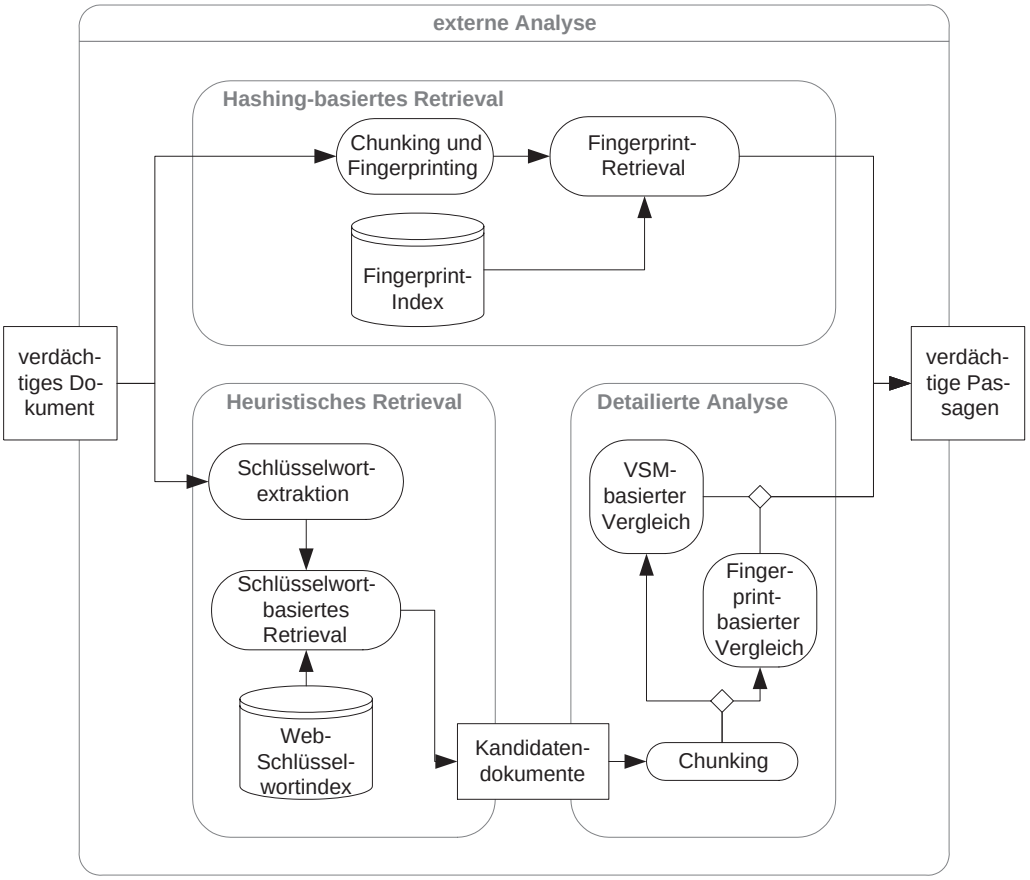


***Abbildung 3.3:*** *Externe Analyse im Detail.*

### Heuristisches Retrieval

Die erste Methode basiert auf Anfragen mit Schlüsselwörtern. Dazu werden aus dem verdächtigen Dokument Schlüsselwörter extrahiert[1] und mit Hilfe geeigneter Heuristiken die Schlüsselwortindizes herkömmlicher Suchmaschinen angefragt. So wird die Kollektion $D$ auf eine kleinere Menge Dokumente $D'$ reduziert, die thematisch sehr ähnliche Dokumente bezüglich des Ausgangsdokuments enthält. Aufgrund der hohen Ähnlichkeit sind sie wahrscheinliche Quelle für ein Plagiatvergehen. In einem weiteren Analyseschritt werden die Dokumente in $D'$ eingehender untersucht.

[1]Typische Verfahren dazu werden in Klüger (2006) beschrieben und miteinander verglichen.

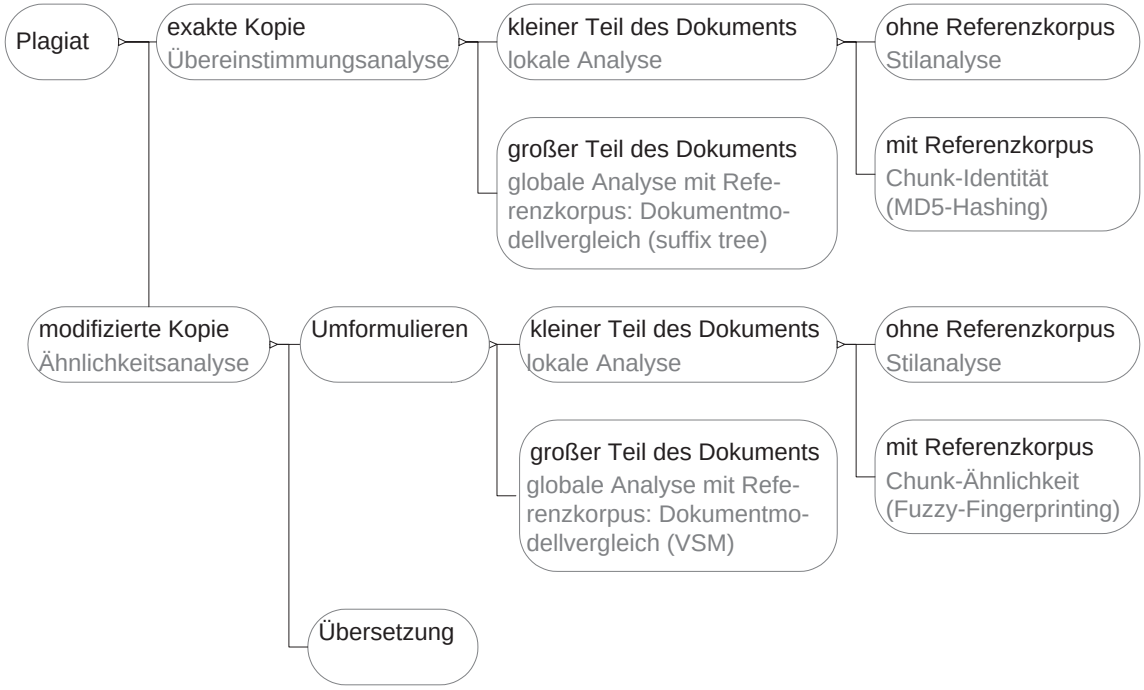


***Abbildung 3.4:*** *Taxonomie der Plagiatvergehen mit zugeordneten konkreten Methoden der Erkennung (vgl. Meyer zu Eißen und Stein (2006)).*

### Detaillierte Analyse

Die genauere Analyse kann beispielsweise auf einem Vergleich des Dokumentmodells beruhen. Mit Hilfe des Vektorraummodells werden sowohl vom verdächtigen Dokument $d_q$ als auch vom möglichen Quelldokument $d_x$ die Indexterme, wie in Kapitel 2.2.1 beschrieben, als Vektoren repräsentiert und die Ähnlichkeit zwischen ihnen durch das Kosinusmaß bestimmt. Die Ähnlichkeit steigt mit der Menge der übereinstimmenden Wörter. Ist sie sehr hoch, liegt der Verdacht nahe, daß $d_q$ ein Plagiat von $d_x$ ist. Da im Vektorraummodell jegliche Information über die Reihenfolge der Terme verloren geht, kann es zur Erkennung von Plagiaten eingesetzt werden, die durch Umformulieren entstanden sind. Es ist zu bemerken, daß zwei Dokumente maximale Ähnlichkeit aufweisen können, obwohl kein einziger ihrer Sätze identisch ist. Zur Identifikation identischer Passagen ist ein auf Suffixbäumen basierendes Dokumentmodell geeignet, da es die Wortreihenfolge abbilden kann. Für nähere Informationen dazu wird auf Meyer zu Eißen u. a. (2005) verwiesen. Diese Vergleichsmethoden sind besonders für größere Textabschnitte oder Plagiate, die das gesamte Dokument umfassen, geeignet.

Eine weitere Möglichkeit der Entdeckung von Plagiaten ist der Vergleich basierend auf Fingerprints. Identische Kopien werden ausfindig gemacht, indem das Dokument in kleine Abschnitte (Chunks) zerteilt wird und für jeden ein Hashwert (beispielsweise durch den MD5-Algorithmus (Rivest (1992))) berechnet wird. Identische Hashwerte verweisen auf identische Chunks. Ein einziges verändertes Zeichen kann bei diesem Ansatz bereits einen völlig anderen Hashwert bewirken. Daher sind möglichst kleine Chunks notwendig, was den Aufwand des Vergleichs und auch den Speicherbedarf erhöht. Zu kleine Chunks bewirken, daß eine Übereinstimmung angezeigt wird, obwohl dies unter den Gesichtspunkten der Plagiatanalyse nicht gerechtfertigt ist. Fuzzy-Fingerprinting erlaubt das Erkennen ähnlicher Chunks, durch die Anwendung einer ähnlichkeitssensitiven Hashfunktion. Hierbei wird die Analyse wesentlich größerer Abschnitte, die auch das komplette Dokument umfassen können, ermöglicht.

#### Hashing-basiertes Retrieval

Der zweite Retrieval-Ansatz nutzt Hashing-basierte Indizierung. Aus $\mathbf{D}$ wird ein neuer Index erstellt, der jedes Dokument als Fingerprint repräsentiert, wie dies in Kapitel 2.2.2 beschrieben wurde. Auch das verdächtige Dokument wird durch einen Fingerprint dargestellt. Mit dessen Hilfe können sehr schnell alle ähnlichen Dokumente ermittelt werden. Werden statt des kompletten Dokuments einzelne Abschnitte indiziert, können die verdächtigen Passagen direkt durch Anfragen des Indexes identifiziert werden und der Schritt der detaillierten Analyse entfällt.

Im Gegensatz zu herkömmlichen Hashing-Verfahren, bei denen identische Hashwerte mit sehr großer Sicherheit auf identische Abschnitte verweisen, ist bei ähnlichkeitssensitivem Hashing die Wahrscheinlichkeit höher, daß Abschnitte fälschlicherweise als ähnlich identifiziert (falsch positiv) oder Passagen zu unrecht als nicht ähnlich bewertet wurden (falsch negativ). Die Ergebnisse von Meyer zu Eißen und Stein (2006) zeigen jedoch, daß ähnlichkeitssensitives Hashing, insbesondere Fuzzy-Fingerprinting, robust und für die Plagiatanalyse sehr gut geeignet ist. Gerade für größere Textabschnitte ist es den herkömmlichen Hashing-Verfahren überlegen.

### 3.2.2 Intrinsische Analyse

Ausgehend von der Annahme, daß jeder Mensch einen eigenen Schreibstil hat[2], können Plagiate auch durch Analyse eines einzelnen Dokuments ohne Vergleiche mit einem Referenzkorpus entdeckt werden. Grundlage der Erkennung sind Veränderungen in Zeitform oder Person, aber auch Änderungen in der Ausdrucksweise oder der Zeichensetzung. Die Berechnung des Schreibstils innerhalb eines Dokuments erfolgt auf Basis einer Menge von Stilmerkmalen, die auf Textstatistiken basieren, die beispielsweise die Häufigkeit bestimmter Satzzeichen ermitteln oder die durchschnittliche Wortlänge auswerten. Auch syntaktische Merkmale wie die Satzlänge oder Strukturmerkmale wie die durchschnittliche Abschnitts- oder Kapitellänge werden bestimmt. Aussagen über die Komplexität eines Textes und den Umfang des Wortschatzes eines Autors können besonders durch die Analyse verschiedener Wortklassen getroffen werden. Dazu werden die Häufigkeit einzelner Wortklassen und die Anzahl von Stopwörtern oder Fremdwörtern bestimmt. Aus diesen Merkmalen ergibt sich eine individuelle Charakteristik.

Nach der Bestimmung des globalen Stils eines Dokuments wird dieses in einzelne Abschnitte von 40 bis 200 Wörtern zerlegt und für jeden der Abschnitte ebenfalls eine Stilanalyse durchgeführt (Meyer zu Eißen und Stein (2006)). Die Werte werden mit denen des gesamten Dokuments verglichen und der Grad der Abweichung bestimmt. Ist die Differenz für einen Großteil der Merkmale eines Abschnitts nahe der maximal möglichen Differenz, deutet dies auf unterschiedliche Autoren und ein Plagiatvergehen hin. Kopiert ein Autor allerdings ein komplettes Dokument, dann scheitert dieser Ansatz, da der Schreibstil innerhalb des Dokuments einheitlich ist. Für diese Art des Plagiats können nur Verfahren erfolgreich sein, die auf einem Vergleich mit anderen Dokumenten basieren.

[2] In ihrer Untersuchung haben Baayen u. a. (2002) gezeigt, daß sich der Schreibstil verschiedener Autoren voneinander unterscheidet, selbst wenn die Texte das gleiche Thema behandeln und die Autoren über ähnliches Hintergrundwissen verfügen. Auch die Untersuchungen von Koppel und Schler (2003) zeigen, daß sich Autoren anhand unterschiedlicher stilistischer Merkmale eines Textes erkennen lassen.

## 3.3 Sprachübergreifende Analyse

Die sprachübergreifende Plagiatanalyse beschäftigt sich mit der Entdeckung von Plagiaten, die durch Übersetzungen entstanden sind. Formal läßt sich dies wie folgt beschreiben: Es existiert ein verdächtiges Dokument $d_q$ in der Sprache $L$. Ebenfalls gegeben ist eine Kollektion von Dokumenten $D$ in der Sprache $L'$. Analog zur monolingualen Analyse stellt sich die Frage, ob $d_q$ einen Abschnitt $p_q$ enthält, zu dem ein Dokument $d_x \in D$ gefunden werden kann, das einen Abschnitt $p_x$ enthält, so daß unter einem Retrievalmodell $\mathcal{R}$ die Ähnlichkeit $\varphi_{\mathcal{R}}$ zwischen $p_q$ und $p_x$ nahe 1 ist.

Prinzipiell sind alle Verfahren, die für die Plagiatart „Umformulieren“ anwendbar sind, auch für Übersetzungsplagiate denkbar. Die Schwierigkeit besteht darin, sprachübergreifend die Ähnlichkeit zu bestimmen und während des Retrievals unterschiedliche Sprachen von Anfrage und Ergebnis zu unterstützen. Abbildung 3.5 zeigt schematisch die zusätzlichen Schritte, die für die sprachübergreifende externe Analyse notwendig sind. Dargestellt sind drei mögliche Alternativen des Retrievals, welche in den nächsten Abschnitten im Detail vorgestellt werden. Voraussetzung für alle drei Varianten ist die Erkennung der Sprache $L$ des zu untersuchenden Dokuments $d_q$, um sprachspezifische Algorithmen, die während der Analyse notwendig sind, anwenden zu können. Nach dem Retrieval und der damit verbundenen Bestimmung der Menge der Kandidatendokumente $D'$, erfolgt deren detaillierte Analyse, welche jedoch nicht Gegenstand dieser Arbeit ist.

### 3.3.1 Monolinguales heuristisches Retrieval mit übersetzten Schlüsselwörtern

Eine Möglichkeit, sprachübergreifend nach eventuellen Ursprungsdokumenten zu suchen, besteht darin, aus dem verdächtigen Dokument $d_q$ eine Menge von $n$ Schlüsselwörtern zu extrahieren, die das Dokument charakterisieren, und diese dann in die Zielsprache $L'$ zu übersetzen. Der Schlüsselwortindex einer Kollektion $D$ in $L'$ wird dann mit den Schlüsselwörtern angefragt. Das heißt, nach der Übersetzung werden monolinguale, heuristische Retrievalstrategien angewandt, die beispielsweise auf

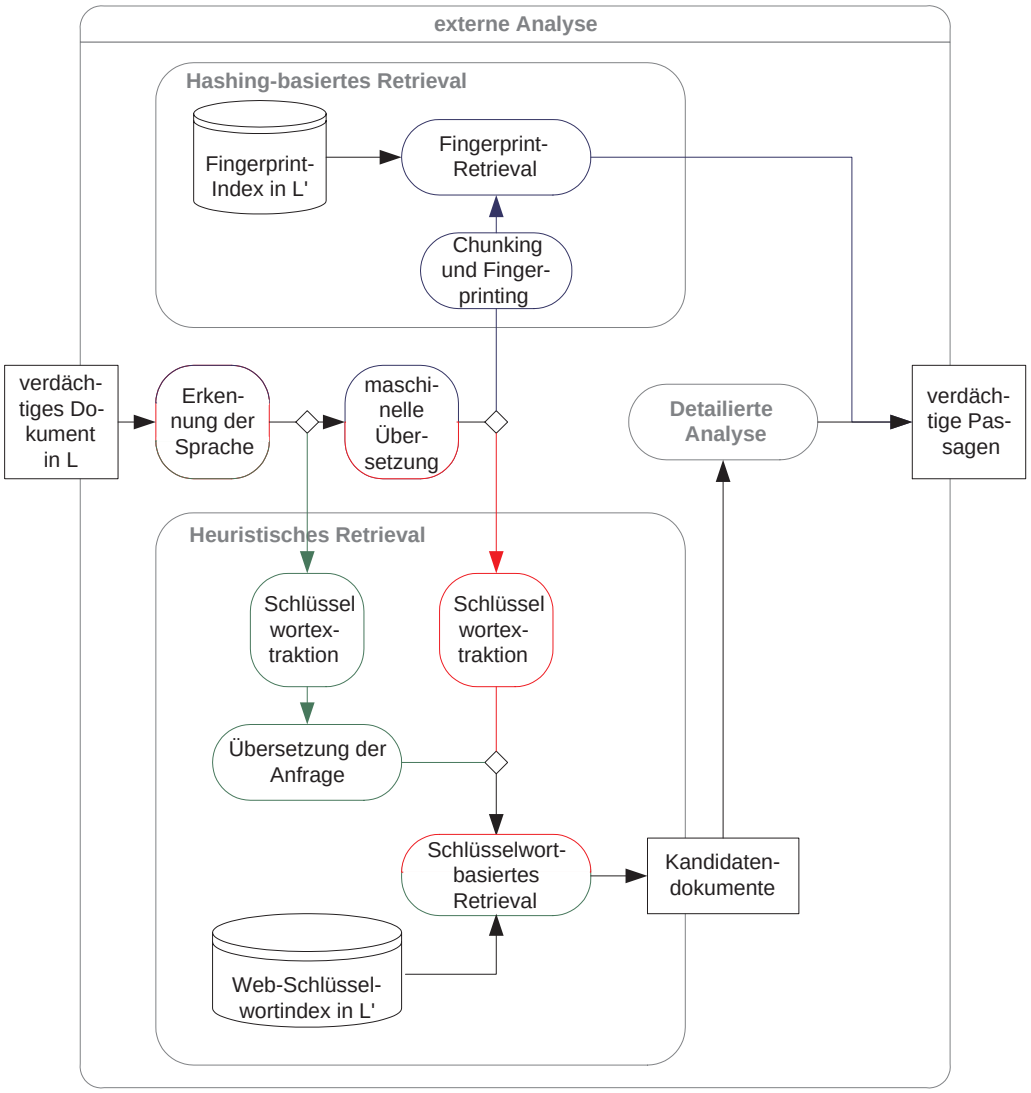


***Abbildung 3.5:*** *Retrievalschritte in der sprachübergreifenden Plagiatanalyse. Die farbliche Differenzierung kennzeichnet die drei verschiedenen Ansätze.*

dem Vektorraummodell basieren. Diejenigen Dokumente $d_x$, die am besten zu den gegebenen Schlüsselwörtern passen, werden so in die Menge der Kandidatendokumente $D'$ aufgenommen und anschließend näher untersucht.

In dieser Arbeit wurde ein wörterbuchbasiertes Übersetzungsverfahren implementiert. Der komplette Prozeß basiert auf folgenden Schritten und ist in Abbildung 3.6 bis zum Schlüsselwort-Retrieval dargestellt:

- ***Extraktion der Schlüsselwörter***: Die das verdächtige Dokument beschreibenden Schlüsselwörter werden mit Hilfe der von Klüger (2006) implementierten Bibliothek zur Schlüsselwortextraktion aus den Dokumenten entnommen. Von

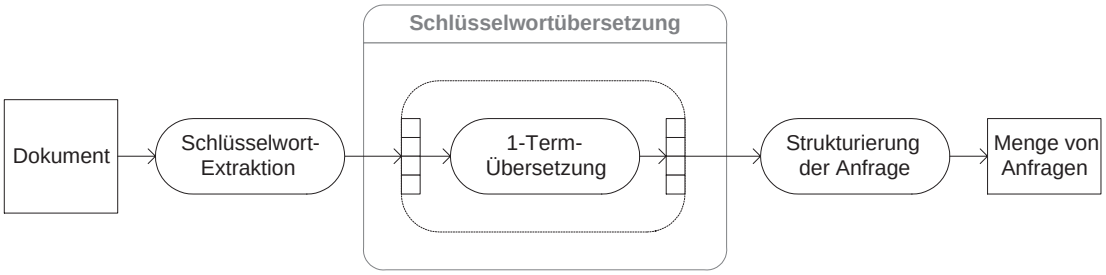


***Abbildung 3.6:*** *UML-Aktivitätsdiagramm. Schritte zur Übersetzung von Schlüsselwörtern.*

den zur Verfügung stehenden Extraktoren wurde der „CooccurrenceExtractor“ verwendet, welcher die Schlüsselbegriffe mittels Kookkurrenzanalyse ermittelt. Als Schlüsselwörter können wahlweise einzelne Wörter oder Wortgruppen extrahiert werden.

- ***Übersetzung***: Im zweiten Schritt werden die Schlüsselbegriffe übersetzt. Dazu wird jeder Begriff zu Beginn in einem Wörterbuch nachgeschlagen (siehe Abbildung 3.7). Kann der Begriff gefunden werden, war der Übersetzungsvorgang erfolgreich und alle Übersetzungsvarianten werden im nächsten Schritt weiterverarbeitet. Ist der Begriff nicht im Lexikon enthalten, kann abhängig von der Sprache seine Zerlegung in einzelne Komponenten (engl. decompounding) erfolgversprechend sein. Das einfachste Verfahren ist es, den Begriff bei bestimmten Zeichen wie Leerzeichen oder Bindestrich zu trennen. Komplexere Methoden, die jedoch nicht implementiert wurden, können auch zusammengesetzte Substantive wie „*Nachrichtensendung*“ in „*Nachrichten*“ und „*Sendung*“ zerlegen. Für diese Terme findet dann eine *n-Term-Übersetzung* statt (siehe Abbildung 3.8).

  Während der n-Term-Übersetzung werden zunächst die einzelnen Terme getrennt voneinander übersetzt. Anschließend werden die Übersetzungskandidaten aller Wörter miteinander kombiniert und aus diesen Kombinationen wird mittels Kookkurrenzanalyse die wahrscheinlichste Variante ausgewählt. Hier wurde auf die Methode von Monz und Dorr (2005) zurückgegriffen. Sie bauen ein sogenanntes Kookkurrenznetzwerk für alle Begriffe auf und ermitteln iterativ, welche Kombination am häufigsten auftritt. In ihrer Arbeit gewichten sie die einzelnen Varianten und verwerfen keine der Möglichkeiten. Die Imple-

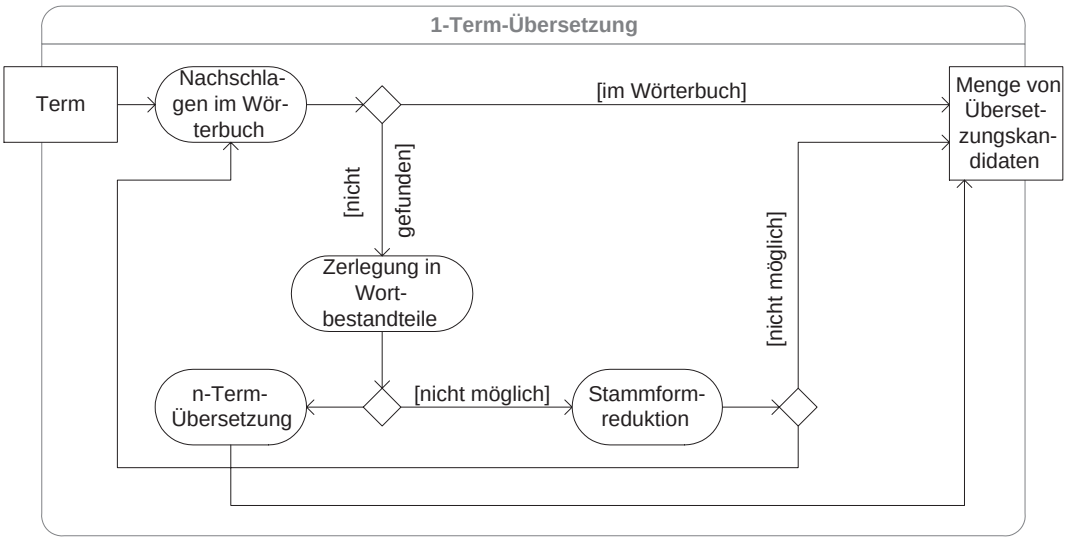


***Abbildung 3.7:*** *UML-Aktivitätsdiagramm. Übersetzung eines Schlüsselwortes im Detail.*

mentierung der n-Term-Übersetzung verwendet nur die wahrscheinlichste Kombination, da Termgewichte für Anfragen an herkömmliche Suchmaschinen nicht geeignet sind. Kookkurrenzwerte werden mit Hilfe des von Trenkmann (2008) erstellten Google-n-Gramm-Indexes ermittelt.

Ist auch nach der Zerlegung der Begriffe in ihre Bestandteile keine erfolgreiche Übersetzung möglich, wird als letztes versucht, den Term auf seine Stammform zu reduzieren (engl. stemming) und diese im Wörterbuch nachzuschlagen. Nicht übersetzbare Terme werden in der Ausgangssprache übernommen. Die einzelnen Komponenten wurden so generisch implementiert, daß einzelne Verfahren bequem ausgetauscht und an die jeweils verwendeten Sprachen angepaßte Methoden verwendet werden können. In dieser Arbeit ist die Ausgangssprache Deutsch, gesucht sind englische Dokumente. Daher wird ein Deutsch-Englisches-Wörterbuch benötigt. Verwendung findet das Wörterbuch „BEOLINGUS“ der TU-Chemnitz (TU Chemnitz (2007)), das über 394.000 Übersetzungen enthält. Dies können sowohl einzelne Begriffe als auch Wortgruppen oder Redewendungen sein. Daher wird zu Beginn jedes Schlüsselwort, egal ob einzelner Term oder mehrteiliger Begriff, im Wörterbuch nachgeschlagen, bevor die n-Term-Übersetzung stattfindet. Für das Stemming deutscher Begriffe wird der

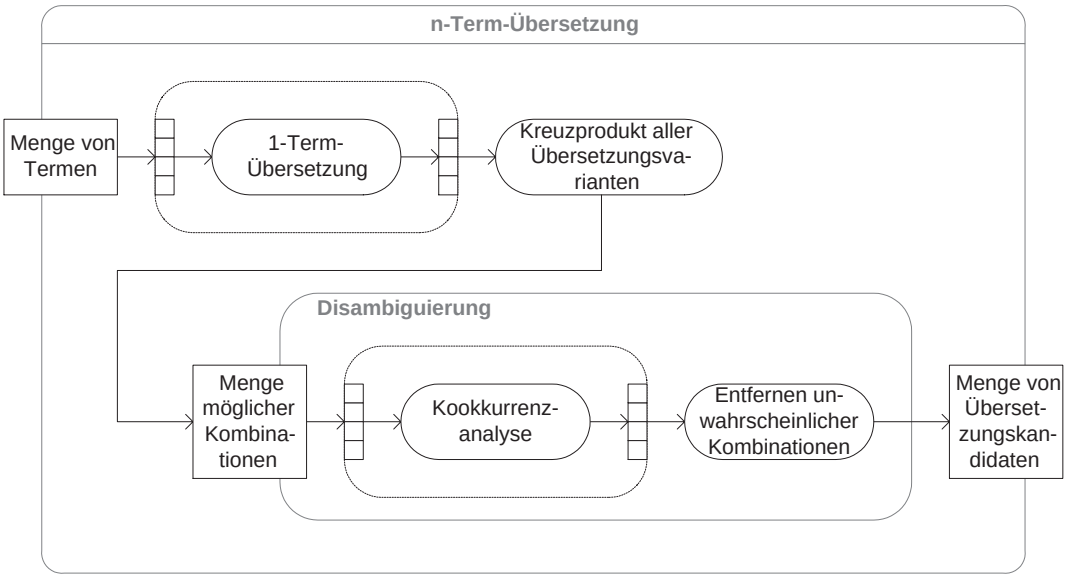


***Abbildung 3.8:*** *UML-Aktivitätsdiagramm. Übersetzung eines mehrteiligen Schlüsselwortes.*

Snowball-Stemmer[3] verwendet. Nach diesem Schritt ist jeder der $i$ extrahierten Schlüsselbegriffe $s_i$ in der Sprache $L$ in die Sprache $L'$ übersetzt worden:

$$s_i^L \Rightarrow \{s_{i;1}^{L'}, s_{i;2}^{L'}, \ldots, s_{i;k}^{L'}\}$$

Die Menge $S^L$ enthält $n$ Schlüsselbegriffe, die Menge $S^{L'}$ umfaßt $m = \sum_{i=1}^{n} m_i$ Begriffe, wobei $m_i$ die Anzahl der Übersetzungen des $i$-ten Schlüsselwortes angibt.

- ***Strukturierung der Anfrage und monolinguales Retrieval***: Aus den übersetzten Schlüsselbegriffen wird nun nicht eine einzige Anfrage erzeugt, sondern eine Menge von Anfragen, die jeweils eine Teilmenge aller Schlüsselbegriffe enthalten. Die Ergebnisse aller Anfragen werden kombiniert. Dazu wird das Verfahren „Close-End-Query“ von Shapiro und Taksa (2003) mit geringfügigen Erweiterungen verwendet.

  Der Algorithmus erzeugt aus $x$ Schlüsselbegriffen jede mögliche Kombination $q \in Q$, die mindestens $y$ der $x$ Wörter enthält. So entstehen $|Q| = \sum_{i=y}^{x} \frac{x!}{(x-i)! \cdot i!}$ Anfragen, die der Suchmaschine übergeben werden. Diese liefert dann $|Q|$ Listen mit sortierten Suchergebnissen (eine pro Anfrage), welche zu einer einzigen sortierten Liste zusammengeführt werden müssen. Das Problem wird als

[3] http://snowball.tartarus.org/algorithms/german/stemmer.html

Ranking-Fusionierung bezeichnet. Eine einfache Heuristik zu dessen Lösung ist die folgende: Ausgehend von der Annahme, daß ein Suchergebnis umso relevanter ist, je häufiger es in den Listen auftaucht, werden die einzelnen Ergebnisse nach ihrer Häufigkeit absteigend sortiert. Dabei werden allerdings nicht alle Ergebnisse einer Liste betrachtet, sondern beispielsweise nur die ersten 10 oder die ersten 100.

Damit eine Anfrage nicht mehrere Übersetzungsvarianten des gleichen Terms enthält, wurde der Algorithmus so variiert, daß für das $i$-te Schlüsselwort $s_i$ jeweils ein Term aus den $k$ Übersetzungen von $s_i^L$ ausgewählt wird. Damit steigt die Anzahl der möglichen Kombinationen weiter an auf:

$$|Q| = \begin{cases} \sum_{s=y}^{x}(\sum_{i=1}^{x-s+1}(\prod_{t=i}^{i+s-2} m_t \sum_{j=i+s-1}^{x} m_j)) & y > 1 \\ \sum_{l=1}^{x} m_l + \sum_{s=y+1}^{x}(\sum_{i=1}^{x-s+1}(\prod_{t=i}^{i+s-2} m_t \sum_{j=i+s-1}^{x} m_j)) & y = 1 \end{cases}$$

Beispielsweise hat die deutsche Anfrage

$$S^L = \{\textit{Entwicklung, Ölpreis, weltweit}\}$$

die englische Übersetzung

$$\begin{aligned} S^{L'} &= \{S_1^{L'}, S_2^{L'}, S_3^{L'}\} \text{ mit} \\ S_1^{L'} &= \{\textit{development, growth, deployment}\}, \\ S_2^{L'} &= \{\textit{oil price}\}, \\ S_3^{L'} &= \{\textit{global, universal, worldwide}\} \end{aligned}$$

Daraus werden die Teilanfragen nach dem oben beschriebenen Verfahren erstellt. Sie sollen 1 bis 3 Begriffe enthalten, d.h. $y = 1, x = 3$. Folgende 31 Kombinationen der Schlüsselwörter sind möglich:

*Q = {{development}, {growth}, {deployment}, {oil price}, {global}, {universal}, {worldwide}, {development, oil price}, {development, global}, {development, universal}, {development, worldwide}, {growth, oil price}, {growth, global}, {growth, universal}, {growth, worldwide}, {deployment, oil price}, {deployment, global}, {deployment, universal}, {deployment, worldwide}, {oil price, global}, {oil price, universal}, {oil price, worldwide}, {development, oil price, global},*

*{development, oil price, universal}, {development, oil price, worldwide}, {growth, oil price, global}, {growth, oil price, universal}, {growth, oil price, worldwide}, {deployment, oil price, global}, {deployment, oil price, universal}, {deployment, oil price, worldwide}}*

Die hier beschriebene Vorgehensweise basiert auf einem Verfahren von Ballesteros (2001), das nach eingehender Literaturrecherche und einem Vergleich der in Kapitel 2.3 genannten Übersetzungsverfahren von Schlüsselwörtern hinsichtlich Qualität, Komplexität und Umsetzbarkeit am besten geeignet erscheint. Die endgültige Implementierung unterscheidet sich jedoch in einigen Punkten:

- Übersetzung: Die Methode von Ballesteros basiert auf Anfragen in Satzform wie „Hat der Teddybär in der Welt an Popularität gewonnen oder verloren?“. Daher verwendet sie einen Part-of-Speech-Tagger, um Wortgruppen zu identifizieren und die Anzahl der Übersetzungsvarianten anhand der Wortart zu minimieren. Ein POS-Tagger ist in dieser Arbeit nicht einsetzbar, da dieser anhand einzelner Wörter keine oder nur sehr ungenaue Rückschlüsse über die Wortart ziehen kann. In Ballesteros' Anwendungsfall wurden zudem spanische Anfragen ins Englische übersetzt. Daher verwendet sie andere Methoden zur morphologischen Verarbeitung. Möglicherweise sind diese anspruchsvoller als die in dieser Arbeit umgesetzten. Sie macht weiterhin keine Angaben zur Art der verwendeten Kookkurrenzanalyse.

- Strukturierung der Anfrage: Ballesteros baut aus allen Übersetzungskandidaten eine einzige Anfrage auf. Das von ihr verwendete Retrieval-System INQUERY (Callan u. a. (1992)) erlaubt verschiedene Operatoren zur Gruppierung einzelner Begriffe. Der wichtigste ist der sogenannte „#syn“-Operator, der alle Übersetzungsvarianten eines einzelnen Terms als Synonyme betrachtet. In herkömmlichen Suchmaschinen entspräche dieser einer ODER-Verknüpfung der Terme. Das Problem dieser Methode ist, daß Suchmaschinen wie Google[4], Yahoo[5] oder LiveSearch[6], die u. a. während der Plagiatanalyse verwendet werden sollen, diese Operatoren zwar unterstützen, jedoch nicht in dem Maße wie dies erfor-

[4] http://www.google.com
[5] http://search.yahoo.com
[6] http://www.live.com/

derlich ist. Um sicherzustellen, in welcher Reihenfolge die so entstandenen logischen Ausdrücke abgearbeitet werden, ist eine Klammerung der Klauseln notwendig, die jedoch nur LiveSearch unterstützt. Hinzu kommt eine Limitierung der Anzahl der Suchbegriffe. So unterstützt Google 32 Suchbegriffe, LiveSearch lediglich 10. Allein für das in Kapitel 2.3 genannte Beispiel „operating system“ wäre eine Anfrage mit 15 Schlüsselwörtern nötig. Die Anfrage sähe folgendermaßen aus:

> *(Betrieb OR betrieblich OR „in Betrieb befindlich“ OR Betriebs-) AND (Anlage OR Anordnung OR Aufbau OR Methode OR Ordnung OR System OR Systematik OR Verfahren OR wesen)*

Während des Retrieval-Prozesses in der sprachübergreifenden Plagiatanalyse ist von wesentlich mehr Schlüsselwörtern auszugehen, da diese ein gesamtes Dokument repräsentieren sollen. Daher wurde hier auf den zuvor beschriebenen Ansatz „Close-End-Query“ von Shapiro und Taksa zurückgegriffen. Zwar löst dieser das nicht das Problem der großen Menge von Schlüsselbegriffen, welche durch die Ambiguität entsteht. Jedoch ist die Länge der Teilanfragen begrenzt, weshalb er für herkömmliche Suchmaschinen geeignet ist.

- Ballesteros erweitert die Anfrage nach der Strukturierung um weitere Begriffe durch sogenannte lokale Kontextanalyse. Dadurch sollen zusätzliche potentiell relevante Dokumente gefunden werden. Dieser Schritt ist hier nicht notwendig. Eine Anfrage wird immer aus einem Dokument erzeugt, indem aus diesem Schlüsselwörter entnommen werden. Daher kann einfach die Anzahl der Begriffe erhöht werden, um mehr Kontext zur Verfügung zu stellen.

### 3.3.2 Monolinguales heuristisches Retrieval mit übersetzten Dokumenten

Um den Mangel an Kontextwissen während der Übersetzung zu umgehen, bietet sich eine zweite Möglichkeit an. Hier wird zunächst das komplette verdächtige Dokument $d_q$ in $L$ in die Sprache $L'$ der Referenzkollektion $D$ übersetzt. Anschließend erfolgt die

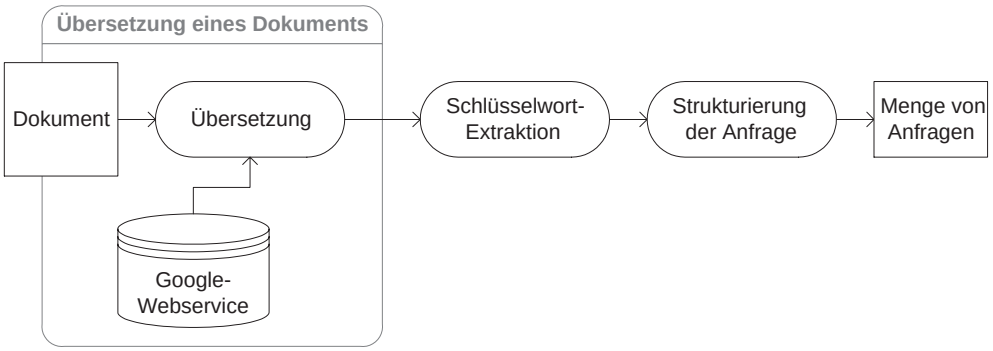


***Abbildung 3.9:*** *UML-Aktivitätsdiagramm. Übersetzung eines Dokuments.*

Extraktion von Schlüsselwörtern, anhand derer erneut monolinguales Schlüsselwortretrieval betrieben wird, um die Menge $D'$ der Kandidatendokumente zu ermitteln.

Die einzelnen Schritte bis zum Schlüsselwort-Retrieval sind im UML-Aktivitätsdiagramm in Abbildung 3.9 dargestellt. Die Implementierung greift auf den Übersetzungs-Web-Service von Google[7] zurück, der zur Übersetzung von Fließtext ein maschinelles Lernverfahren verwendet, über das jedoch keine detaillierten Informationen bekannt sind. Bisher existiert kein Übersetzungsverfahren für Fließtext, das Dokumente einer Sprache fehlerfrei und grammatisch korrekt in die Zielsprache transferieren kann. Es muß daher untersucht werden, ob die unter Umständen fehlerhaften Sätze die Retrieval-Qualität negativ beeinflussen. Aus dem übersetzten Dokument werden, wie im vorherigen Abschnitt beschrieben, mit Hilfe der Bibliothek von Klüger (2006) Schlüsselwörter entnommen. Die Strukturierung der Anfrage basiert auf dem zuvor erläuterten Verfahren „Close-End-Query“ von Shapiro und Taksa (2003) ohne die zusätzliche Erweiterung.

### 3.3.3 Hashing-basiertes Retrieval

Die dritte Alternative nutzt das vollständig übersetzte verdächtige Dokument $d_q$ für Hashing-basiertes Retrieval. Die Referenzkollektion $D'$ in $L'$ wird mittels Fuzzy-Fingerprinting indiziert. Für das verdächtige Dokument, das in $L'$ übersetzt wurde, wird ebenfalls ein Fingerprint berechnet. Bei einer Kollision der Hashwerte liegt der Ver-

[7] http://www.google.com/translate_t

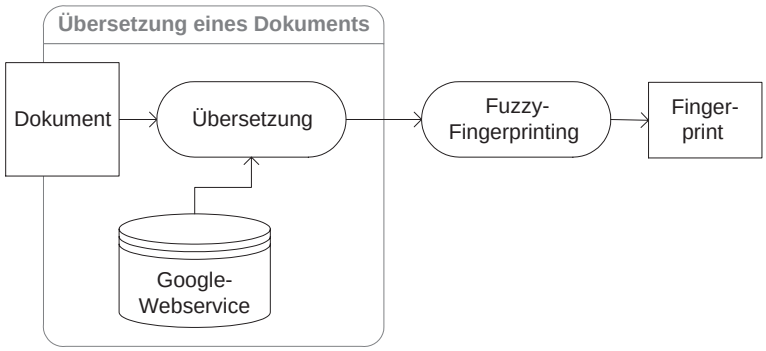


***Abbildung 3.10:*** *UML-Aktivitätsdiagramm. Vorbereitung des Dokuments für das Hashing-basierte Retrieval.*

dacht des Plagiats nahe. Abbildung 3.10 zeigt die Schritte bis zum Fingerprint-Retrieval.

Denkbar wäre auch ein Fingerprinting-Ansatz, dessen Hashfunktion sprachübergreifende Ähnlichkeit berechnet. Dieser würde die maschninelle Übersetzung ersparen und einen weiteren Geschwindigkeitsvorteil mit sich bringen. Eine solche Funktion existiert jedoch nicht, da bisher kein Konzept für sprachübergreifende Äquivalenzklassen gefunden werden konnte. Daher muß für das sprachübergreifende Retrieval mittels Hashing zunächst immer der Umweg über die Übersetzung des Anfragedokuments genommen werden.

# 4 Experimente zum Vergleich der Ansätze

In diesem Kapitel werden die drei zuvor beschriebenen Ansätze zum sprachübergreifenden Retrieval von ähnlichen Dokumenten genauer hinsichtlich ihrer Retrieval-Qualität und Laufzeit untersucht.

Dabei sind vor allem folgende Fragen von Interesse:

- Welches Verfahren hat die besten Retrieval-Eigenschaften? Mit welchem Verfahren können Übersetzungen am sichersten gefunden werden?
- Welches Verfahren zeigt das beste Laufzeitverhalten?
- Beeinflußt die Textlänge die Retrieval-Qualität?
- Wie viele Terme sollten für die schlüsselwortbasierten Verfahren extrahiert werden? Wie stark ist deren Anzahl von der Länge der Texte abhängig?
- Sollten Wortgruppen oder einzelne Wörter entnommen werden?
- Inwiefern beeinflußt die Art der Übersetzung (Schlüsselwörter vs. kompletter Text) die Retrieval-Qualität?
- Ist die Retrieval-Qualität abhängig von der Genauigkeit der Übersetzung?
- Wie gut geeignet ist Fuzzy-Fingerprinting für maschinell übersetzte Texte?

Das Kapitel unterteilt sich in vier Abschnitte. Zunächst wird der Aufbau der Testkollektionen beschrieben. Danach erfolgt die Beschreibung der Vorexperimente, die der Überprüfung der Übersetzungsqualität dienen. Schließlich werden die die Forschungsfragen betreffenden Experimente erläutert und deren Ergebnisse diskutiert.

## 4.1 Korpora und Testkollektionen

Für die Experimente werden bilinguale Korpora benötigt, die Dokumente sowohl in englischer als auch in deutscher Sprache enthalten, wobei jeweils ein deutsches einem englischen Dokument zugeordnet sein muß und umgekehrt. Ausgehend von einem deutschen Dokument soll mit Hilfe der drei vorgestellten Verfahren für das sprachübergreifende Retrieval das zugehörige englische Dokument unter allen übrigen Dokumenten gefunden werden. Um unterschiedliche Grade der Genauigkeit der Übersetzung zu simulieren, werden zwei Korpora verwendet. Der eine enthält Wikipedia-Dokumente und ist ein vergleichbarer Korpus (kurz als „Wikipedia“ bezeichnet). Die Dokumentpaare sind keine Übersetzungen voneinander, sondern behandeln lediglich das gleiche Thema. Der andere ist ein paralleler Korpus, dessen Dokumentpaare exakte Übersetzungen voneinander sind. Er beinhaltet Dokumente der „Acquis Communautaire“ und wird nachfolgend „JRC-Acquis“ genannt. Bevor die drei Verfahren direkt miteinander verglichen werden, muß die Qualität der Übersetzung getestet werden. Zur Auswertung der Übersetzungsqualität von Schlüsselwörtern wird ein von (Ballesteros, 2001, Kapitel4 und 6) beschriebenes Experiment wiederholt. Dazu sind zwei weitere Testkollektionen notwendig, die sogenannten TREC-Kollektionen, die in Abschnitt 4.1.1 beschrieben werden. Die Übersetzung vollständiger Dokumente wird anhand der Wikipedia- und der JRC-Acquis-Kollektionen überprüft, deren Aufbau in den Abschnitten 4.1.2 und 4.1.3 erläutert wird.

### 4.1.1 TREC-Kollektionen

Die TREC-Testkollektionen bestehen aus Testdokumenten, einer Menge von Anfragen sowie Informationen über die Relevanz der Dokumente zu einer bestimmten Anfrage. Ballesteros verwendet in den Experimenten, die in dieser Arbeit wiederholt werden, die für die Text-Retrieval-Konferenzen[1] zusammengestellten Testkollektionen „English ad hoc queries“ der TREC-3, im Folgenden vereinfacht als TREC-3 bezeichnet, sowie die Testkollektion „Cross-language ad hoc queries“ der TREC-6, vereinfacht TREC-6 genannt. Tabelle 4.1 zeigt die wichtigsten Eigenschaften der Kollektionen.

[1] Die Text REtrieval Conference (TREC, http://trec.nist.gov/) wird vom „National Institute of Standards and Technology“ (NIST) organisiert.

| Kollektion $D$ | Sprache | Größe (GB) | $\|D\|$ | Terme pro $d_x$ | $\|Q\|$ | Terme in $q_i$ | $\|D_{rel}\|$ pro $q_i$ |
|---|---|---|---|---|---|---|---|
| TIPSTER (TREC-3) | englisch | 2.2 | 741'856 | 260 | 20 | 10.75 | 167 |
| AP (TREC-6) | englisch | 0.75 | 242'918 | 458 | 21 | 7.8 | 59 |

***Tabelle 4.1:*** *Übersicht der TREC-Testkollektionen (Ballesteros, 2001, Seite 63). Stopwörter sind nicht in den Angaben über Anzahl der Terme pro Dokument bzw. pro Anfrage enthalten.*

***TREC-3***: Die Testkollektion enthält englische Dokumente der Nachrichtenagentur „Associated Press“ (AP) aus den Jahren 1988 bis 1990, Abstracts des „Department of Energy“, Dokumente des „Federal Register“ von 1988 bis 1989, Artikel des „Wall Street Journal“ aus den Jahren 1987 bis 1992 sowie Dokumente aus „Ziff Davis Computer-Select“. Von den 50 gegebenen englischen Anfragen verwendet Ballesteros 20 (Nr. 151-162, 164-171). Da die Kollektion nicht für die Evaluierung von CLIR-Verfahren erstellt wurde, müssen die Anfragen zunächst manuell in die Ausgangssprache übersetzt werden. Ballesteros' Ausgangssprache ist Spanisch, in dieser Arbeit ist es Deutsch[2].

```
1 <top>
2   <num> Number: 151
3   <title> Topic: Coping with overcrowded prisons
4   <desc> Description: The document will provide information on jail and prison
        overcrowding and how inmates are forced to cope with those conditions; or it
        will reveal plans to relieve the overcrowded condition.
5   <narr> Narrative: A relevant document will describe scenes of overcrowding that have
        become all too common in jails and prisons around the country. The document will
         identify how inmates are forced to cope with those overcrowded conditions, and/
        or what the Correctional System is doing, or planning to do, to alleviate the
        crowded condition.
6 </top>
```

***Listing 4.1:*** *Beispielanfrage der TREC-3-Kollektion*

[2]Prinzipiell sind die Ergebnisse eines CLIR-Verfahrens unterschiedlich je nach verwendetem Sprachpaar (Adriani und van Rijsbergen (1999)). Da Spanisch und Deutsch jedoch beide zur Sprachfamilie der indo-germanischen Sprachen gehören, wird angenommen, daß die Ergebnisse nicht extrem voneinander abweichen.

***TREC-6***: In dieser Kollektion sind englische, deutsche und französische Dokumente und Anfragen enthalten. Für den Versuch werden nur die englischen Artikel benötigt. Das sind die bereits in der TREC-3-Kollektion enthaltenen Dokumente der „Associated Press“ von 1988 bis 1990. Von den 25 Anfragen dieser Kollektion verwendet Ballesteros 19 (Nr. 1, 2, 4-7, 9-14, 16-21, 23).

Die Anfragen der Testkollektionen setzen sich aus vier Teilen zusammen, wie das Beispiel in Listing 4.1 zeigt: der ID der Anfrage, dem Titel, einer kurzen Beschreibung und einer langen Beschreibung. Nachfolgend wird von einer kurzen Anfrage gesprochen, wenn der Titel als Anfrage verwendet wird, und von einer langen Anfrage, wenn das Element `<desc>` als Anfrage dient. Weder von Ballesteros noch in dieser Arbeit wird das Element `<narr>` benutzt.

Statt des INQUERY-Retrieval-Systems wurde die Java-Bibliothek „Lucene“ [3] zum Indizieren der Testkollektion verwendet. Die Bibliothek enthält eine Komponente zur Erzeugung eines Indexes, dem verschiedenartige Textdokumente hinzugefügt werden können, sowie eine Komponente zum Durchsuchen dieses Indexes. Während der Suche werden verschiedene Operatoren unterstützt. Dazu zählen binäre Operatoren wie „AND“ und „OR", aber auch die Gewichtung verschiedener Terme innerhalb einer Anfrage ist möglich. Außerdem können Begriffe durch Klammern gruppiert werden. Die Suchergebnisse werden sortiert ausgegeben.

### 4.1.2 Wikipedia-Korpus

Der Wikipedia-Korpus besteht aus Artikeln der Online-Enzyklopädie Wikipedia. Diese existiert in einer Vielzahl von Sprachversionen. Artikel in verschiedenen Sprachen, die dasselbe Thema behandeln, sind durch sogenannte Interlanguage-Links miteinander verbunden. Werden nur diejenigen Artikel ausgewählt, die in der Sprache $L$ einen Interlanguage-Link auf die Sprache $L'$ sowie in der Sprache $L'$ einen Interlanguage-Link auf die Sprache $L$ enthalten, läßt sich ein vergleichbarer bilingualer Korpus erzeugen, in dem jedem Dokument der Sprache $L$ ein Dokument der Sprache $L'$ zugeordnet ist und umgekehrt.

[3] http://lucene.apache.org/

Der komplette Inhalt jeder Wikipedia-Sprachversion wird in Form eines Dumps zum Herunterladen zur Verfügung gestellt[4]. Für den Evaluierungskorpus wurden die englische sowie die deutsche Version des Dumps verwendet (deutscher XML-Dump vom 22.04.2008, englischer XML-Dump vom 12.03.2008). Daraus erfolgt die Extraktion aller Artikel, welche ein Sprachpaar bilden und zusätzlich die folgenden Anforderungen erfüllen (Anderka (2007), S. 48):

- Ein Artikel muß mehr als 100 Wörter enthalten, die keine Stopwörter sind.
- Bei dem Artikel handelt es sich nicht um eine „Weiterleitung“ oder eine „Begriffsklärung“.
- Der Titel des Artikels ist keine Zahl und kein Datum.

Daraus ergeben sich 138’324 Dokumentpaare, die domänenunabhängig sind. Aus allen englischen Dokumenten wurde ein Schlüsselwortindex mit Hilfe von Lucene sowie ein Fingerprint-Index erstellt, anhand derer die Evaluierung stattfand. Da außerdem davon auszugehen ist, daß alle Wikipedia-Dokumente von herkömmlichen Suchmaschinen indiziert sind, fand eine zusätzliche Evaluierung mit Hilfe des Schlüsselwort-Indexes von Yahoo statt.

### 4.1.3 JRC-Acquis-Korpus

Die JRC-Acquis-Kollektion (Steinberger u. a. (2006)) enthält Dokumente des „Acquis Communautaire“ (AC, dt.: gemeinschaftlicher Besitzstand) der Europäischen Union. Der AC umfaßt den Gesamtbestand an Rechten und Pflichten, der für die Mitgliedsstaaten der EU verbindlich ist. Dazu gehören u.a. Verträge, Entscheidungen des Europäischen Gerichtshofes, Erklärungen und bestimmte Abkommen. Die Texte liegen in den 22 Sprachen der EU vor[5]. Die JRC-Acquis-Kollektion wurde von der „Language Technology Group“ des „Joint Research Centre“ der EU-Kommission zusammengestellt. Dabei spielten folgende Auswahlkriterien eine Rolle:

[4] http://download.wikimedia.org/

[5] Bulgarisch, Tschechisch, Dänisch, Deutsch, Griechisch, Englisch, Spanisch, Estnisch, Finnisch, Französisch, Ungarisch, Italienisch, Litauisch, Lettisch, Maltesisch, Holländisch, Polnisch, Portugiesisch, Rumänisch, Slovakisch, Slowenisch und Schwedisch

- Ein Dokument muß in mindestens 10 Sprachen (außer Bulgarisch und Rumänisch) vorliegen.
- Zu diesen 10 Sprachen müssen mindestens 3 Sprachen gehören, die erst seit 2004 offizielle Sprachen der EU sind (Tschechisch, Estnisch, Ungarisch, Litauisch, Lettisch, Maltesisch, Polnisch, Slovakisch und Slowenisch).

Ziel war die Erstellung eines großen, parallelen Korpus für sprachwissenschaftliche und computerlinguistische Forschungen. Die Dokumente sind einander auf Satzebene zugeordnet.

Für diese Arbeit wurde die deutsche Sammlung mit 20'666 Dokumenten sowie die englische mit 20'686 Dokumenten verwendet. Die endgültige Testkollektion enthält nur noch diejenigen Texte, die sowohl in englischer als auch in deutscher Sprache vorhanden sind. Daraus ergeben sich 20'593 Dokumentpaare. Aus den englischen Dokumente wurden wieder ein Schlüsselwortindex und ein Fingerprint-Index erstellt, anhand derer die Evaluierung stattfand.

## 4.2 Vorexperimente

Bevor die drei Verfahren miteinander verglichen werden, wird die Qualität der Übersetzungsmethoden ausgewertet. Um zu überprüfen, wie gut die Nachimplementierung der Übersetzung von Schlüsselwörtern funktioniert, wurden einige der von Ballesteros beschriebenen Experimente durchgeführt (Ballesteros (2001), Kapitel 4 und 6). Die Übersetzung vollständiger Texte kann ausgewertet werden, indem die Ähnlichkeit zwischen Original und maschineller Übersetzung über das Kosinusmaß bestimmt wird.

### 4.2.1 Qualität der Schlüsselwortübersetzung

Der Versuch gestaltet sich folgendermaßen: Es sind Anfragen in der Sprache $L$ gegeben sowie eine Kollektion mit Testdokumenten in der Sprache $L'$. Weiterhin ist bekannt, welche Dokumente zu welcher Anfrage relevant sind. Im Versuch werden nun die Anfragen aus $L$ in $L'$ übersetzt und mit diesen der Index der TREC-Testdokumente

| Kollektion | mono-lingual | Schlüsselwort-übersetzung | Übersetzung und Query-Expansion |
|---|---|---|---|
| TIPSTER (TREC-3) | 0.2259 | 0.1583 (70%) | 0.2489 (110%) |
| AP (TREC-6) | 0.3869 | 0.3057 (79%) | 0.3623 (94%) |

***Tabelle 4.2:*** *Mean-Average-Precision (MAP), welche Ballesteros für die englischen TREC-3- und TREC-6-Kollektionen und spanische Anfragen erreicht (Ballesteros (2001)). „Schlüsselwortübersetzung“ umfaßt eine Wort-für-Wort-Übersetzung der Anfrage sowie eine Kookkurrenzanalyse für mehrteilige Begriffe. In Klammern ist die Qualität der multilingualen Verfahren im Vergleich zum monolingualen Retrieval angegeben.*

angefragt. Da bekannt ist, welche Dokumente in der Ergebnismenge enthalten sein müssen, können Precision und Recall bestimmt werden. Auf diese Weise sind die Ergebnisse des multilingualen Retrievals jedoch abhängig vom verwendeten Retrieval-System und lassen sich schlecht mit Ergebnissen anderer Systeme vergleichen. Dies läßt sich umgehen, indem die relative Veränderung der Retrieval-Werte in Bezug auf einen monolingualen Referenzwert bestimmt wird. Daher sind die Anfragen zusätzlich in der Sprache $L'$ gegeben. Mit diesen wird ebenfalls der Index der Testdokumente in $L'$ angefragt und die Retrieval-Qualität bestimmt. Die prozentuale Anteil des multilingualen Retrievals am monolingualen ist ein guter Indikator für die Qualität des Übersetzungsverfahrens.

Tabelle 4.2 zeigt die von Ballesteros in ihrer Arbeit ermittelten Retrieval-Werte. Demnach erreicht das Übersetzungsverfahren mit anschließender Strukturierung der Anfrage eine Mean-Average-Precision von 70 Prozent (TREC-3) bzw. 79 Prozent (TREC-6) im Vergleich zu der des monolingualen Retrievals. Werden zusätzlich Query-Expansion-Techniken angewandt, steigen die Zahlen auf 94 Prozent (TREC-3) bzw. sogar 110 Prozent (TREC-6). Bei ihren Werten ist nicht immer klar, ob sich die Ergebnisse auf kurze oder lange Anfragen beziehen. Auch unterscheiden sich teilweise die Angaben für ein und dasselbe Verfahren. Mitunter sind keine monolingualen Vergleichswerte angegeben. Die plausibelsten Werte sind in der Tabelle aufgeführt.

In Tabelle 4.3 sind die Werte dargestellt, die das Verfahren mit der eigenen Implementierung des Ansatzes erreicht. Zu Beginn wurde der Ansatz mit Ballesteros’ Art der Strukturierung der Anfrage evaluiert (a). Die Werte sind in beiden Kollektionen signi-

| Kollektion | Anfrage-länge | | mono-lingual$^{(a)}$ | Schlüsselwort-übersetzung$^{(a)}$ | mono-lingual$^{(b)}$ | Schlüsselwort-übersetzung$^{(b)}$ |
|---|---|---|---|---|---|---|
| TIPSTER (TREC-3) | kurz | Recall | 0.1637 | 0.0725 (44.28%) | 0.3127 | 0.1029 (32.91%) |
| | | Precision | 0.1750 | 0.0034 (1.96%) | 0.0509 | 0.0186 (36.56%) |
| | | MAP | 0.2299 | 0.0358 (15.56%) | 0.0509 | 0.0168 (32.91%) |
| | lang | Recall | 0.0095 | 0 (0%) | 0.2504 | 0.0628 (25.08%) |
| | | Precision | 0.1694 | 0 (0%) | 0.1690 | 0.1770 (104.70%) |
| | | MAP | 0.0872 | 0 (0%) | 0.0423 | 0.0111 (26.26%) |
| AP (TREC-6) | kurz | Recall | 0.3697 | 0.2598 (70.28%) | 0.6632 | 0.4563 (68.80%) |
| | | Precision | 0.0676 | 0.0174 (25.70%) | 0.0468 | 0.0349 (74.52%) |
| | | MAP | 0.5008 | 0.2925 (58.40%) | 0.0869 | 0.0736 (84.69%) |
| | lang | Recall | 0.0192 | 0.0016 (8.33%) | 0.4338 | 0.1371 (31.61%) |
| | | Precision | 0.3529 | 0.4000 (113.33%) | 0.0258 | 0.0086 (33.19%) |
| | | MAP | 0.2350 | 0.0952 (40.53%) | 0.0258 | 0.0081 (31.61%) |

***Tabelle 4.3:*** *Ergebnisse der eigenen Implementierung von Ballesteros' Ansatz mit (a) deren Methode zur Strukturierung der Anfrage und (b) Shapiros Methode zur Strukturierung der Anfrage für die englischen TREC-3- und TREC-6-Kollektionen und deutsche Anfragen. In Klammern ist die Qualität des multilingualen Retrievals im Vergleich zum monolingualen angegeben. Für die Strukturierung nach Shapiro wurden folgende Parameter verwendet: mindestens ein Term pro Teilanfrage, 1000 Ergebnisse pro Teilanfrage, 1000 Ergebnisse insgesamt.*

fikant schlechter. Während Ballesteros für die TREC-3-Kollektion 70 Prozent des monolingualen Retrievals erreicht, sind es hier nur 15.56 Prozent bei kurzen Anfragen. Für lange Anfragen scheitert das Verfahren völlig. Auch für die TREC-6-Kollektion schneidet das Verfahren mit 58.4 Prozent für kurze Anfragen bzw. 40.53 Prozent für lange Anfragen schlechter ab als bei Ballesteros (79 Prozent). Absolut betrachtet, ist die Retrieval-Qualität ebenfalls nicht gut. Bereits die monolingualen Referenzwerte sind nicht zufriedenstellend: Der Recall liegt zwischen 0.0095 und 0.3697, die Precision zwischen 0.0676 und 0.3529.

Das in Abschnitt 3.3.1 vorgestellte Verfahren zur Strukturierung von Anfragen nach Shapiro ist besser geeignet (b). Die absoluten Zahlen sowohl für das monolinguale als auch für das multilinguale Retrieval sind besser. Für die TREC-6-Kollektion und kurze Anfragen schneidet das veränderte Verfahren mit 84.69 Prozent der Qualität des monolingualen Retrievals sogar etwas besser ab als das von Ballesteros'; lange Anfragen

und alle Anfragen der TREC-3-Kollektion verbessern die Mean-Average-Precision auf Werte zwischen 26.26 Prozent und 32.91 Prozent der monolingualen Ergebnisse.

Die Ursachen für das schlechtere Abschneiden im Vergleich zu Ballesteros' Ergebnissen können in der unterschiedlichen Art der Implementierung liegen. Das deutlich schlechtere Abschneiden der TREC-3-Kollektion ist möglicherweise in der deutlich höheren Komplexität der Anfragen begründet, wie Listing 4.2 und 4.3 zeigen. Daher wurden deutlich weniger Wörter im Wörterbuch gefunden. Außerdem mußten die Anfragen der TREC-3-Kollektion zunächst manuell ins Deutsche übersetzt werden. Dies ist eine mögliche Fehlerquelle, da die Anfragen eventuell nicht immer korrekt übersetzt wurden.

Die Ergebnisse von Ballesteros konnten nicht nachvollzogen werden. Es scheint, daß das Verfahren entweder nur für spanische Texte oder unter perfekten Rahmenbedingungen brauchbare Resultate liefert. Im schlimmsten Fall funktioniert es überhaupt nicht. All dies spricht gegen die Robustheit des Übersetzungsverfahrens. Aufgrund der vielen zu lösenden Teilprobleme und der Abhängigkeit von der Qualität der verfügbaren Ressourcen, ist es nur schwer zu kontrollieren. Trotzdem wird es in den weiteren Experimenten berücksichtigt, da die in dieser Arbeit zu lösende Problem einfacher ist als die Aufgabenstellung der TREC und der Ansatz in dieser Retrieval-Aufgabe möglicherweise besser Resultate erzielt werden.

```
1 <num> Number: 170
2 <title> Topic: The Consequences of Implantation of Silicone Gel Breast Devices
3 <desc> Description: Document will discuss the use of silicone gel breast implants
      manufactured by Dow-Corning Corporation and recipient's concerns regarding
      future medical problems that may occur as a result of manufacturing defects or
      lack of proper testing before approval for general public use.
```

***Listing 4.2:*** *Beispielanfrage der TREC-3-Kollektion*

```
1 <num> Number: CL14
2 <title> Topic: International Terrorism
3 <desc> Description: What efforts are being taken to combat international terrorism?
```

***Listing 4.3:*** *Beispielanfrage der TREC-6-Kollektion*

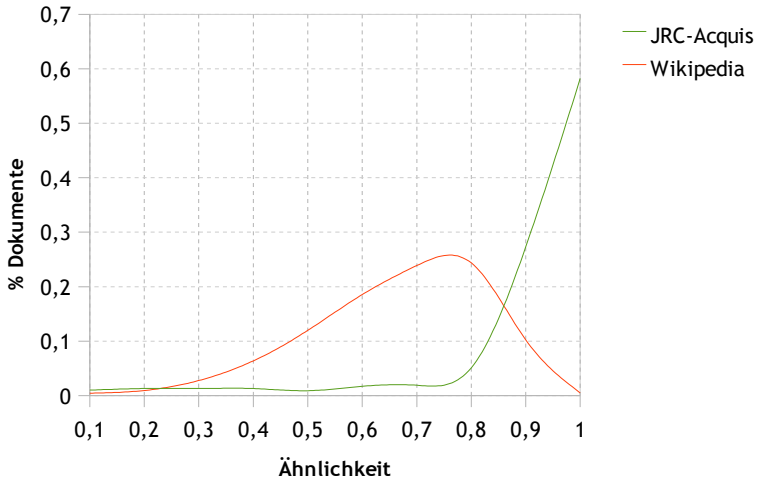


***Abbildung 4.1:*** *Ähnlichkeit zwischen maschinell übersetztem Text und Original.*

### 4.2.2 Qualität der Übersetzung von Dokumenten

Zur Bestimmung der Qualität der Übersetzung von kompletten Dokumenten genügt es, die Kosinus-Ähnlichkeit zwischen maschinell übersetztem Text und Originaldokument zu bestimmen. Je größer die Ähnlichkeit, desto besser ist die Übersetzung. Dazu werden je etwa 1000 zufällig aus der Wikipedia- und der JRC-Acquis-Kollektion ausgewählte deutsche Dokumente mit Hilfe des Web-Service von Google ins Englische übersetzt und die Kosinus-Ähnlichkeit zwischen Übersetzung und englischem Original ermittelt. Die Kurven in Abbildung 4.1 zeigen diese paarweise Ähnlichkeit.

Die Übersetzung der JRC-Acquis-Dokumente funktioniert hervorragend. 86 Prozent aller maschinell übersetzten Dokumente weisen zum Original eine Ähnlichkeit auf, die über 0.8 liegt, bei 58 Prozent ist sie größer als 0.9. Immerhin 59 Prozent aller maschinell übersetzten Wikipedia-Artikel weisen zum Original eine Ähnlichkeit auf, die über 0.6 liegt. Wird berücksichtigt, daß zwei zufällig ausgewählte Wikipedia-Dokumente in nur 4.5 Prozent der Fälle eine Ähnlichkeit aufweisen, die über 0.1 liegt (siehe Abbildung 3.1), können auch die Wikipedia-Dokumente gut übersetzt werden.

## 4.3 Experimente

### 4.3.1 Ablauf der Experimente

Die Qualität und die Laufzeit von

- Retrieval mit übersetzten Schlüsselwörtern (kurz Schlüsselwortübersetzung, SÜ),
- Retrieval mit übersetzten Dokumenten (kurz Textübersetzung, TÜ) und
- Hashing-basiertem Retrieval mit auf Fuzzy-Fingerprinting (kurz FFP)

wurden miteinander verglichen. Dazu wurden aus den beiden Korpora, Wikipedia und JRC-Acquis, 1000 deutsche Artikel zufällig ausgewählt und es wurde überprüft, ob das zugehörige englische Dokument gefunden werden konnte. Die drei Schlüsselwortindizes liefern sortierte Ergebnislisten zurück. Daher konnte hier auch die Position des gefundenen Dokuments ausgewertet werden.

Der Ablauf der Experimente gestaltete sich wie folgt:

- Schlüsselwortübersetzung (SÜ)
  1. Auswahl von jeweils 1000 zufälligen deutschen Dokumenten aus der Wikipedia-Kollektion und der JRC-Acquis-Kollektion
  2. Extraktion der Schlüsselwörter
  3. Übersetzung der Schlüsselwörter
  4. Anfrage des jeweiligen Schlüsselwortindexes
  5. Bestimmung von Rang, Recall, Precision, MAP und F-Measure
- Textübersetzung (TÜ)
  1. Auswahl von jeweils 1000 zufälligen deutschen Dokumenten aus der Wikipedia-Kollektion und der JRC-Acquis-Kollektion
  2. Übersetzung der Dokumente
  3. Extraktion der Schlüsselwörter
  4. Anfrage des jeweiligen Schlüsselwortindexes

5. Bestimmung von Rang, Recall, Precision, MAP und F-Measure

- Fuzzy-Fingerprinting (FFP)

  1. Auswahl von jeweils 1000 zufälligen deutschen Dokumenten aus der Wikipedia-Kollektion und der JRC-Acquis-Kollektion
  2. Übersetzung der Dokumente
  3. Anfrage des jeweiligen Fingerprint-Indexes
  4. Bestimmung von Recall, Precision und F-Measure

In den verschiedenen Durchläufen fand eine Variation der zu untersuchenden Parameter statt, die im nächsten Abschnitt näher erörtert werden.

Für SÜ und TÜ sind außerdem monolinguale Vergleichswerte notwendig, um die einzelnen Ansätze besser miteinander vergleichen zu können. Diese wurden durch die Auswahl von 1000 englischen Dokumenten, die Extraktion der Schlüsselwörter daraus und anschließender Anfrage der Schlüsselwortindizes gewonnen.

### 4.3.2 Parameterbestimmung

Alle drei Verfahren besitzen eine Reihe von Parametern zur Beeinflussung der Retrieval-Qualität. Für diese wurden vor Beginn der Experimente geeignete Werte ermittelt, um die Verfahren möglichst optimal einzustellen.

***Strukturierung der Anfrage***: Die Verfahren SÜ und TÜ, welche auf einem schlüsselwortbasierten Retrieval beruhen, beinhalten jeweils den Schritt zur Strukturierung von Anfragen, während dessen aus den extrahierten (und übersetzten) Schlüsselbegriffen eines Dokuments verschiedene Teilanfragen an den Suchindex erzeugt werden. Hier muß eingestellt werden, wieviele Terme eine Anfrage minimal enthält (minTermsPerQuery). Daraus ergibt sich die Anzahl der Teilanfragen. Weiterhin muß die maximale Anzahl der Ergebnisse pro Teilanfrage (documentsPerQuery) sowie die maximale Anzahl der Ergebnisse nach Zusammenführung der Resultate der Teilanfragen (maxResults) festgelegt werden. Die Retrieval-Qualität ist außerdem davon abhängig, wieviele Schlüsselwörter aus dem Dokument entnommen wurden und welcher Art

| Anzahl der extrahierten Terme | | 2 | 5 | 10 |
|---|---|---|---|---|
| documentsPerQuery | | 100 | 100 | 100 |
| maxResults | | 100 | 100 | 100 |
| minTermsPerQuery | | 1 | 1 | 1 |
| 1-Term | Rang | 4.99 | 1.82 | 1.08 |
| | Recall | 0.8904 | 0.9985 | 1 |
| | Precision | 0.0089 | 0.0100 | 0.0100 |
| | MAP | 0.5339 | 0.8421 | 0.9700 |
| n-Term | Rang | 4.14 | 1.58 | 1.06 |
| | Recall | 0.9762 | 1 | 1 |
| | Precision | 0.0098 | 0.0100 | 0.0100 |
| | MAP | 0.6210 | 0.8553 | 0.9804 |

***Tabelle 4.4:*** *Retrieval-Qualität des monolingualen Retrievals für die Wikipedia-Kollektion abhängig von den Parametern zur Strukturierung der Anfrage sowie der Anzahl und Art der Schlüsselbegriffe.*

diese Begriffe (einzelne Wörter oder Wortgruppen) sind. Anhand eines monolingualen Referenzwertes werden die einzelnen Varianten miteinander verglichen, indem die Abweichung der Qualität des bilingualen Retrievals vom monolingualen ermittelt wird. Um hier aussagekräftige Zahlen zu erhalten, müssen die monolingualen Werte hinreichend gut sein.

Tabelle 4.4 zeigt die Werte des englischen Retrievals abhängig von der Anzahl der Schlüsselbegriffe und deren Art (1-Terme: einzelne Wörter, n-Terme: Wortgruppen) für die zu Beginn gewählten Parameter minTermsPerQuery $= 1$, documentsPerQuery $= 100$ und maxResults $= 100$. Es ist zu erkennen, daß diese für einen Vergleich ausreichend gut sind und daher nicht weiter optimiert werden müssen. Der Recall liegt zwischen 0.8904 für zwei einzelne Wörter und 1 für fünf bzw. zehn Begriffe. Das gesuchte Dokument befindet sich durchschnittlich auf Rang 1 bis 5. Die Precision wird stark über den Parameter maxResults gesteuert und spielt daher keine große Rolle.

***Fuzzy-Fingerprinting***: Die Qualität des Fuzzy-Fingerprintings hängt u.a. von der Anzahl der Präfixklassen, der Anzahl der Fuzzifizierungsschemata sowie der Anzahl der linguistischen Variablen pro Schema ab. Die Anzahl der Präfixklassen wird über einen Parameter „dimension“ eingestellt. Wird beispielsweise dimension = 50 festgelegt, werden alle $26^2$ Präfixklassen mit Präfixen der Länge von zwei Buchstaben gebildet

und zu 50 Klassen rekombiniert. Mit Hilfe des Parameters „thresholds“ werden die linguistischen Variablen definiert. Daraus ergibt sich eine Vielzahl an Kombinationsmöglichkeiten der einzelnen Parameter, die jedoch erst im eigentlichen Experiment bestimmt werden.

### 4.3.3 Textübersetzung vs. Schlüsselwortübersetzung

Das erste Experiment widmete sich dem Vergleich der beiden schlüsselwortbasierten Verfahren. Insbesondere wurde der Einfluß der Art und Anzahl der Schlüsselwörter auf die Qualität der Ergebnisse untersucht.

Die Abbildungen 4.2 und 4.3 zeigen die Kurven für Recall und Precision von Schlüsselwortübersetzung und Textübersetzung für die JRC-Acquis-Kollektion, die Abbildungen 4.4 und 4.5 die Kurven für den Wikipedia-Korpus. Untersucht wurden beide Verfahren mit jeweils zwei, fünf und zehn extrahierten Schlüsselbegriffen. Als Referenz dienen die Kurven für das monolinguale Retrieval.

Für die Wikipedia-Kollektion erreicht die Textübersetzung immer bessere Werte als die Schlüsselwortübersetzung. Bei Anfragen mit zwei Termen erreichen die Anfragen einen besseren Recall, wenn die Schlüsselwörter Wortgruppen sind. Möglicherweise liegt das daran, daß die Anfragen dann länger sind, wodurch das gesuchte Dokument besser beschrieben wird. Anfragen mit fünf oder zehn Termen zeigen kaum unterschiedliche Werte für Wortgruppen oder einzelne Wörter. Werden aus den Ergebnismengen jeweils nur die ersten zehn Dokumente betrachtet, so erreicht die Textübersetzung bei fünf Schlüsselwörtern einen Recall von 70 Prozent, die Schlüsselwortübersetzung knapp 55 Prozent. Der Recall der monolingualen Suche liegt bei knapp 100 Prozent.

Die Ergebnisse für die JRC-Acquis-Kollektion sind insgesamt schlechter als die der Wikipedia-Kollektion. Das monolinguale Retrieval erreicht lediglich für Anfragen mit zehn Schlüsselwörtern knapp 100 Prozent Recall. Auch zeigen sich hier größere Unterschiede zwischen Anfragen mit Wortgruppen und Anfragen, die aus einzelnen Wörtern bestehen. Bei zwei Schlüsselwörtern pro Anfrage schneidet sogar die Textübersetzung besser ab als das monolinguale Retrieval mit 1-Termen, wenn Wortgruppen

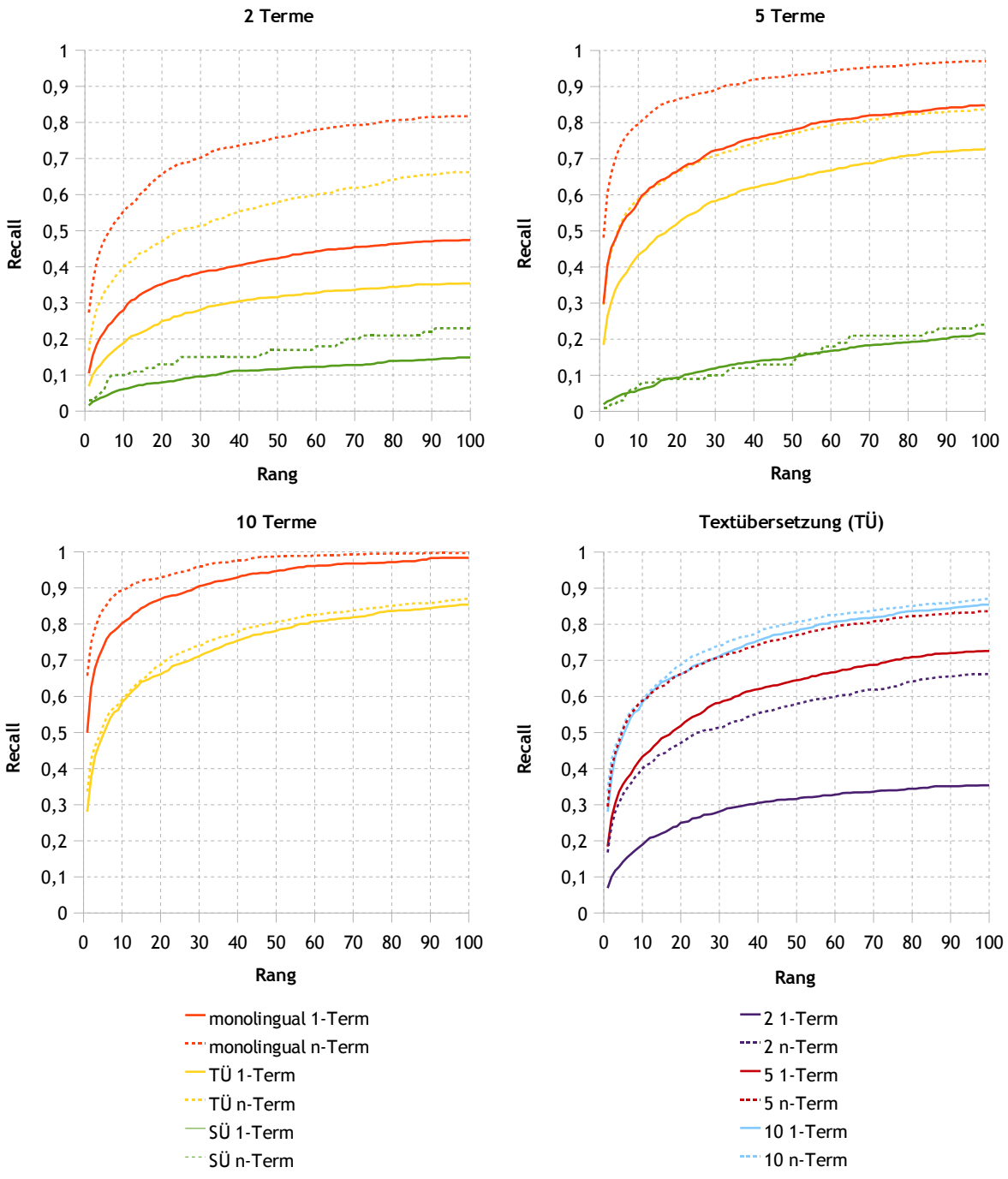


***Abbildung 4.2:*** *Die Diagramme zeigen den Recall von Schlüsselwortübersetzung (SÜ) und Textübersetzung (TÜ) abhängig vom Rang des gesuchten Dokuments für die JRC-Acquis-Testkollektion. Dargestellt sind die Kurven für zwei, fünf und zehn Begriffe.*

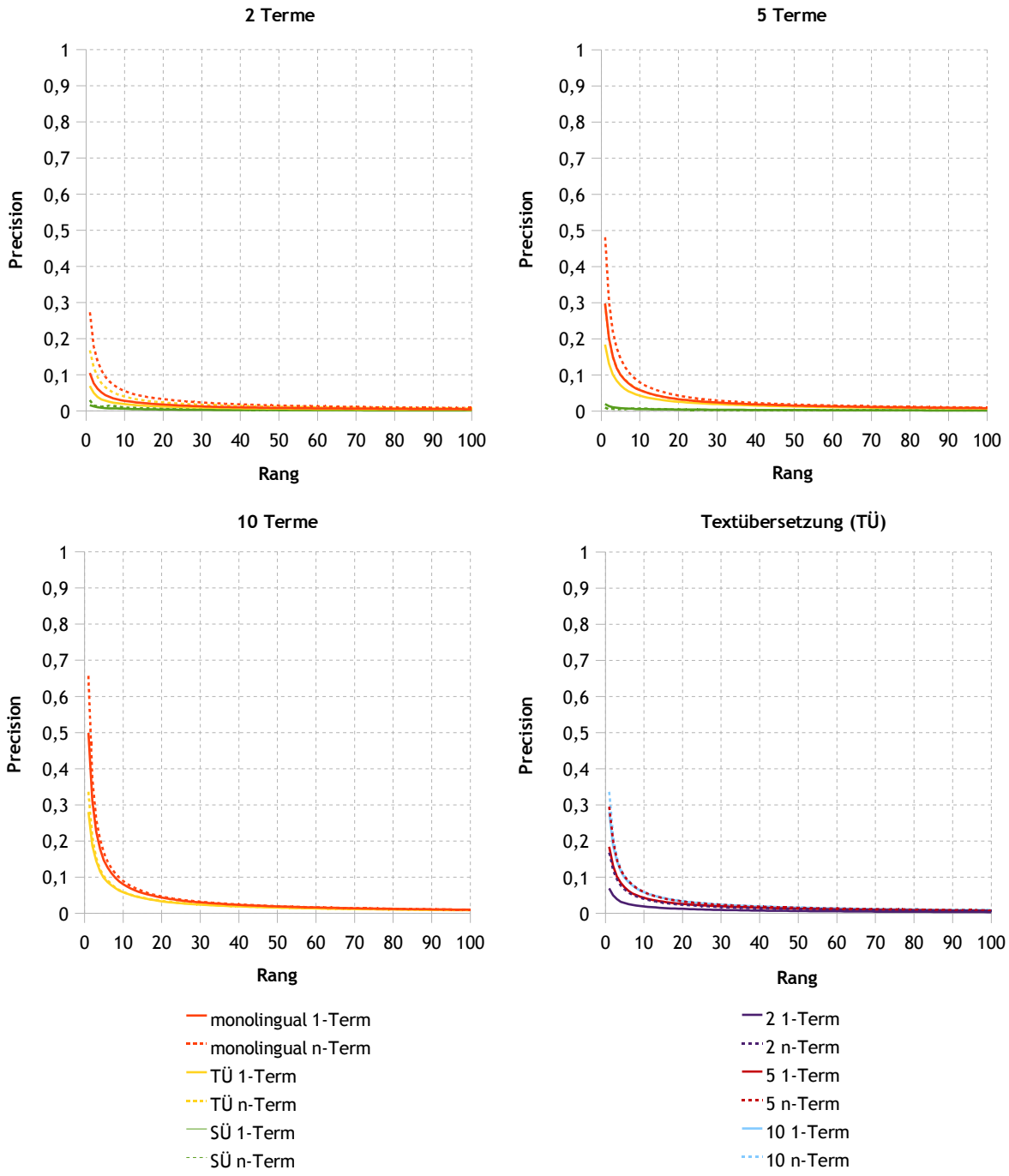


***Abbildung 4.3:*** *Die Diagramme zeigen die Precision von Schlüsselwortübersetzung (SÜ) und Textübersetzung (TÜ) abhängig vom Rang des gesuchten Dokuments für die JRC-Acquis-Testkollektion. Dargestellt sind die Kurven für zwei, fünf und zehn Begriffe.*

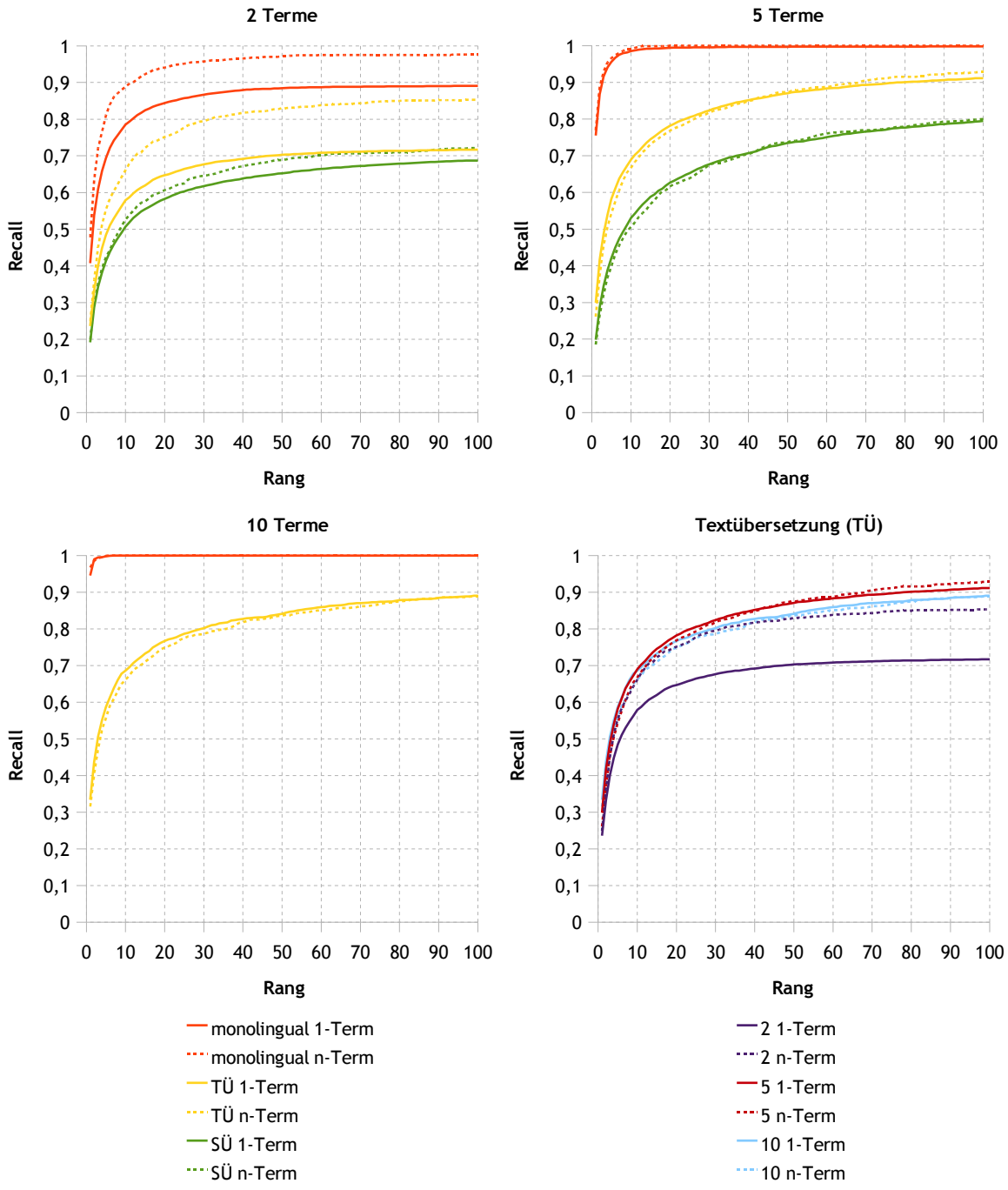


***Abbildung 4.4:*** *Die Diagramme zeigen den Recall von Schlüsselwortübersetzung (SÜ) und Textübersetzung (TÜ) abhängig vom Rang des gesuchten Dokuments für die Wikipedia-Testkollektion. Dargestellt sind die Kurven für zwei, fünf und zehn Begriffe.*

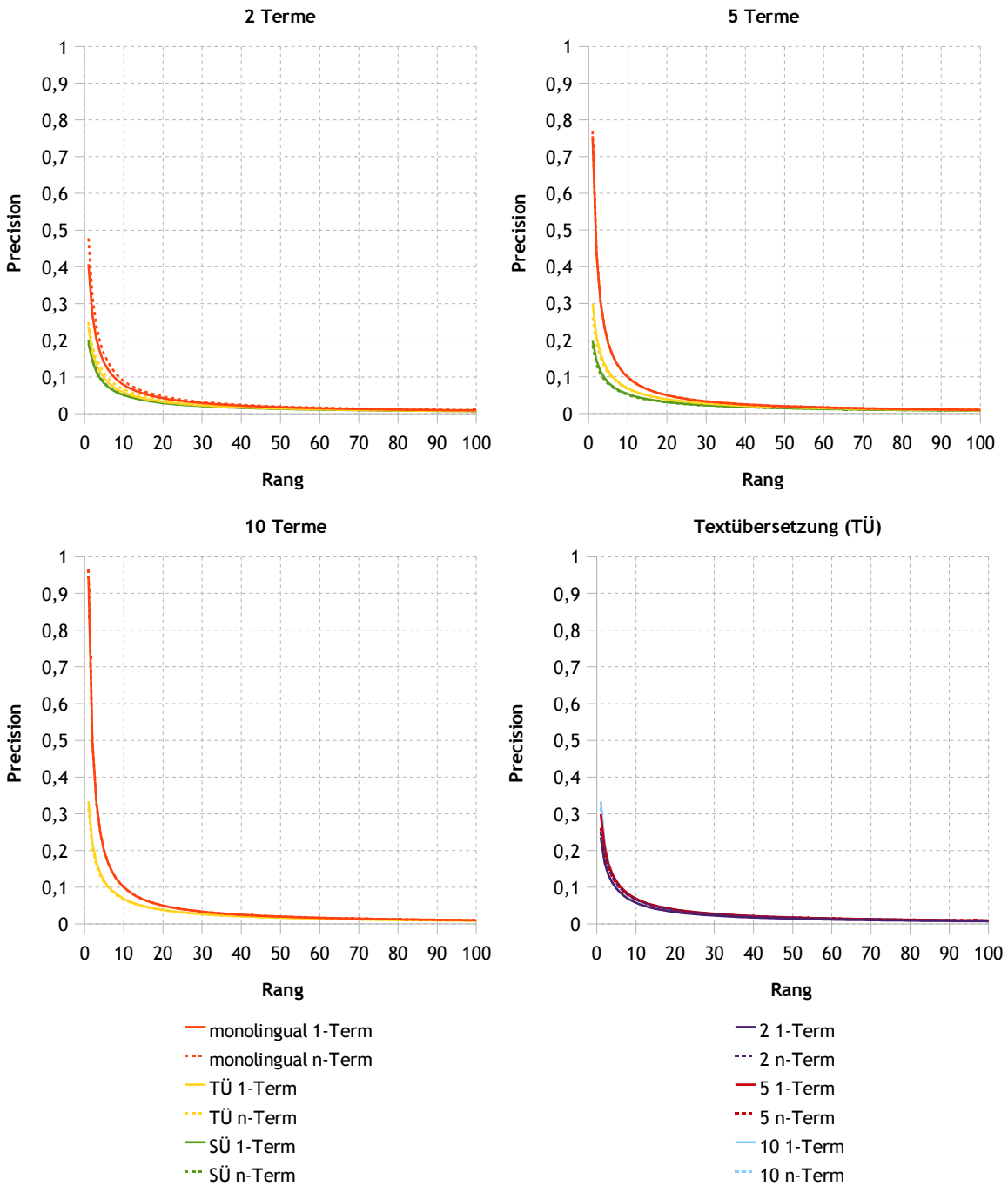


***Abbildung 4.5:*** *Die Diagramme zeigen die Precision von Schlüsselwortübersetzung (SÜ) und Textübersetzung (TÜ) abhängig vom Rang des gesuchten Dokuments für die Wikipedia-Testkollektion. Dargestellt sind die Kurven für zwei, fünf und zehn Begriffe.*

verwendet werden. Das zeigt, daß zwei Begriffe zu wenig sind, um die gewünschten Dokumente zu finden. Unabhängig von Art und Anzahl der Schlüsselwörter schneidet wieder die Textübersetzung besser ab als die Schlüsselwortübersetzung.

Die Schlüsselwortübersetzung stößt bei zehn Begriffen an ihre Grenzen. Die Auflösung der Ambiguität bzw. der Umgang mit vielen Übersetzungsvarianten gestaltet sich schwierig. Problematisch ist einerseits die Übersetzung von Wortgruppen. Hier liegen die Schwierigkeiten in der Kookkurrenzanalyse. Wie im Kapitel 3.3.1 beschrieben, werden zunächst die einzelnen Wörter der Wortgruppe übersetzt. Anschließend wird mit Hilfe einer Kookkurrenzanalyse ermittelt, welche Kombination der Übersetzungsvarianten die häufigste ist. Die Laufzeit dieser Analyse steigt exponentiell mit der Anzahl der möglichen Übersetzungen. Zusätzlich treten gehäuft Speicherprobleme auf, deren Ursache der zur Kookkurrenzanalyse verwendete Index ist, der sich während der Durchführung der Experimente noch in der Entwicklung befand. Daher ist es mitunter unmöglich die einzelnen Teilanfragen aufzubauen. Andererseits treten auch während der Strukturierung der Anfragen Probleme auf. Je mehr Übersetzungsvarianten existieren, umso mehr Teilanfragen müssen konstruiert und deren Ergebnisse ausgewertet werden. Wenn jeder der zehn Begriffe nur eine Übersetzung hätte, dann müßten daraus 1023 Anfragen gebildet werden. Angenommen die zehn extrahierten Schlüsselwörter besitzen je zwei Übersetzungen, dann ergäben sich nach der Formel in Kapitel 3.3.1 schon 7944 Teilanfragen. Häufig haben Wörter mehr als zwei Übersetzungen, wodurch die Zahl der Teilanfragen weiter stark steigt. Während der Evaluierung traten Beispiele mit mehr als 1 Million Anfragen auf. Diese Menge ist nur noch schwer zu bewältigen, die Laufzeit steigt enorm an. Aufgrund dieser Schwierigkeiten konnten für die Schlüsselwortübersetzung von zehn Termen keine aussagekräftigen Retrieval-Werte ermittelt werden.

Tabelle 4.5 zeigt die Ergebnisse noch einmal im Überblick. Aufgeführt sind der durchschnittliche Rang, auf dem das gesuchte Dokument gefunden wird, der Recall sowie die Mean-Average-Precision. Zusätzlich ist in Klammern angegeben, wieviel Prozent der Qualität des monolingualen Retrievals erreicht werden konnte.

| Terme | | | Textübersetzung (TÜ) Wikipedia | Textübersetzung (TÜ) JRC-Acquis | Schlüsselwortübersetzung (SÜ) Wikipedia | Schlüsselwortübersetzung (SÜ) JRC-Acquis |
|---|---|---|---|---|---|---|
| 2 | 1-Term | Rang | 7.64 (+2.65) | 17.97 (+1.55) | 10.69 (+5.71) | 27.83 (+11.40) |
| | | Recall | 0.7170 (80.52%) | 0.3540 (74.68%) | 0.6872 (77.18%) | 0.1490 (31.43%) |
| | | MAP | 0.3457 (64.75%) | 0.1088 (66.35%) | 0.2947 (55.21%) | 0.0306 (18.65%) |
| | n-Term | Rang | 8.60 (+4.46) | 17.85 (+5.10) | 10.76 (+6.63) | 29.35 (+16.60) |
| | | Recall | 0.8530 (87.39%) | 0.6620 (81.03%) | 0.7210 (73.86%) | 0.2300 (28.15%) |
| | | MAP | 0.3829 (61.65%) | 0.2473 (67.05%) | 0.3064 (49.34%) | 0.0483 (13.10%) |
| 5 | 1-Term | Rang | 10.04 (+8.22) | 16.98 (+3.74) | 13.94 (+12.12) | 34.20 (+20.96) |
| | | Recall | 0.9115 (91.29%) | 0.7260 (85.61%) | 0.7949 (87.07%) | 0.2150 (25.11%) |
| | | MAP | 0.4274 (50.76%) | 0.2703 (39.70%) | 0.3050 (60.09%) | 0.0350 (5.18%) |
| | n-Term | Rang | 11.69 (+10.11) | 12.98 (+5.08) | 14.81 (+13.23) | 39.00 (+31.10) |
| | | Recall | 0.9285 (92.85%) | 0.8360 (86.19%) | 0.8004 (86.20%) | 0.2400 (27.85%) |
| | | MAP | 0.3928 (45.92%) | 0.3970 (66.89%) | 0.2891 (62.96%) | 0.0249 (3.73%) |
| 10 | 1-Term | Rang | 9.99 (+8.91) | 13.85 (+5.90) | – | – |
| | | Recall | 0.8903 (89.07%) | 0.8540 (86.79%) | – | – |
| | | MAP | 0.4492 (46.31%) | 0.3829 (62.82%) | – | – |
| | n-Term | Rang | 11.02 (+9.96) | 13.09 (+8.45) | – | – |
| | | Recall | 0.8888 (88.88%) | 0.8700 (87.62%) | – | – |
| | | MAP | 0.4268 (43.53%) | 0.4265 (57.66%) | – | – |

***Tabelle 4.5:*** *Gegenüberstellung der Retrieval-Werte von Schlüsselwortübersetzung (SÜ) und Textübersetzung (TÜ) für beide Testkollektionen für zwei, fünf und zehn Terme (1-Term: einzelnes Wort; n-Term: Wortgruppe).*

***Webexperiment***: Mit jeweils fünf extrahierten Termen pro Dokument wurde das zuvor beschriebene Experiment auch auf dem Web-Schlüsselwort-Index von Yahoo durchgeführt. Die Testmenge beschränkt sich hier auf jeweils 100 Dokumente. Abbildung 4.6 zeigt die Kurven für Recall und Precision. Als Referenz dient erneut das monolinguale Retrieval. Wieder schneidet die Textübersetzung besser ab als die Schlüsselwortübersetzung. Jedoch liefert die Verwendung von einzelnen Wörtern statt Wortgruppen im Gegensatz zu den vorherigen Experimenten bessere Ergebnisse. In Tabelle 4.6 sind die Werte des Web-Indexes denen des Lucene-Indexes gegenübergestellt. Insgesamt sind die Zahlen für den Yahoo-Index wie erwartet schlechter. Der Lucene-Index enthält nur Wikipedia-Dokumente, das heißt keine zwei Dokumente behandeln das gleiche Thema. Im Gegensatz dazu enthält der Web-Index weitaus mehr Dokumente, vor allem mehr Dokumente zum gleichen Thema. Daher steigt die Wahr-

| | | Textübersetzung (TÜ) | | Schlüsselwortübersetzung (SÜ) | |
|---|---|---|---|---|---|
| | | Web-Index | Lucene-Index | Web-Index | Lucene-Index |
| 1-Term | Rang | 20.39 (+6.85) | 10.04 (+8.22) | 24.94 (+11.40) | 13.94 (+12.12) |
| | Recall | 0.76 (93.83%) | 0.9115 (91.29%) | 0.53 (65.43%) | 0.7949 (87.07%) |
| | MAP | 0.1819 (66.25%) | 0.4274 (50.76%) | 0.0842 (30.66%) | 0.3050 (60.09%) |
| n-Term | Rang | 19.2 (+2.49) | 11.69 (+10.11) | 25.92 (+9.21) | 14.81 (+13.22) |
| | Recall | 0.65 (82.28%) | 0.9285 (92.85%) | 0.4848 (61.37%) | 0.8004 (86.20%) |
| | MAP | 0.1654 (61.80%) | 0.3928 (45.92%) | 0.0542 (20.27%) | 0.2891 (62.96%) |

***Tabelle 4.6:*** *Vergleich der Retrieval-Werte der Wikipedia-Testkollektion zwischen Web-Index und Lucene-Index.*

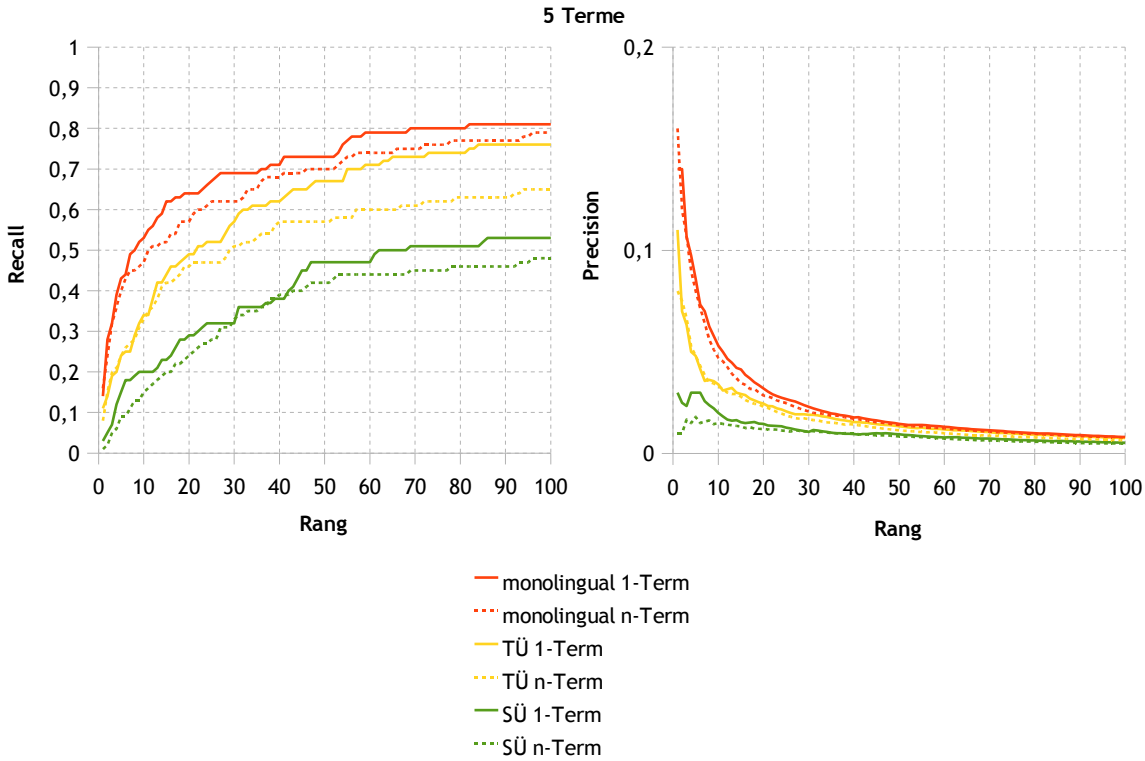


***Abbildung 4.6:*** *Die Diagramme zeigen den Recall und die Precision abhängig vom Rang des gesuchten Dokuments von Schlüsselwortübersetzung und Textübersetzung für die Wikipedia-Testkollektion getestet mit dem Web-Index von Yahoo. Dargestellt sind die Kurven für fünf Begriffe.*

| | | 1.25 (65) | 1.55 (65) | 1.85 (65) | 1.25 (60), 1.55 (65) | 1.25 (65), 1.55 (60) | 1.25 (65), 1.55 (65) | 1.25 (65), 1.85 (65) |
|---|---|---|---|---|---|---|---|---|
| Wikipedia | Recall | 0.0373 | 0.0820 | 0.1281 | 0.1070 | 0.1167 | 0.1028 | 0.1614 |
| | Precision | 0.0043 | 0.0013 | 0.0004 | 0.0014 | 0.0010 | 0.0015 | 0.0005 |
| | F-Measure | 0.0077 | 0.0025 | 0.0008 | 0.0027 | 0.0019 | 0.0030 | 0.0010 |
| JRC-Acquis | Recall | 0.0690 | 0.1130 | 0.1880 | 0.1830 | 0.2200 | 0.1710 | 0.2400 |
| | Precision | 0.1133 | 0.0386 | 0.0129 | 0.0430 | 0.0237 | 0.0523 | 0.0160 |
| | F-Measure | 0.0858 | 0.0576 | 0.0242 | 0.0696 | 0.0428 | 0.0801 | 0.0300 |

***Tabelle 4.7:*** *Retrieval-Werte des Fuzzy-Fingerprinting unter Verwendung unterschiedlicher Parameter.*

scheinlichkeit, daß auch andere Dokumente zu einer Anfrage gefunden werden. In der Auswertung wurde anhand der URL überprüft, ob die gesuchten Wikipedia-Artikel in der Ergebnismenge auftauchen. Da Kopien dieser Artikel jedoch auch unter anderen Adressen zu finden sind, ist es möglich, daß die gesuchten Artikel in den Ergebnissen auftauchen, jedoch nicht beachtet wurden.

### 4.3.4 Retrieval-Qualität des Fuzzy-Fingerprintings

Die Wahl der richtigen Parameter für einen möglichst guten Recall gestaltet sich schwierig. In Tabelle 4.7 sind die Werte einiger geprüfter Varianten zusammengestellt. Getestet wurden verschiedene Werte für den Parameter „threshold“ (jeweils ein einzelner Wert zwischen 1.25 und 1.85) in Kombination mit verschiedenen Werten für den Parameter „dimension“ (60 bis 65) sowie eine Verknüpfung mehrerer Funktionen.

Der Recall ist für alle getesteten Parameterkombinationen gering. Er reicht von 0.04 bis 0.16 für die Wikipedia-Kollektion und 0.07 bis immerhin 0.24 für die JRC-Acquis-Kollektion. Die Precision ist ebenfalls niedrig. Insgesamt betrachtet schneiden die verknüpften Hashfunktionen besser ab, als die unverknüpften. Die Resultate der JRC-Acquis-Kollektion übertreffen in allen Fällen die der Wikipedia-Kollektion. Grund dafür ist die höhere Ähnlichkeit zwischen maschinell übersetztem Text und Originaldokument. Dadurch ist die Wahrscheinlichkeit höher, daß für Übersetzung und Original

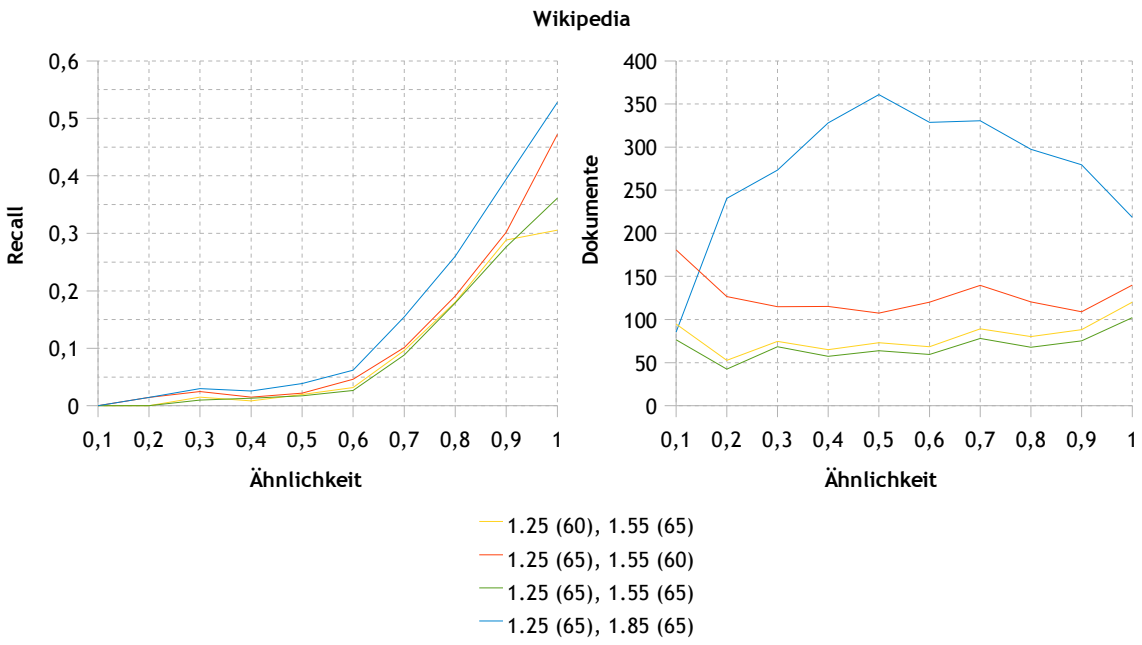


***Abbildung 4.7:*** *Das linke Diagramm zeigt den Recall für das Fuzzy-Fingerprinting für verschiedene Parameter mit der Wikipedia-Testkollektion. Im rechten Diagramm ist aufgeführt, wieviele Dokumente die Ergebnismenge durchschnittlich umfaßt. Beide Maße sind in Abhängigkeit von der Ähnlichkeit zwischen Übersetzung und Original aufgetragen.*

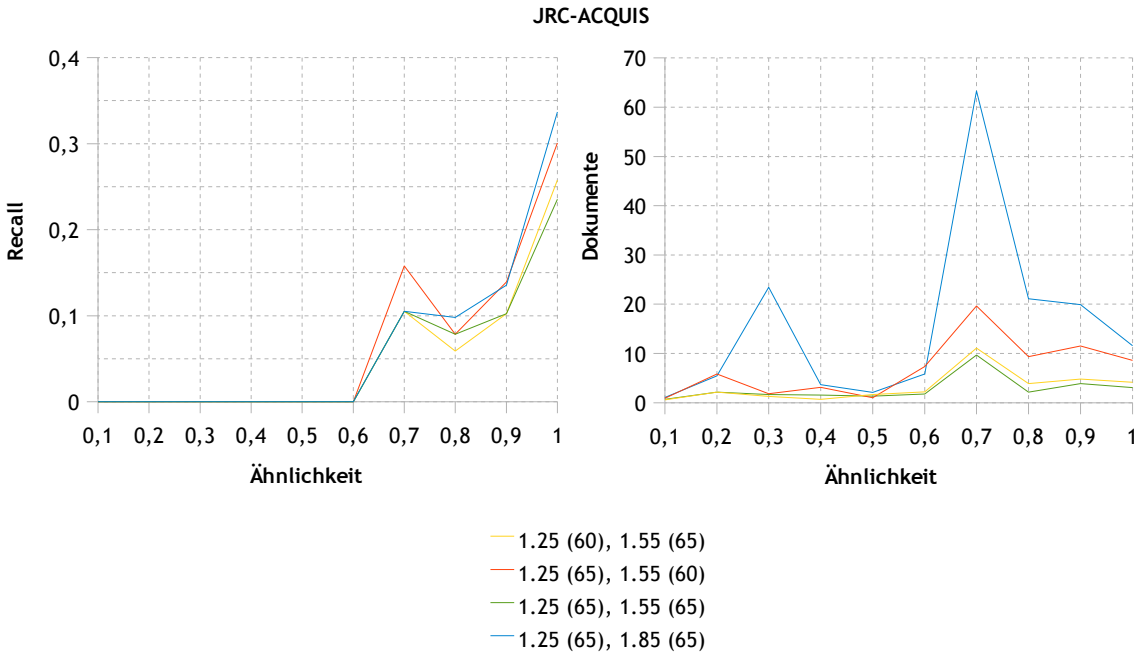


***Abbildung 4.8:*** *Die Diagramme zeigen den Recall und die Größe der Ergebnismenge des Fuzzy-Fingerprintings mit der JRC-Acquis-Testkollektion analog zu den Diagrammen der Wikipedia-Kollektion.*


***Abbildung 4.9:*** *Durchschnittliche Ähnlichkeit zwischen übersetztem Text und englischem Original abhängig von der Textlänge.*

der gleiche Fingerprint berechnet wird, somit beide Dokumente als sehr ähnlich erkannt werden und der Recall steigt.

Die Abbildungen 4.7 und 4.8 zeigen die Ergebnisse der kombinierten Hashfunktionen im Detail. In den linken Grafiken ist jeweils die Abhängigkeit des Recalls von der Ähnlichkeit zwischen Übersetzung und Original erkennbar. Die Kurven der rechten Diagramme zeigen jeweils die durchschnittliche Größe der Ergebnismenge abhängig von der genannten Ähnlichkeit. Darin ist zu erkennen, daß die Precision der Kombination „1.25 (65) und 1.85 (65)“ unverhältnismäßig stark im Vergleich zum Recall sinkt, während sich die Kurven der drei anderen Varianten ähneln.

Die Ergebnisse erwecken den Anschein, daß Fuzzy-Fingerprinting für maschinell übersetzte Texte wenig geeignet ist. Möglicherweise existiert eine Kombination von Parametern, die noch bessere Resultate liefert. Ähnlich gute Ergebnisse wie beim Verfahren „Textübersetzung“ scheinen aber unwahrscheinlich.

### 4.3.5 Einfluß der Textlänge

Schließlich wurde für alle drei Verfahren überprüft, inwiefern ihre Retrieval-Eigenschaften abhängig von der Länge des verdächtigen Textes sind.

Zunächst wurde ermittelt, ob und wie sich die Qualität der Übersetzung eines kompletten Dokuments in Abhängigkeit von der Textlänge verändert. In Abbildung 4.9 ist

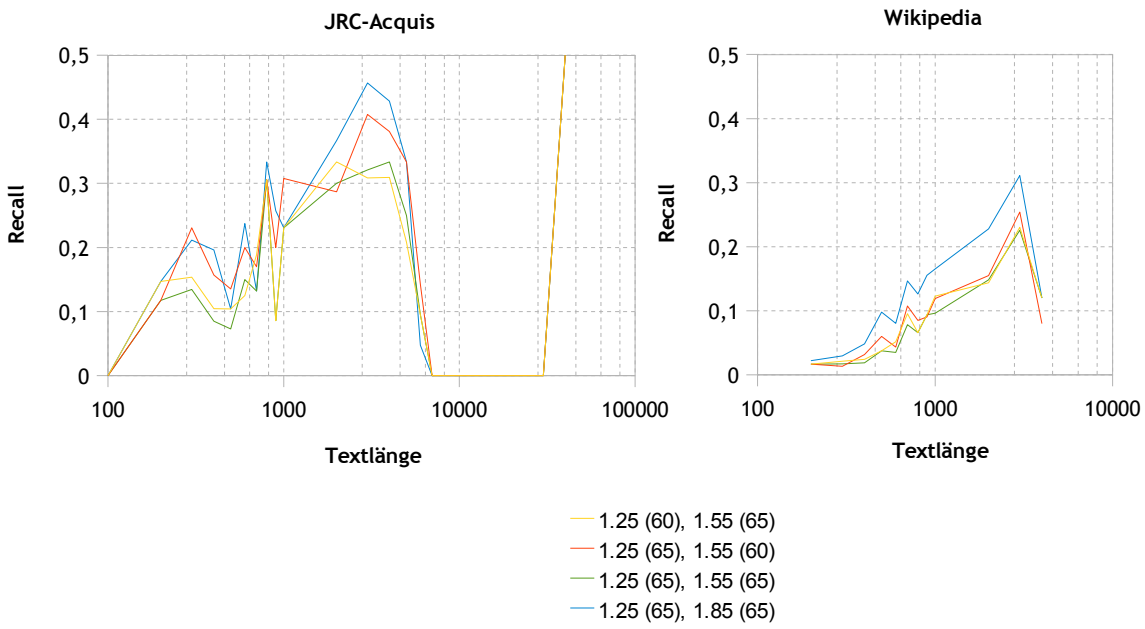


***Abbildung 4.10:*** *Recall des Fuzzy-Fingerprintings abhängig von der Textlänge.*

zu erkennen, daß die Ähnlichkeit zwischen maschineller Übersetzung und Original für die JRC-Acquis-Dokumente bei einer Textlänge zwischen 200 und 3000 Wörtern konstant bei etwa 0.9 liegt, während die Ähnlichkeit für die Wikipedia-Artikel in diesem Bereich von etwa 0.55 kontinuierlich auf 0.7 ansteigt. Für alle anderen Textlängen liegen für beide Kollektionen zu wenige Daten vor, um verläßliche Aussagen treffen zu können.

Im nächsten Schritt wurde die Abhängigkeit des Fuzzy-Fingerprintings von der Dokumentlänge untersucht. Abbildung 4.10 zeigt die Ergebnisse. Es ist ein deutlicher Anstieg des Recalls sowohl für die Wikipedia- als auch die JRC-Acquis-Kollektion zu erkennen, je mehr Wörter die Texte enthalten. Das Maximum liegt jeweils bei etwa 2000 Wörtern. Da die Ähnlichkeit zwischen Übersetzung und Original für die JRC-Acquis-Dokumente unabhängig von der Textlänge ist, kann geschlußfolgert werden, daß das Fuzzy-Fingerprinting am besten für Dokumente mit etwa 2000 Wörtern funktioniert.

Anders verhält es sich bei den beiden Verfahren „Schlüsselwortübersetzung“ und „Textübersetzung“. Hier bestand die Annahme, daß fünf Schlüsselwörter ein Dokument besser beschreiben, welches 200 Wörter enthält, als eines mit 3000 Wörtern,

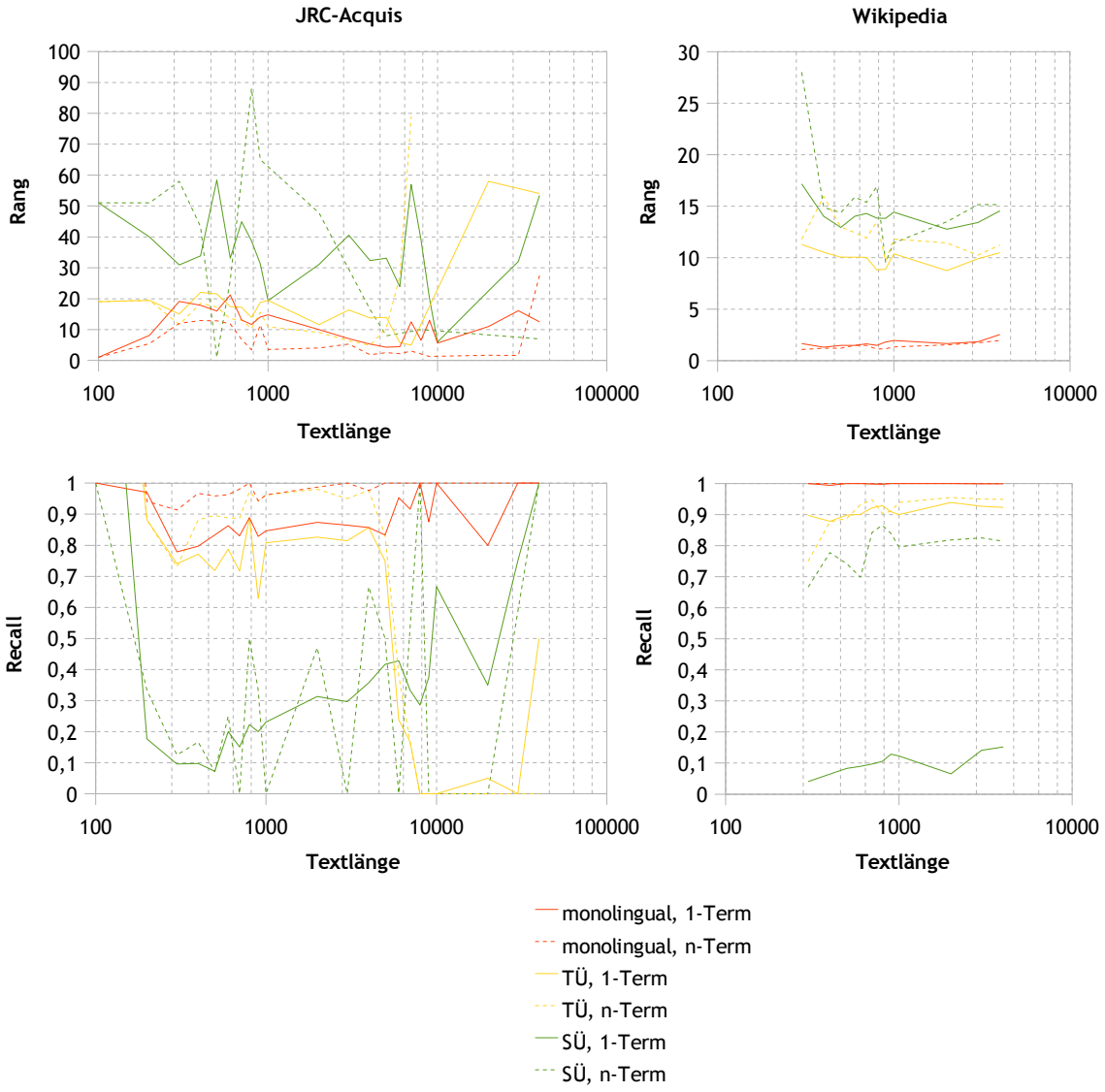


***Abbildung 4.11:*** *Recall und Rang für Schlüsselwortübersetzung (SÜ) und Textübersetzung (TÜ) abhängig von der Textlänge.*

und daß fünf Wortgruppen besser geeignet sind als fünf einzelne Wörter. Jedoch scheint dies nicht der Fall zu sein, wie aus den Kurven des monolingualen Retrievals der Wikipedia-Kollektion in Abbildung 4.11 ersichtlich ist. Die Kurven für Recall liegen konstant bei 1 und auch der Rang ändert sich nicht. Die Kurven der JRC-Acquis-Kollektion schwanken stärker, allerdings ohne erkennbaren Trend. Die Schwankungen sind durch die geringe Anzahl an Dokumenten pro Längen-Intervall zu erklären. Hier schneidet das Retrieval mit Wortgruppen besser ab, als das mit einzelnen Begriffen. Weder für Schlüsselwortübersetzung noch für Textübersetzung ist eine klare Abhängigkeit von der Textlänge erkennbar. Trotz allem ist davon auszugehen, daß die Anzahl der Schlüsselbegriffe anzupassen ist, je länger die Dokumente sind.

### 4.3.6 Laufzeitverhalten der drei Verfahren

Letztes Bewertungskriterium der Verfahren war deren Laufzeit, die in Abbildung 4.12 dargestellt ist.

Fuzzy-Fingerprinting ist eindeutig das schnellste Verfahren. Nachdem ein Fingerprint-Index der Referenzkollektion erstellt wurde, ist die Suche in konstanter Zeit durchführbar. Einzig die Übersetzung der Dokumente nimmt etwas Zeit in Anspruch.

Im Gegensatz dazu steigt die Laufzeit der Textübersetzung je mehr Schlüsselwörter zur Suche verwendet werden. Je mehr Schlüsselwörter vorhanden sind, umso mehr Teilanfragen müssen während der Strukturierung der Anfrage erzeugt und ausgewertet werden. Außerdem zeigt das Diagramm die Laufzeitunterschiede zwischen 1-Termen und n-Termen. Daraus ist zu entnehmen, daß auch die Auswertung einer einzelnen Anfrage länger dauert, je mehr Schlüsselwörter sie enthält. Ist die Anzahl der Terme gleich, werden für n-Terme ebensoviele Teilanfragen konstruiert wie für 1-Terme. Die Zeitdifferenz kann also nicht aus der unterschiedlichen Anzahl an Teilanfragen resultieren, sondern ist nur durch die größere Anzahl an einzelnen Wörtern (entstehend durch die Wortguppen) zu erklären.

Problematisch ist die Schlüsselwortübersetzung. Verläßlich sind hier nur die Zahlen für einzelne Wörter als Schlüsselbegriffe. Die Zeit, die zur Extraktion und Übersetzung der Schlüsselwörter benötigt wird, ist sehr gering. Jedoch benötigt die Struktu-

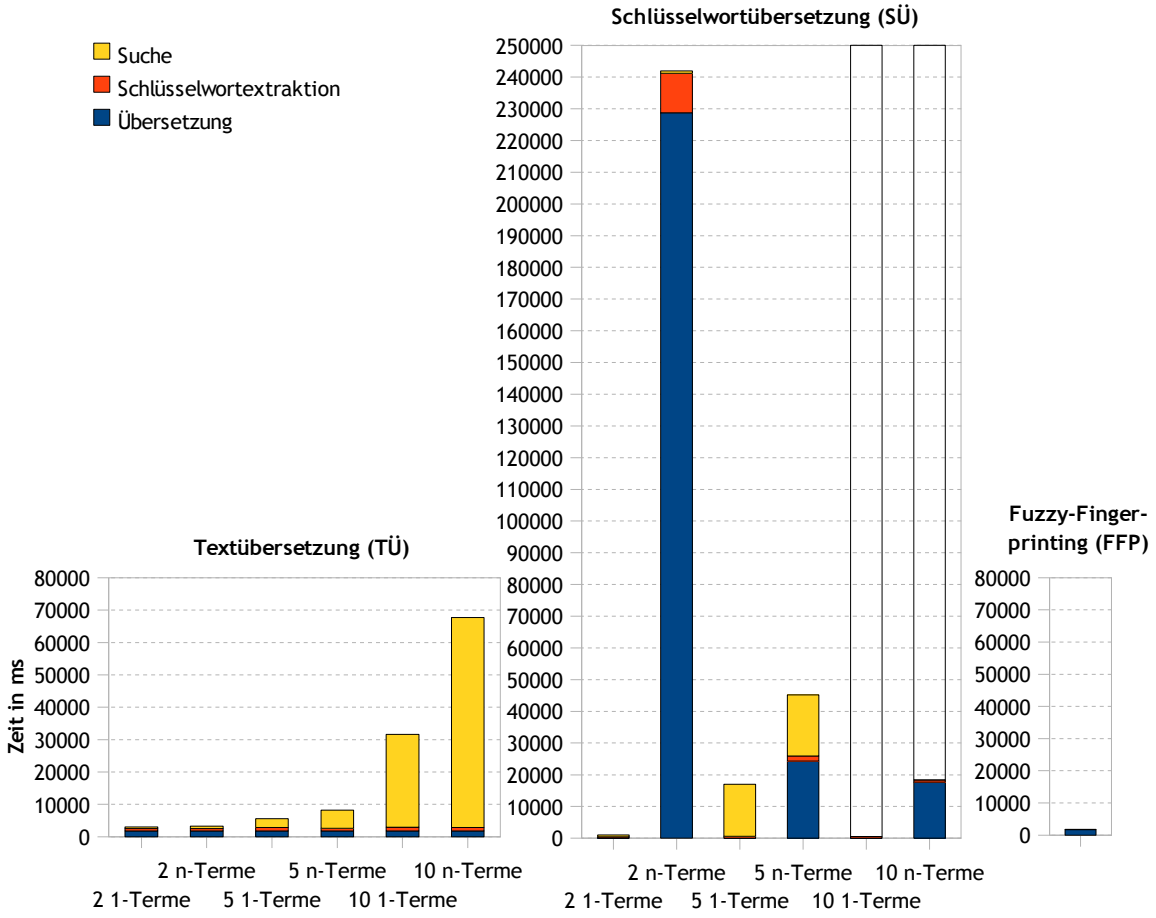


***Abbildung 4.12:*** *Laufzeiten der einzelnen Ansätze.*

rierung der Anfrage einige Zeit. Ursache dafür ist die Ambiguität der Übersetzungen. Durch die vielen Übersetzungsvarianten steigt die Anzahl der Teilanfragen enorm an. Bei der Verwendung von Wortgruppen als Schlüsselbegriffe tritt das Problem schon während der Übersetzungsphase auf. Ein Teil der Ambiguität wird durch Kookkurrenzanalyse beseitigt, indem die wahrscheinlichste Kombination der Übersetzungsvarianten der einzelnen Bestandteile der Wortgruppe ausgewählt wird. Wie bereits in Abschnitt 4.3.3 beschrieben, traten hier einige Schwierigkeiten auf. Im Diagramm zur Laufzeit der Schlüsselwortübersetzung ist zu sehen, daß fünf n-Terme weniger Zeit beanspruchen als zwei n-Terme. Das stimmt so nicht. Für zwei Terme konnten mehr Anfragen erfolgreich ausgewertet werden als für fünf. Vor allem Anfragen mit großen Wortgruppen bestehend aus drei und mehr Wörtern konnten aufgrund von Speicherproblemen nicht ausgewertet werden, so daß in der Auswertung von 5 n-

| | Übersetzungen | | | |
|---|---|---|---|---|
| | genau 1 | max. 3 | max. 5 | unbegrenzt |
| Laufzeit | 6508 ms | 22986 ms | 78826 ms | ? |
| Rang | 8.33 | 10.77 | 11.79 | ? |
| Recall | 0.93 | 0.93 | 0.88 | ? |
| MAP | 0.4769 | 0.4268 | 0.4106 | ? |
| Teilanfragen | 1023 | 5204 | 16543 | 178991 |
| - min. | 1023 | 1023 | 1023 | 1023 |
| - max. | 1023 | 36863 | 155519 | 3157415 |

***Tabelle 4.8:*** *Vergleich der Retrieval-Werte bei Beschränkung der Schlüsselwörter (10 1-Terme) auf solche, die nur eine, drei oder fünf Übersetzungen haben.*

Termen hauptsächlich Anfragen mit geringer Ambiguität enthalten sind, die also wenig Laufzeit beanspruchen und somit die Ergebnisse verfälschen.

Das Problem der vielen Teilanfragen während des Strukturierungsvorgangs kann unter Umständen mit einer veränderten Auswahl der Schlüsselbegriffe vermindert werden. Zu Schwierigkeiten führen vor allem Begriffe mit vielen Übersetzungsvarianten. Unter der Annahme, daß „gute“ Schlüsselwörter nicht mehrdeutig sind und dementsprechend wenige Übersetzungskandidaten existieren, läßt sich die Anzahl der Teilanfragen einschränken. Dazu werden aus einem Dokument nur Begriffe extrahiert, die nicht mehr als eine fest vorgegebene Anzahl an Übersetzungen besitzen. Tabelle 4.8 zeigt die Ergebnisse eines entsprechenden Versuchs. Aus 100 deutschen Wikipedia-Artikeln wurden zehn Schlüsselwörter entnommen, die nicht mehr als eine, drei oder fünf Übersetzungen hatten, und ermittelt, wie gut mit diesen übersetzten Schlüsselwörtern das entsprechende englische Dokument gefunden werden konnte. Es scheint, daß die Ergebnisse minimal besser sind, je weniger Übersetzungen die Schlüsselwörter haben. Dies könnte auch die These bestätigen, daß es sich hierbei um die geeigneteren Schlüsselwörter handelt. Trotz allem kann auf diese Weise nur die Anzahl der Teilanfragen, die bei der Textübersetzung entstehen, angenähert werden. Daher kann sich die Laufzeit diesem Verfahren nur annähern und nicht besser sein. Die Zeit, die zur Übersetzung des kompletten Dokuments benötigt wird, ist vergleichsweise gering und fällt daher nicht ins Gewicht.

| Kollektion | | FFP | SÜ | TÜ |
|---|---|---|---|---|
| Wikipedia | Rang | - | 14.81 | 11.69 |
| | Recall | 0.1028 | 0.8004 | 0.9285 |
| | Precision | 0.0015 | 0.0080 | 0.0092 |
| | MAP | - | 0.2891 | 0.3982 |
| | F-Measure | 0.0030 | 0.0158 | 0.0184 |
| JRC-Acquis | Rang | - | 39 | 12.98 |
| | Recall | 0.1710 | 0.2400 | 0.8360 |
| | Precision | 0.0523 | 0.0024 | 0.0084 |
| | MAP | - | 0.0249 | 0.3970 |
| | F-Measure | 0.0801 | 0.0048 | 0.0166 |
| gesamt | Laufzeit | 1770 ms | 45202 ms | 8157 ms |

***Tabelle 4.9:*** *Vergleich aller drei Verfahren anhand wichtiger Retrieval-Eigenschaften. Für jedes Verfahren wurde die Variante mit den besten Resultaten ausgewählt: Fuzzy-Fingerprinting mit zwei Hashfunktionen (1.25 (65) und 1.55 (65)), Schlüsselwortübersetzung und Textübersetzung mit jeweils fünf n-Termen.*

## 4.4 Fazit

In Tabelle 4.9 sind alle drei Verfahren mit den jeweils besten Parametern einander noch einmal gegenübergestellt. Abschließend läßt sich folgendes Fazit ziehen:

***Fuzzy-Fingerprinting***: Fuzzy-Fingerprinting ist in Bezug auf das Laufzeitverhalten unschlagbar. Aufwendigster Schritt ist die Indizierung, die nur einmal durchgeführt werden muß. Die Suche ist dann in konstanter Zeit möglich. Jedoch ist die Retrieval-Qualität wenig befriedigend. Möglicherweise ist eine andere Kombination der Parameter erfolgreicher. Im Rahmen der Untersuchungen konnte allerdings keine geeignetere Variante gefunden werden. Entweder ist die Ähnlichkeit zwischen maschineller Übersetzung und Originaltext nicht hoch genug (wobei das im Fall der JRC-Acquis-Kollektion unwahrscheinlich ist) oder das Verfahren ist für maschinell übersetzte Texte nicht gut geeignet. Zudem ist beim Fuzzy-Fingerprinting im Gegensatz zu den beiden anderen Verfahren die Retrieval-Qualität stark von der Länge des zu untersuchenden Textes abhängig.

***Schlüsselwortübersetzung***: Die Schlüsselwortübersetzung scheitert an der Ambiguität. Je mehr Übersetzungsvarianten existieren, desto mehr Teilanfragen sind wäh-

rend der Anfragestrukturierung notwendig und umso stärker steigt die benötigte Laufzeit bzw. umso größer ist die Gefahr, daß die Implementierung zu keinem Ergebnis kommt. Weiterhin ist das Verfahren von der Qualität verschiedener Ressourcen, wie beispielsweise dem Wörterbuch, abhängig und, wie bereits das Vorexperiment gezeigt hat, nur schwer unter Kontrolle zu bringen.

***Textübersetzung***: Insgesamt betrachtet, schneidet die Textübersetzung hinsichtlich Laufzeitverhalten und Retrieval-Qualität am besten ab. Prinzipiell hat dieses Verfahren auch ein Ambiguitätsproblem. Vorteilhaft ist jedoch, daß die Übersetzungsqualität durch das zugrundeliegende statistische Lernverfahren umso besser wird, je länger der Web-Service im Einsatz ist. Außerdem steht Google wesentlich mehr Rechenleistung zur Verfügung.

# 5 Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurden drei Verfahren zum sprachübergreifenden Retrieval von ähnlichen Dokumenten aus großen Textkollektionen untersucht. Im Kontext der Plagiatanalyse sollten sie hinsichtlich Umsetzbarkeit, Retrieval-Eigenschaften und Laufzeitverhalten miteinander verglichen werden. Zielstellung war es, zu einem Dokument, welches verdächtigt wird, ein Übersetzungsplagiat zu sein, das Original in seiner Ausgangssprache zu finden.

Zunächst wurden grundlegende Begriffe des Information-Retrieval erläutert und Verfahren der Ähnlichkeitssuche, also der Suche mit Hilfe eines Beispieldokuments, vorgestellt. Weiterhin erfolgte die Defintion eines Plagiats, insbesondere eines Plagiats, welches durch Übersetzungen entstanden ist. Die Plagiatanalyse wurde als Retrieval-Aufgabe beschrieben und Möglichkeiten zur Entdeckung verschiedener Arten von Plagiaten wurden aufgezeigt.

Zur sprachübergreifenden Plagiatanalyse wurden drei verschiedene Strategien vorgeschlagen, ausgehend von einem verdächtigen Dokument ähnliche Dokumente in anderen Sprache zu finden. Der erste Ansatz, Schlüsselwortübersetzung, basiert auf monolingualem Schlüsselwort-Retrieval. Dazu müssen dem verdächtigen Dokument Schlüsselbegriffe entnommen und diese übersetzt werden. Hierzu wurde auf ein Verfahren aus dem Bereich des sprachübergreifenden Information-Retrieval zur Übersetzung von Anfragen, also Schlüsselwörtern, zurückgegriffen. Die Schwierigkeiten lagen in der Auswahl der geeigneten Übersetzung eines Begriffs, ohne daß Kontextwissen über dessen Bedeutung vorliegt. Der zweite Ansatz, Textübersetzung, beruht ebenfalls auf monolingualem Schlüsselwort-Retrieval. Jedoch wird hier erst das komplette Dokument übersetzt, bevor die zur Suche verwendeten Schlüsselbegriffe extrahiert werden. Interessant war hierbei, inwiefern falsche Übersetzungen und gramma-

tikalisch nicht korrekte Sätze die Retrieval-Qualität beeinflussen. Der dritte Ansatz, Fuzzy-Fingerprinting, nutzt Hashing-basierte Ähnlichkeitssuche. Mit Hilfe einer ähnlichkeitssensitiven Hashfunktion wird für jedes Dokument ein Hashwert berechnet, so daß ähnliche Dokumente mit hoher Wahrscheinlichkeit auf denselben Hashwert abgebildet werden. Im Falle der Plagiatanalyse stellt eine Kollision der Hashwerte ein Indiz für ein Plagiatvergehen dar. Durch Hashing-basierte Suche können die Laufzeiteigenschaften von Hashtabellen ausgenutzt werden. Es stellte sich die Frage, inwiefern Fuzzy-Fingerprinting für maschinell übersetzte Texte und damit sprachübergreifende Plagiatanalyse geeignet ist.

Die Evaluierung fand anhand zweier bilingualer Korpora statt. Einer enthielt Dokumente in zwei Sprachen, welche exakte Übersetzungen voneinander sind. Der andere bestand aus Dokumenten, die paarweise dasselbe Thema in unterschiedlicher Sprache behandeln. Verglichen wurden alle drei Verfahren anhand verschiedener Kriterien wie Retrieval-Qualität, Laufzeitverhalten und Abhängkeit von der Textlänge. Die schlüsselwortbasierten Verfahren wurden speziell dahingehend ausgewertet, wie die Anzahl und Art der extrahierten Begriffe die Ergebnisse beeinflußt. Im Mittelpunkt der Untersuchungen stand außerdem der Einfluß der Übersetzungsmethodik. Das Verfahren zur Übersetzung von Schlüsselwörtern löst das Problem der Ambiguität nicht vollständig auf, sondern verwendet mehrere Übersetzungskandidaten zum Retrieval. Laut Kraaij u. a. (2003) soll sich dadurch der Einfluß falscher Übersetzungen verringern. Die alternativen Übersetzungen sollen den Effekt von Query-Expansion-Verfahren haben und den Recall verbessern. Im Gegensatz dazu steht bei der Übersetzung von Dokumenten wesentlich mehr Kontext zur Verfügung und es wird jeweils genau eine Übersetzung ausgewählt. Jedoch können die entstehenden Sätze sowohl syntaktisch als auch semantisch inkorrekt sein. Untersucht wurde, welches Verfahren trotz der genannten Schwierigkeiten die besseren Resultate liefert. Schließlich wurde die Eignung von Fuzzy-Fingerprinting für die sprachübergreifende Suche überprüft.

Die Schlüsselwortübersetzung scheitert an der Ambiguität. Je mehr Übersetzungsvarianten existieren, desto mehr Teilanfragen sind während der Anfragestrukturierung notwendig und umso stärker steigt die benötigte Laufzeit bzw. umso größer ist die Gefahr, daß die Implementierung zu keinem Ergebnis kam. Das Verfahren ist zudem

stark von der Qualität verschiedener Ressourcen, wie beispielsweise dem Wörterbuch, abhängig und aufgrund der vielen einzeln zu optimierenden Komponenten nur schwer zu kontrollieren.

Fuzzy-Fingerprinting ist in Bezug auf das Laufzeitverhalten unschlagbar. Aufwendigster Schritt ist die Indizierung, die nur einmal durchgeführt werden muß. Die Suche ist dann in konstanter Zeit möglich. Die Ergebnisse der Suche sind jedoch nicht befriedigend. Im Rahmen der Untersuchungen konnte keine Parameterkombination gefunden werden, die das Retrieval signifikant verbessert. Daher lag die Schlußfolgerung nahe, daß das Verfahren für maschinell übersetzte Texte nicht geeignet ist. Es zeigte sich außerdem, daß die Qualität der Ergebnisse stark von der Länge des untersuchten Textes abhängig ist. Für die Zukunft stellt sich die Frage, ob eine Hashfunktion gefunden werden kann, die sprachübergreifend Ähnlichkeit bestimmen kann. In diesem Fall wäre der Fingerprinting-Ansatz sehr vielversprechend.

Insgesamt schnitt das Verfahren zur Textübersetzung am besten ab. Prinzipiell hat dieses Verfahren auch ein Ambiguitätsproblem. Vorteilhaft ist jedoch, daß die Übersetzungsqualität durch das zugrundeliegende statistische Lernverfahren umso besser wird, je länger der Web-Service im Einsatz ist. Ungünstig an diesem Ansatz ist die Verwendung eines Services eines Fremdanbieters, wenn Texte mit vertraulichem Inhalt untersucht werden sollen.

Problematisch bei beiden schlüsselwortbasierten Verfahren ist der Schritt der Anfragestrukturierung, da hier sehr viele Teilanfragen an Suchmaschinen erzeugt werden. Sollen wesentlich längere Texte als die in dieser Arbeit verwendeten analysiert werden, müßte die Anzahl der extrahierten Schlüsselbegriffe erhöht werden, wodurch die Anzahl der generierten Teilanfragen exponentiell ansteigt. Daher muß eine bessere Heuristik zur Erzeugung von Anfragen gefunden werden, welche die Anzahl der Anfragen stark verringert ohne den Recall zu senken.

Abschließend bleibt festzustellen, daß es – unabhängig vom verwendeten Verfahren – nahezu unmöglich ist, das Original eines Übersetzungsplagiats zu finden, wenn die Sprache, aus der übersetzt wurde, unbekannt ist.

# Literaturverzeichnis

[Adriani und van Rijsbergen 1999] Adriani, Mirna ; Rijsbergen, C. J. van: Term Similarity-Based Query Expansion for Cross-Language Information Retrieval. In: *ECDL '99: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries*. London, UK : Springer-Verlag, 1999, S. 311–322. – ISBN 3-540-66558-7

[Anderka 2007] Anderka, Maik: *Methoden zur sprachübergreifenden Plagiaterkennung*, Universität Paderborn, Fakultät für Elektrotechnik, Informatik und Mathematik, Masterarbeit, Oktober 2007

[Baayen u. a. 2002] Baayen, Harald ; van Halteren, Hans ; Neijt, Anneke ; Tweedie, Fiona: An Experiment in Authorship Attribution. In: *6es Journées internationales d'Analyse statistique des Données Textuelles*. St. Malo, 2002, S. 29–37

[Baeza-Yates und Ribeiro-Neto 1999] Baeza-Yates, Ricardo A. ; Ribeiro-Neto, Berthier A.: *Modern Information Retrieval*. Addison Wesley, 1999. – ISBN 0-201-39829-X

[Ballesteros und Croft 1997] Ballesteros, Lisa ; Croft, W. B.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM, 1997, S. 84–91. – ISBN 0-89791-836-3

[Ballesteros 2001] Ballesteros, Lisa A.: *Resolving ambiguity for cross-language information retrieval: a dictionary approach*, Graduate School of the University of Massachusetts Amherst, Department of Computer Science, Dissertation, 2001. – W. Bruce Croft

[Brockhaus 1992] Brockhaus, Enzyklopädie in 24 Bänden: *Eintrag: Plagiat*. 1992. – Band 17, 19., völlig neu bearbeitete Auflage, F. A. Brockhaus GmbH, Mannheim

[Brown u. a. 1990] Brown, Peter F. ; Cocke, John ; Pietra, Stephen A. D. ; Pietra, Vincent J. D. ; Jelinek, Fredrick ; Lafferty, John D. ; Mercer, Robert L. ; Roossin, Paul S.: A statistical approach to machine translation. In: *Comput. Linguist.* 16 (1990), Nr. 2, S. 79–85. – ISSN 0891-2017

[Callan u. a. 1992] Callan, James P. ; Croft, W. B. ; Harding, Stephen M.: The Inquery Retrieval System. In: *Proceedings of the Third International Conference on Database and Expert Systems Applications*, Springer-Verlag, 1992, S. 78–83

[Carbonell u. a. 1997] Carbonell, Jaime G. ; Yang, Yiming ; Frederking, Robert E. ; Brown, Ralf D. ; Geng, Yibing ; Lee, Danny: Translingual Information Retrieval: A Comparative Evaluation. In: *IJCAI*, 1997, S. 708–715

[Chen u. a. 1999] Chen, Hsin-Hsi ; Bian, Guo-Wei ; Lin, Wen-Cheng: Resolving translation ambiguity and target polysemy in cross-language information retrieval. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Morristown, NJ, USA : Association for Computational Linguistics, 1999, S. 215–222. – ISBN 1-55860-609-3

[Dumais u. a. 1997] Dumais, Susan T. ; Letsche, Todd A. ; Littman, Michael L. ; Landauer, Thomas K.: Automatic cross-language retrieval using latent semantic indexing. In: *AAAI'97 Cross-Language Text and Speech Retrieval*, 1997

[Ebbertz 2002] Ebbertz, Martin: *Netz-Tipp-Studie: Das Internet spricht Englisch ... und neuerdings auch Deutsch.Sprachen und ihre Verbreitung im World-Wide-Web*. 2002. – http://www.netz-tipp.de/sprachen.html am 07.08.2008

[Gao u. a. 2002] Gao, Jianfeng ; Zhou, Ming ; Nie, Jian-Yun ; He, Hongzhao ; Chen, Weijun: Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM, 2002, S. 183–190. – ISBN 1-58113-561-0

[Hart und Friesner 2004] Hart, Mike ; Friesner, Tim: Plagiarism and Poor Academic Practice - A threat to the Extension of e-Learning in Higher Education? In: *Electronic Journal of e-Learning* 2 (2004), dec, Nr. 2. – ISSN 1479-4403

[Hull 1997] Hull, David A.: Using structured queries for disambiguation in cross-language information retrieval. In: *American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval*, 1997, S. 84–98

[Jang u. a. 1999] Jang, Myung-Gil ; Myaeng, Sung H. ; Park, Se Y.: Using mutual information to resolve query translation ambiguities and query term weighting. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Morristown, NJ, USA : Association for Computational Linguistics, 1999, S. 223–229. – ISBN 1-55860-609-3

[Jurafsky und Martin 2000] Jurafsky, Daniel ; Martin, James H.: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ, USA : Prentice Hall PTR, 2000. – ISBN 0130950696

[Klüger 2006] Klüger, Karsten: *Automatische Extraktion von Schlüsselwörtern aus Text*, Bauhaus-Universität Weimar, Fakultät Medien, Mediensysteme, Diplomarbeit, Juni 2006

[Koppel und Schler 2003] Koppel, Moshe ; Schler, Jonathan: Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In: *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*. Acapulco, Mexico, 2003, S. 69–72

[Kraaij u. a. 2003] Kraaij, Wessel ; Nie, Jian-Yun ; Simard, Michel: Embedding web-based statistical translation models in cross-language information retrieval. In: *Comput. Linguist.* 29 (2003), Nr. 3, S. 381–419. – ISSN 0891-2017

[Lavrenko u. a. 2002] Lavrenko, Victor ; Choquette, Martin ; Croft, W. B.: Cross-lingual relevance models. In: *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM, 2002, S. 175–182. – ISBN 1-58113-561-0

[Liu und Jin 2005] Liu, Yi ; Jin, Rong: Query translation disambiguation as graph partitioning. In: *AAAI'05: Proceedings of the Twentieth National Conference in Artificial Intelligence, Seventeenth Conference on Innovative Applications of Artificial Intelligence*. Pittsburgh, PA : AAAI Press, 2005

[Lumb 2008] Lumb, Wolf: *Das Urheberrecht*. 2008. – http://home.arcor.de/wolf... ...lumb/verlag/lehrmaterial/verlagsbetriebslehre/urheberrecht.htm am 27.08.2008

[Lyman und Varian 2003] Lyman, Peter ; Varian, Hal R.: *How Much Information*. 2003. – http://www.sims.berkeley.edu/how-much-info-2003 am 31.07.2008

[Maeda u. a. 2000] Maeda, Akira ; Sadat, Fatiha ; Yoshikawa, Masatoshi ; Uemura, Shunsuke: Query term disambiguation for Web cross-language information retrieval using a search engine. In: *IRAL '00: Proceedings of the fifth international workshop on on Information retrieval with Asian languages*. New York, NY, USA : ACM, 2000, S. 25–32. – ISBN 1-58113-300-6

[Markengesetz 1994] *Markengesetz vom 25. Oktober 1994 (BGBl. I S. 3082 (1995, 156); 1996, 682), zuletzt geändert durch Artikel 12 Abs. 3 des Gesetzes vom 13. Dezember 2007 (BGBl. I S. 2897)*. 1994. – http://bundesrecht.juris.de/markeng/BJNR308210994.html am 27.08.2008

[Maurer u. a. 2006] Maurer, Hermann ; Kappe, Frank ; Zaka, Bilal: Plagiarism - A Survey. In: *Journal of Universal Computer Science* 12 (2006), aug, Nr. 8, S. 1050–1084

[McEnery und Xiao 2007] McEnery, Anthony ; Xiao, Zhonghua: *Incorporating Corpora: Translation and the Linguist. Translating Europe.* Kap. XX - Parallel and comparable corpora: What are they up to? Clevedon, UK : Multilingual Matters, 2007. – ISBN 978-1-85359-986-6

[Meyer zu Eißen und Stein 2006] Meyer zu Eißen, Sven ; Stein, Benno: Intrinsic Plagiarism Detection. In: Lalmas, Mounia (Hrsg.) ; MacFarlane, Andy (Hrsg.) ; Rüger, Stefan (Hrsg.) ; Tombros, Anastasios (Hrsg.) ; Tsikrika, Theodora (Hrsg.) ; Yavlinsky, Alexei (Hrsg.): *Advances in Information Retrieval: Proceedings of the 28th European Conference on IR Research, ECIR 2006, London* Bd. 3936 LNCS, Springer, 2006, S. 565–569. – ISBN 3-540-33347-9

[Meyer zu Eißen u. a. 2005] Meyer zu Eißen, Sven ; Stein, Benno ; Potthast, Martin: The Suffix Tree Document Model Revisited. In: Tochtermann, Klaus (Hrsg.) ; Maurer, Hermann (Hrsg.): *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz, Austria*, Know-Center, Juli 2005 (Journal of Universal Computer Science), S. 596–603. – ISSN 0948-695x

[Monz und Dorr 2005] Monz, Christof ; Dorr, Bonnie J.: Iterative translation disambiguation for cross-language information retrieval. In: *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM, 2005, S. 520–527. – ISBN 1-59593-034-5

[Pirkola u. a. 2001] Pirkola, Ari ; Hedlund, Turid ; Keskustalo, Heikki ; Järvelin, Kalervo: Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings. In: *Inf. Retr.* 4 (2001), Nr. 3-4, S. 209–230. – ISSN 1386-4564

[Pirkola u. a. 2003] Pirkola, Ari ; Puolamäki, Deniz ; Järvelin, Kalervo: Applying query structuring in cross-language retrieval. In: *Inf. Process. Manage.* 39 (2003), Nr. 3, S. 391–402. – ISSN 0306-4573

[Potthast u. a. 2008] Potthast, Martin ; Stein, Benno ; Anderka, Maik: A Wikipedia-Based Multilingual Retrieval Model. In: Macdonald, Craig (Hrsg.) ; Ounis, Iadh (Hrsg.) ; Plachouras, Vassilis (Hrsg.) ; Ruthven, Ian (Hrsg.) ; White, Ryen W. (Hrsg.): *30th European Conference on IR Research, ECIR 2008, Glasgow* Bd. 4956 LNCS. Berlin Heidelberg New York : Springer, 2008, S. 522–530. – ISBN 978-3-540-78645-0

[Rivest 1992] Rivest, Ronald L.: *The MD5 Message Digest Algorithm*. April 1992. – Internet RFC 1321

[Sadat 2002] Sadat, Fatiha: Cross-Language Information Retrieval via Hybrid Combination of Query Expansion Techniques. In: *Proceedings of the ACL Student Research Workshop*. Philadelphia : Association for Computational Linguistics, July 2002, S. 48–53

[Salton u. a. 1975] Salton, G. ; Wong, A. ; Yang, C. S.: A vector space model for automatic indexing. In: *Communications of the ACM* 18 (1975), Nr. 11, S. 613–620. – ISSN 0001-0782

[Shapiro und Taksa 2003] Shapiro, Jacob ; Taksa, Isak: Constructing Web search queries from the user's information need expressed in a natural language. In: *SAC '03: Proceedings of the 2003 ACM symposium on Applied computing*. New York, NY, USA : ACM, 2003, S. 1157–1162. – ISBN 1-58113-624-2

[Sheridan u. a. 1997] Sheridan, Paraic ; Braschler, Martin ; Schäuble, Peter: Cross-Language Information Retrieval in a Multilingual Legal Domain. In: *ECDL '97: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*. London, UK : Springer-Verlag, 1997, S. 253–268. – ISBN 3-540-63554-8

[Stein 2005] Stein, Benno: Fuzzy-Fingerprints for Text-Based Information Retrieval. In: Tochtermann, Klaus (Hrsg.) ; Maurer, Hermann (Hrsg.): *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz, Austria*, Know-Center, Juli 2005 (Journal of Universal Computer Science), S. 572–579. – ISSN 0948-695x

[Stein 2007] Stein, Benno: *Vorlesungsskript: Information Retrieval - Unit. Modelle und Prozesse im IR*. 2007. – http://www.uni-weimar.de/medien/webis/... .../teaching/lecturenotes/information-retrieval/part-retrieval-models/unit-retrie... ....val-models.ps.pdf am 05.08.2008

[Stein u. a. 2008] Stein, Benno ; Loose, Fabian ; Becker, Steffen ; Potthast, Martin: *Retrieval-Technologien für die Plagiaterkennung in Programmen*. 2008

[Stein u. a. 2007] Stein, Benno ; Meyer zu Eissen, Sven ; Potthast, Martin: Strategies for Retrieving Plagiarized Documents. In: Charles Clarke, Noriko Kando Wessel K. (Hrsg.) ; Vries, Arjen de (Hrsg.): *30th Annual International ACM SIGIR Conference*, ACM, July 2007, S. 825–826. – ISBN 987-1-59593-597-7

[Stein und Potthast 2007] Stein, Benno ; Potthast, Martin: Applying Hash-based Indexing in Text-based Information Retrieval. In: Marie-Francine Moens and Tinne Tuytelaars and Arjen P. de Vries (Hrsg.): *7th Dutch-Belgian Information Retrieval*

*Workshop (DIR 2007).* Leuven, Belgium : Faculty of Engineering, Universiteit Leuven, März 2007, S. 29–35. – ISBN 978-90-5682-771-7

[Steinberger u. a. 2006] Steinberger, Ralf ; Pouliquen, Bruno ; Widiger, Anna ; Ignat, Camelia ; Erjavec, Tomaž ; Tufiş, Dan ; Varga, Dániel: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006).* Genoa, Italien, 2006. – Download der Kollektion am 19.05.2008 von http://wt.jrc.it/lt/Acquis/

[The Official Google Blog 2008] The Official Google Blog: *We knew the web was big...* 2008. – http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html am 31.07.2008

[Trenkmann 2008] Trenkmann, Martin: Netspeak - *Ein Assistent zum Verfassen fremdsprachiger Texte*, Bauhaus-Universität Weimar, Fakultät Medien, Mediensysteme, Bachelorarbeit, Juni 2008. – Kapitel 3.2

[TU Chemnitz 2007] TU Chemnitz: *BEOLINGUS*. 2007. – Deutsch-englische Version vom 06.09.2007. http://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/

[Urheberrechtsgesetz 1965] *Urheberrechtsgesetz vom 9. September 1965 (BGBl. I S. 1273), zuletzt geändert durch Artikel 12 Abs. 4 des Gesetzes vom 13. Dezember 2007 (BGBl. I S. 2897).* 1965. – http://bundesrecht.juris.de/urhg/BJNR012730965.html am 27.08.2008

[van Rijsbergen 1979] van Rijsbergen, C. J.: *Information Retrieval.* 2nd Edition. Newton, MA, USA : Butterworth-Heinemann, 1979. – ISBN 0408709294

[Weber-Wulff 2007] Weber-Wulff, Debora: *Fremde Federn Finden - Eine E-Learning Einheit. Plagiat in der Literatur.* 2007. – http://plagiat.fhtw-berlin.de/ff/vielfalt/... .../2_1/literatur am 27.08.2008

[Welt Online 2004] Welt Online: *Plagiate. Haben Nabokov und Thomas Mann geklaut?* 2004. – http://www.welt.de/print-welt/article311124/Haben_Nabokov... ..._und_Thomas_Mann_geklaut.html am 27.08.2008