Norman Meuschke, Nicole Walger & Bela Gipp

F 5 Plagiat

1 Einleitung

Plagiate sind kein neues Phänomen, wie ein Blick in die Literatur (Theisohn 2009) belegt. Zahlreiche seit dem Jahr 2011 aufgedeckte Plagiatsfälle, wovon einige prominente Politiker*innen betrafen, verdeutlichen die unverändert hohe Relevanz des Themas bis heute. Viele dieser Fälle wurden durch Freiwillige enthüllt, die auf eigens gegründeten Plattformen wie VroniPlag Wiki (2021) in Deutschland (bislang 212 Fälle) und Dissernet (2021) in Russland (bislang 1216 Fälle) wissenschaftliche Plagiate dokumentieren. Das gemeinnützige Projekt Retraction Watch (2021) berichtet über zurückgezogene Artikel in wissenschaftlichen Fachzeitschriften. Für 4037 der 31460 hier bis November 2021 erfassten Fälle (ca. 13 %) waren Plagiate ursächlich.

Unerkannte Plagiate schaden dem Kompetenzerwerb und der Beurteilung des Lernerfolgs, was zu ungerechtfertigten Karrierevorteilen für Plagiator*innen führen kann. Plagiate in Forschungspublikationen behindern die Nachvollziehbarkeit von Ideen, Überprüfung von Behauptungen, Replikation von Experimenten und Korrektur von Ergebnissen (Weber-Wulff 2014, S. 22). Auch können Forschungsgelder zu Unrecht für plagiierte Ideen vergeben oder plagiierte Publikationen als Ergebnis von Forschungsprojekten unerkannt akzeptiert werden. Die Prüfung und Sanktionierung plagiierter Arbeiten verursacht einen hohen Arbeitsaufwand. Hochschulen, Verlage, wissenschaftliche Fachzeitschriften und Konferenzen sowie Fördermittelgeber stehen daher vor der Herausforderung, tragfähige Lösungen für die Erkennung und Prävention wissenschaftlicher Plagiate zu finden.

2 Der Plagiatsbegriff und seine rechtliche Verortung

Als griffige und sämtliche Formen[1] des Wissenschaftsplagiats umfassende Definition ist die Formulierung von Teddi Fishman zu sehen:

> Plagiarismus tritt auf, wenn jemand (1) Worte, Ideen oder Arbeitsergebnisse verwendet, (2) die einer anderen identifizierbaren Person oder Quelle zugeordnet werden können, (3) ohne die Quelle, aus der übernommen wurde, auszuweisen. (4) Dies in einem Zusammenhang, in dem die berechtigte Erwartung eigenständiger Autorschaft besteht. (5) Mit dem Ziel, einen Vorteil, Ansehen oder einen Gewinn zu erlangen, der nicht unbedingt monetär sein muss. (übersetzt aus: Fishman, T. (2009))

Auch wenn der Volksmund mit Plagiat allgemein die unrechtmäßige Übernahme fremden geistigen Eigentums im Rahmen wissenschaftlicher Arbeiten gleichsetzt, ist der Plagiatsbegriff nicht für alle Fachdisziplinen identisch und in Deutschland als Rechtsbegriff auch nicht eindeutig definiert. In Abhängigkeit vom jeweiligen Sach- und Rechtsgebiet bedarf er der Auslegung.

1 Vollplagiat, Übersetzungsplagiat, Strukturplagiat, Selbstplagiat, Paraphrase, ungenügendes Zitieren, Ideenplagiat.

https://doi.org/10.1515/9783110769043-069

Unterschiedliche Rechtsgebiete, wie das Urheberrecht[2] und das Hochschulrecht betreffend, kann ein Plagiat verschiedene Gesetze tangieren, auch strafbar sein und neben Schadensersatzforderungen des plagiierten Urhebers bzw. der Urheberin universitätsrechtliche Sanktionen nach sich ziehen. Letztere sind von den Hochschulen selbst festzulegen. Die Folgen wissenschaftlichen Fehlverhaltens in Form des Plagiats können dabei vielfältig sein. Sie werden in den Regelwerken zum wissenschaftlichen Arbeiten festgeschrieben und reichen von Geldbußen über Prüfungswiederholung, Nichtbestehen und Exmatrikulation bis Aberkennung von Titeln und Freiheitsstrafe.

Entsprechend der den Hochschulen obliegenden Aufgabe der Einhaltung wissenschaftlicher Redlichkeit umfasst ihr Auftrag neben der Vermittlung und Förderung der Regeln zur Sicherung der guten wissenschaftlichen Praxis auch die Kontrolle ihrer Einhaltung und das Aufdecken von Täuschungsversuchen. Der Einsatz unterstützender technologischer Verfahren im Beurteilungsverfahren zum automatisierten Abgleich von Texten auf Ähnlichkeit mit anderen Quellen gewinnt dabei aufgrund der kaum noch überschaubaren Menge an Quellen zunehmend an Attraktivität (dazu auch Dagli-Yalcinkaya 2021, S. 16). An manchen Hochschulen kommt Plagiatserkennungssoftware bereits standardmäßig für die Aufdeckung von Täuschungsversuchen in studentischen Arbeiten zum Einsatz, während andere aufgrund prüfungs-, urheber- und datenschutzrechtlicher Unsicherheiten noch zögerlich sind.

3 Rechtlicher Rahmen für den Einsatz von Plagiatserkennungssoftware

Die Verwendung von Plagiatserkennungssoftware zur Unterstützung im Beurteilungsverfahren ist den Hochschulen unter Einhaltung folgender Voraussetzungen sowohl aus prüfungs-, urheber- und datenschutzrechtlicher Sicht möglich[3] (Dagli-Yalcinkaya 2021, S. 21):

Studierende und Promovierende treten mit der Immatrikulation in ein öffentlich-rechtliches Rechtsverhältnis mit ihrer Hochschule. Sie erkennen damit die in den Prüfungs- und Studienordnungen geregelten rechtlichen Rahmenbedingungen der Prüfungsverfahren an und verpflichten sich zur Einhaltung der Regeln guter wissenschaftlicher Praxis. Damit willigen sie in die Überprüfung im Rahmen der Leistungsüberprüfung ein. Da die Verwendung von Plagiatserkennungssoftware, z. B. mit Upload-, Übermittlungs-, Speicher- und Veranschaulichungsprozessen, aufseiten der Softwareanbieter und Suchmaschinenbetreiber in das Vervielfältigungsrecht des Urhebers oder der Urheberin eingreift, ist die Einräumung der Nutzungsrechte durch die Verfasser*innen der zur prüfenden Arbeit erforderlich.

Ebenso bedarf die automatisierte Verarbeitung personenbezogener Daten der Prüflinge gemäß europäischer Datenschutzgrundverordnung gesetzlichen Rechtfertigungsgründen. Ihre vorherige Information bspw. über die Aufnahme einer Datenschutzinformation in die Studien-, Prüfungs- oder Promotionsordnung sowie der Abschluss einer Vereinbarung zur Auftragsverarbeitung gemäß Art. 28 Abs. 3 DSGVO mit dem Software-

2 Hier insb. § 13 UrhG, § 15 UrhG, § 23 UrhG, 51 UrhG und § 63 UrhG.

3 Zu diesem Ergebnis kommt auch ein vom DH-NRW-geförderten Projekt PlagStop.nrw beauftragtes unveröffentlichtes Rechtsgutachten der KPMG Law Rechtsanwaltsgesellschaft mbH (Dagli-Yalcinkaya 2021, S. 21).

Anbieter schaffen hierfür den erforderlichen Rahmen. Daneben ist unter Bezugnahme auf das vom Europäischen Gerichtshof (EuGH) nicht als ausreichend befundene Datenschutzniveau in den USA sicherzustellen, dass ein Datentransfer in diese unterbleibt und die Funktionalität, im Internet nach Plagiaten zu suchen, nicht über in den USA ansässige Suchmaschinendienste, wie Microsoft oder Google erfolgt.

Die Entscheidung, ob es sich bei den durch Softwareeinsatz identifizierten Textähnlichkeiten um ein Plagiat handelt, obliegt den Prüfenden. Der Softwareeinsatz stellt für sie lediglich eine digitale Unterstützung bei der Identifikation von Textähnlichkeiten dar.

4 Plagiatserkennungstechnologie

Technologische Verfahren für die Plagiatssuche folgen entweder dem extrinsischen oder dem intrinsischen Paradigma (vgl. z. B. Foltýnek, Meuschke & Gipp, 2019):

Das *extrinsische* Paradigma beschreibt Verfahren, die zu überprüfende Dokumente mit einer umfangreichen Dokumentkollektion (der sog. Referenzkollektion) vergleichen. Ziel ist es, alle Dokumente, die einen bestimmten Grad der Ähnlichkeit zum überprüften Dokument überschreiten, für eine intensivere Begutachtung bereitzustellen. Extrinsische Plagiatsanalyse ist damit ein Information-Retrieval-Szenario (vgl. Kapitel C 1 Informationswissenschaftliche Perspektiven des Information Retrieval).

Das *intrinsische* Paradigma umfasst Verfahren, die das Eingabedokument auf unterschiedliche Schreibstile hin analysieren, ohne Vergleiche mit anderen Dokumenten durchzuführen. Stilistische Unterschiede betrachten diese Verfahren als Indikatoren für mögliche Plagiate.

4.1 Extrinsische Plagiatsanalyse

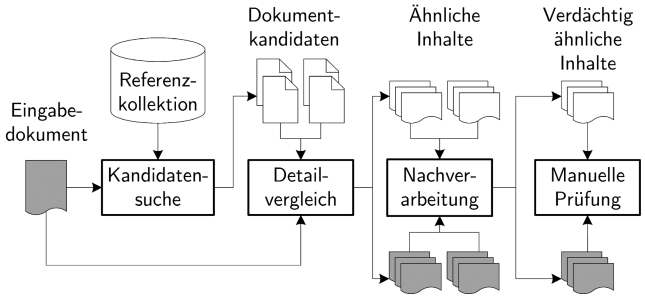

Abb. 1: Ablauf einer extrinsischen Plagiatsanalyse

Extrinsische Plagiatsanalyseverfahren folgen typischerweise dem in Abbildung 1 dargestellten mehrstufigen Prozess. In der Phase der *Kandidatensuche* filtern besonders effiziente Algorithmen die Referenzkollektion nach Dokumenten, die potenziell eine Quelle für Inhalt im überprüften Dokument sein könnten. Während des *Detailvergleichs* wird

das Eingabedokument paarweise mit jedem der zuvor gefundenen Kandidatendokumente verglichen, um das Ausmaß und die Position ähnlicher Inhalte in beiden Dokumenten zu identifizieren. Im Schritt der *Nachverarbeitung* werden als ähnlich identifizierte Inhalte einer wissensbasierten Filterung unterzogen, um typische Fehlalarme zu vermeiden. Korrekte Zitate sind typische Beispiele für fälschlich als verdächtig erkannte Inhalte. Für die *manuelle Prüfung* werden den Nutzenden die als verdächtig ähnlich eingestuften Inhalte im Eingabe- und den potenziellen Quelldokumenten präsentiert.

4.1.1 Kandidatensuche

Ziel dieser Phase ist es, sämtliche Dokumente zu finden, die inhaltliche Ähnlichkeiten zum überprüften Dokument aufweisen. Der Ähnlichkeitsschwellwert ist typischerweise niedrig, da es in diesem Schritt wichtiger ist, möglichst keine Quelldokumente auszuschließen, als unverdächtige Dokumente als Kandidaten zu behandeln und weiterzuverarbeiten. Werden Quelldokumente in diesem Schritt nicht gefunden, können sie auch in den nachfolgenden Prozessschritten, die die Ergebnismenge lediglich weiter einengen, nicht erkannt werden.

Die Referenzkollektion umfasst in der Regel mehrere Millionen bis mehrere hundert Millionen Dokumente, da sie zumeist Internetquellen einschließt. Um diese Datenmengen mit vertretbarem Zeit- und Kostenaufwand durchsuchen zu können, müssen die eingesetzten Algorithmen besonders laufzeit- und speichereffizient sein. Aufgrund dieser Anforderungen kommen zumeist etablierte Ansätze des Information Retrieval, wie Vektorraum-Modelle oder Invertierte Indexe zum Einsatz (vgl. Kapitel C 2 Modelle des Information Retrieval).

Bei *Vektorraum-Modellen* für die Plagiatsanalyse bilden zumeist lexikalische Einheiten wie Worte, Sätze oder *n*-Gramme die Dimensionen des Vektorraums. Lexikalische *n*-Gramme bezeichnen zusammenhängende Folgen von *n* Elementen, typischerweise Zeichen oder Worte, seltener Phrasen oder Sätze. Auch nicht textuelle Inhaltselemente wie Quellenverweise, grafische oder mathematische Merkmale können genutzt werden. Die Kosinus-Ähnlichkeit der Vektoren wird typischerweise als Proxy für die Ähnlichkeit der Dokumente verwendet.

Invertierte Indexe sind universelle Datenstrukturen, die vielseitig für die Plagiatsanalyse und besonders für die Kandidatensuche einsetzbar sind. Ein weitverbreitetes, Index-basiertes Verfahren zur Suche von Dokumentkandidaten mit übereinstimmenden Textteilen ist *n*-gram *Fingerprinting*. Dieser Ansatz unterteilt alle Dokumente der Referenzkollektion in *n*-Gramme, wahlweise mit oder ohne Überlappung. Eine Auswahl der gebildeten *n*-Gramme wird als „Fingerabdruck“ des jeweiligen Dokumentes in einem Index gespeichert.

Viele Plagiatsanalysesysteme nutzen für die Kandidatensuche Programmierschnittstellen von *Websuchmaschinen*, anstatt mit hohem zeitlichem und finanziellem Aufwand eigene Referenzkollektionen aufzubauen und Suchalgorithmen zu entwickeln.

4.1.2 Detailanalyse

Ziel dieser Phase ist es, die zuvor gefundenen Kandidatendokumente eingehend mit dem Eingabedokument zu vergleichen. Da während der Kandidatensuche deutlich weni-

ger Dokumente zu untersuchen sind, können rechentechnisch aufwändigere Verfahren eingesetzt werden. Die Identifikation lexikalisch ähnlichen Texts umfasst typischerweise die folgenden Schritte:

1. **Seeding:** Auffinden von Textteilen des Eingabedokuments (dem Seed) innerhalb eines Kandidatendokuments.
2. **Erweiterung:** Weitestmögliche Ausweitung des Seeds, um die vollständige Passage zu finden, die eventuell übernommen wurde. Ein populärer Ansatz für die Erweiterung ist das Zusammenführen benachbarter Seeds im Eingabe- und Kandidatendokument, wenn deren Abstand (gemessen in Zeichen) unter einem Schwellenwert liegt.
3. **Filterung:** Ausschluss von Fragmenten, die vordefinierte Kriterien nicht erfüllen (z. B. zu kurz sind) und Vereinheitlichung überlappender Passagen.

Die *Identifikation von Paraphrasen*, also von semantisch äquivalentem, aber lexikalisch unterschiedlichem Text, ist oft ein separater Schritt, für den vielfältige Verfahren der *syntaktischen* und *semantischen* Textanalyse zum Einsatz kommen.

Syntaktische Textanalyseverfahren bestimmen mittels Part-of-speech-Tagging die syntaktische Struktur von Sätzen. Die syntaktischen Strukturen helfen, morphologische Mehrdeutigkeiten aufzulösen (vgl. z. B. Hussein 2015) oder den Arbeitsaufwand für eine anschließende semantische Analyse zu reduzieren, indem z. B. nur Wortpaare mit identischen Wortarten verglichen werden (so bspw. bei Gupta, Kanjirangat & Leema L. 2016).

Semantische Textanalyseverfahren lassen sich grob in zwei Kategorien unterteilen (Gomaa & Fahmy 2013). *Wissensbasierte Ansätze* analysieren die Verbindungen zwischen Worten oder Konzepten, die in Wörterbüchern, Enzyklopädien oder Thesauri kodiert sind. *Korpusbasierte Ansätze* folgen der Idee der Verteilungssemantik, d. h., Begriffe, die in ähnlichen Kontexten vorkommen, haben tendenziell eine ähnliche Bedeutung. Umgekehrt geht die Verteilungssemantik davon aus, dass ähnliche Verteilungen von Begriffen auf semantisch ähnliche Texte hinweisen. *Word Embeddings*, *Latent Semantic Analysis (LSA)*, *Semantic Concept Analysis (SCA)* und *Neuronale Sprachmodelle* sind erfolgreiche Methoden für die semantische Textanalyse, die aus der Idee der Verteilungssemantik abgeleitet wurden. Die Methoden unterscheiden sich im Hinblick auf den Bereich, in dem sie das Vorkommen von Begriffen analysieren. Word Embeddings berücksichtigen die unmittelbar angrenzenden Worte, LSA analysiert das gesamte Dokument, SCA und neuronale Sprachmodelle nutzen umfangreiche externe Korpora. Neuere Plagiatsanalyseverfahren kombinieren oft wissensbasierte und korpusbasierte semantische Analyseverfahren durch den Einsatz von Machine Learning.

Die *Identifikation von Ideenplagiaten*[4] ist eine besondere Herausforderung für Plagiatsanalyseverfahren und Schwerpunkt aktueller Forschung. Ansätze zur Lösung dieses Problems kombinieren zumeist semantische Textanalyseverfahren mit Ähnlichkeitsbetrachtungen für weitere Dokumentinhalte. Beispielweise können im Text verwendete *Quellenverweise* auf ähnliche Muster untersucht werden, die auf eine eventuell verdächtige Ähnlichkeit der betreffenden Passagen hindeuten können (bspw. Gipp, Meuschke & Breitinger 2014; Pertile, Moreira & Rosso 2016). Gipp, Meuschke & Beel (2011) zeigten zum Beispiel, dass durch den Vergleich von Quellenverweisen 13 der 16 nachgewiesenen Übersetzungsplagiate in der Dissertation von Karl-Theodor zu Guttenberg auffindbar wa-

4 Ideenplagiate bezeichnen die Verwendung von Konzepten, Daten oder inhaltlichen Strukturen einer Quelle ohne angemessene Kennzeichnung, wobei die fremden Inhalte vollständig in eigenen Worten wiedergegeben werden.

ren. Etablierte Plagiatsanalyseverfahren, die nach lexikalischer Ähnlichkeit suchen, konnten keine dieser übersetzen Passagen finden.

Die Suche nach ähnlichen *Abbildungen* ist ein weiterer Ansatz, um mögliche Ideenplagiate zu erkennen. Bisherige Verfahren für die abbildungsbasierte Plagiatsanalyse bedienen sich überwiegend etablierter Methoden des *Content-based Image Retrieval*, um visuell ähnliche Abbildungen zu finden (bspw. Eisa, Salim & Abdelmaboud 2020; Iwanowski, Cacko & Sarwas 2016). Fortschritte auf dem Gebiet der automatisierten Chart-Analyse erlauben zum Teil die Rekonstruktion der in Diagrammen dargestellten Daten (Davila, Setlur, Doermann, Kota & Govindaraju 2021). Die rekonstruierten Rohdaten können für eine rein datenbasierte Suche nach auffälligen Ähnlichkeiten zwischen wissenschaftlichen Dokumenten genutzt werden. Meuschke et al. (2018) präsentierten bspw. einen Ansatz, der die in Balkendiagrammen dargestellten Werte rekonstruiert und für die Suche nach inhaltlich, jedoch nicht zwangsweise visuell ähnlichen Balkendiagrammen nutzt.

Auch *mathematische Inhalte* werden mittlerweile für die Plagiatsanalyse genutzt (bspw. Meuschke et al. 2017; Meuschke et al. 2019). Die Ähnlichkeitsanalyse mathematischer Inhalte kann auf den Ebenen der Präsentation (identische Symbole), des Inhalts (äquivalente Symbole oder ähnliche Struktur) und der Semantik (zugrundeliegende mathematische Konzepte) erfolgen (Guidi & Sacerdoti Coen 2016). Aktuell beschränken sich mathematikbasierte Plagiatsanalyseverfahren auf die Präsentationsebene. Der Einbezug der Inhalts- und Semantikebene ist Gegenstand aktueller Forschung.

Die Identifikation von *Übersetzungsplagiaten* erfordert sprachübergreifende Ansätze, welche ebenfalls dem in Abbildung 1 dargestellten Prozess folgen. Für die Kandidatensuche können z. B. multilinguale Wort- oder *n*-Gramm-Indexe verwendet werden, die mittels paralleler Korpora oder Verfahren der maschinellen Übersetzung (s. Kapitel B 14 Maschinelle Übersetzung) erstellt wurden (Potthast 2011; Roostaee et al. 2020). Für die Detailanalyse kommen vorrangig statistische maschinelle Übersetzung, multilinguale semantische Textanalyse, z. B. auf Basis konzeptbasierter Wissensgraphen, und multilinguale neuronale Sprachmodelle zum Einsatz (Ferrero et al. 2017; Franco-Salvador et al. 2016). Auch die bereits erläuterten Verfahren, die nicht-textuelle Inhalte untersuchen, liefern Hinweise auf mögliche Übersetzungsplagiate.

4.2 Intrinsische Plagiatsanalyse

Intrinsische Plagiatsanalyseverfahren untersuchen ein Eingabedokument auf stilistische Unterschiede im Verlauf des Textes, indem sie eine Vielzahl linguistischer Textmerkmale quantifizieren und vergleichen. Die meisten intrinsischen Analyseverfahren folgen dabei einem dreistufigen Prozess (Safin & Kuznetsova 2017) bestehend aus:

1. **Textdekomposition**
   Segmentierung des Textes in gleich große Abschnitte (z. B. Passagen, Zeichen- oder Wort-*n*-Gramme), Struktureinheiten (z. B. Absätze oder (überlappende) Sätze), thematische oder stilistische Einheiten.
2. **Konstruktion von Stilmodellen:**
   a. Analyse lexikalischer, syntaktischer und struktureller stilistischer Merkmale für jedes Textsegment, z. B. die Häufigkeiten von *n*-Grammen, Satzzeichen und Wortklassen.

   b. Berechnung quantitativer Maße, z. B. bezüglich des Wortschatzes, der Lesbarkeit und der Komplexität des Textes.
   c. Zusammenfassen der Maße zu stilistischen Merkmalsvektoren.
3. **Ausreißer-Erkennung** Klassifizierung der stilistischen Merkmalsvektoren jedes Textsegments als Mitglieder der Zielklasse, d. h. wahrscheinlich unverdächtig, oder als Ausreißer, d. h. wahrscheinlich von jemand anderem geschrieben.

Textteile mit auffälligen stilistischen Unterschieden können durch extrinsische Plagiatsüberprüfungsverfahren weiter analysiert oder menschlichen Prüfer*innen vorgelegt werden.

Intrinsische Plagiatsanalyse war im Vergleich zu extrinsischer Plagiatserkennung lange von untergeordneter Bedeutung. Der Hauptgrund hierfür ist, dass der rechtssichere Nachweis eines Fehlverhaltens bei der intrinsischen Analyse schwieriger zu führen ist. Anders als extrinsische Verfahren identifizieren intrinsische nicht unmittelbar eine mögliche Quelle für den verdächtigen Inhalt. Eine Zunahme von Contract Cheating[5] im akademischen Umfeld rückt intrinsische Verfahren seit einigen Jahren stärker in den Fokus (Ison 2020; Juola 2017). Insbesondere Auftragsarbeiten kommerzieller Anbieter*innen sind für extrinsische Plagiatserkennungssoftware nicht zugreif- und damit nicht identifizierbar. Für solche Arbeiten stellt Schreibstilanalyse oft die einzige Option für eine computergestützte Überprüfung dar.

5 Plagiatsprävention

Die Ursachen für Plagiat sind vielfältig (dazu ausführlicher Franzky et al. 2016, S. 31). Studien (insb. Sattler 2007) belegen als Einflussfaktoren für Plagiate u. a. Schreibschwierigkeiten, mangelnde Kenntnis und Unklarheit über die korrekte Anwendung der Regeln des wissenschaftlichen Arbeitens sowie ungenügende Betreuung. Neben fehlendem Fehlerbewusstsein als weitere Ursache, nimmt auch der Aspekt Überforderung durch Zeitdruck, Prüfungsangst und fehlende moralische Grundüberzeugung eine bedeutende Rolle ein. Mehrheitlich erweisen sich Plagiatsfälle aber nicht als vorsätzlich herbeigeführt.

Um solche Fälle zu vermeiden, ehe sie entstehen, erscheinen als wirksame Maßnahmen „eine qualitativ hochwertige Ausbildung, ein vertrauensvolles Betreuungsverhältnis und das gelebte Vorbild“ (Glatzmeier 2019, S. 31), neben der abschreckenden Wirkung durch die Erhöhung der Entdeckungswahrscheinlichkeit (Sattler 2007, S. 194) und das konsequente Durchsetzen von Sanktionen (Sattler 2007, S. 199).

Einige Hochschulen begegnen Plagiaten präventiv durch die Etablierung von Unterstützungsangeboten zur Förderung der Methodenkompetenz im Rahmen des wissenschaftlichen Arbeitens (von der Literaturrecherche und der kritischen Lektüre über das Erstellen eines Arbeitsplans und der Inhaltsgliederung bis zum Verfassen und richtigen Zitieren) sowie durch forcierte Bewusstseinsbildung und klare Kommunikation des Umgangs mit Plagiaten. „Die Vermittlung wissenschaftlicher Arbeitstechniken [ist] primär [...] Aufgabe der Fachlehre.“ (Rotzal & Schuh 2016, S. 65) Sie kann sich dabei aber gut ergänzenden Unterstützungsangeboten in Form von Schulungen, Diskussionsveranstaltungen sowie frei nachnutzbaren Lernmaterialien bedienen, die insbesondere mit Hilfe

5 Der Begriff bezeichnet die Beauftragung von Dritten mit der Anfertigung von Studienleistungen, wie z. B. Haus-, Seminar- und Abschlussarbeiten oder Übungsaufgaben.

von Drittmittelförderungen in wissenschaftlichen Bibliotheken aufgebaut und gepflegt werden. „Inhaltlich gesehen leiste[n] [...] Bibliothek[n] [damit] einen Beitrag zur Grundlagenlehre." (Rotzal & Schuh 2016, S. 67) Bibliotheken erschließen sich mit derartigen Projekten im Rahmen der Informationskompetenzvermittlung ein weiteres Handlungsfeld. Das Projekt Akademische Integrität (Johannes Gutenberg-Universität Mainz 2021) mit Fokus auf Präventionsarbeit zielt darauf ab, Standards der Wissenschaft besser sichtbar zu machen und im stetigen Dialog das Problembewusstsein über Verstöße gegen die gute wissenschaftliche Praxis, hier primär mit Blick auf das Plagiat, möglichst früh in der akademischen Ausbildung oder gar in der Schule zu verankern. Das Projekt Refairenz (Universität Konstanz 2021) nimmt das Plagiat in Form von intertextuellen Fehlern sowie Ursachen dafür in den Blick. Als effektivste Maßnahme gilt auch hier die Präventionsarbeit. Ziel des Projekts PlagStop.nrw (Digitale Hochschule NRW 2021), ist die Vermeidung von Plagiaten durch den rechtssicheren und optimierten Einsatz von Plagiatserkennungssoftware im Kontext von Learning-Management-Systemen an den Hochschulen in Kombination mit Präventionsmaßnahmen in Form frei nachnutzbarer Selbstlernmodule. Ein zu beauftragendes Rechtsgutachten soll daneben Klarheit schaffen, ob Aufbau und Einsatz einer landesweiten Datenbank zum Abgleich von Textähnlichkeiten in schriftlichen studentischen Arbeiten rechtlich möglich sind.

Wissenschaftliche Bibliotheken als wichtige Akteure der Projekte bauen damit ihre Rolle als zentrale Informationskompetenzvermittlerin in den Hochschulen aus. Ihr Engagement im Bereich der Plagiatsdetektion und -prävention darf aber nicht dahingehend interpretiert werden, dass sie eine

> Kontrollfunktion über die Richtigkeit bzw. Wahrheit [...] wissenschaftliche[r] Aussage[n] übernehmen [möchten]. Dieses Regulativ bleibt der wissenschaftlichen Selbstkontrolle und damit dem Wissenschaftssystem überantwortet, wo Wissenschaftlichkeit systemimmanent verhandelt wird (Brandtner 2014, S. 37).

6 Fazit

Plagiate sind eine schwere Form akademischen Fehlverhaltens und ein ernstzunehmendes Problem für Bildungs- und Forschungseinrichtungen, Verlage und Fördermittelgeber. Plagiatserkennungssoftware wird für Prüfende zunehmend wichtiger, da die immense Anzahl international verfügbarer wissenschaftlicher Texte im Wege einer manuellen Prüfung nicht mehr erfassbar ist. Die zunehmend freie Verfügbarkeit von Texten und Quellen entpuppt sich dabei eher als Segen denn als Fluch. Wenngleich frei verfügbare wissenschaftliche Texte durch ihre mühelose Kopierbarkeit das Plagiieren erleichtern (Weber-Wulff 2010, S. 57), sind sie auch im Rahmen der Plagiatserkennung einfacher identifizierbar als Inhalte hinter Bezahlschranken.

Andererseits kann und wird Erkennungstechnologie allein die Problematik wissenschaftlicher Plagiate nicht lösen können. Plagiatserkennungssoftware kann nur auf inhaltliche Ähnlichkeiten des überprüften Dokumentes zu den zugreifbaren Quellen hinweisen. Die Beurteilung, ob erkannte Ähnlichkeiten ein Plagiat darstellen, erfordert immer einer Prüfung durch den Menschen. Aktuelle Software beschränkt sich zudem oft noch auf die Suche nach identischem Text. Dadurch stellt die Erkennung von Paraphrasen, Ideen- und Übersetzungsplagiaten viele aktuelle Softwarelösungen vor große Probleme. Neuartige Systeme wie HyPlag – Hybride Plagiatserkennung (Meuschke et al. 2018) adressieren diese Schwäche, indem sie neben semantischer Textähnlichkeit auch

Abbildungen, Grafiken, Formeln und Quellenverweise überprüfen. Doch auch technisch verbesserte Software adressiert überwiegend die Symptome und nicht die zugrundeliegenden Ursachen des Problems Plagiat. Letztere sind mehrheitlich mangelnde Kenntnisse, fehlendes Problembewusstsein und Überforderung als vorsätzliches Fehlverhalten und sollten durch präventive Maßnahme, wie Vermittlung von Methodenkompetenz adressiert werden. Eine umfassende und effektive Behandlung des Problems erfordert das koordinierte Zusammenspiel von technischen und nicht-technischen Maßnahmen der Plagiatsprävention, -erkennung und -sanktion.

7 Literaturverzeichnis

Brandtner, A. (2014). Auf den Schultern von Bibliotheken/On the shoulders of libraries/Sur les épaules des bibliothèques. *Information – Wissenschaft & Praxis*, 65(1). doi: 10.1515/iwp-2014-0013.

Dagli-Yalcinkaya, L. (2021). *PlagStop.nrw: Abschlussbericht*. Hochschule Niederrhein. https://www.dh.nrw/kooperationen/PlagStop.nrw-46.

Davila, K., Setlur, S., Doermann, D., Kota, B. U. & Govindaraju, V. (2021). Chart Mining: A Survey of Methods for Automated Chart Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11), 3799–3819. doi: 10.1109/tpami.2020.2992028.

DFG (Hrsg.) (2013). *Denkschrift: Sicherung Guter Wissenschaftlicher Praxis*. Wiley-VCH. doi: 10.1002/9783527679188.

DFG (2019). *Kodex: Leitlinien zur Sicherung guter wissenschaftlicher Praxis*. doi: 10.5281/zenodo.3923602.

Digitale Hochschule NRW (2021). Plagtop.nrw. https://www.dh.nrw/kooperationen/PlagStop.nrw-46.

Dissernet (2021). *Dissernet Chronicles*. https://dissernet.org/acat_chronicle/.

Eisa, T. A. E., Salim, N. & Abdelmaboud, A. (2020). Content-Based Scientific Figure Plagiarism Detection Using Semantic Mapping. In F. Saeed, F. Mohammed & N. Gazem (Hrsg.), *Emerging Trends in Intelligent Computing and Informatics* (Bd. 1073) (S. 420–427). Springer. doi: 10.1007/978-3-030-33582-3_40.

Ferrero, J., Besacier, L., Schwab, D. & Agnès, F. (2017). Deep Investigation of Cross-Language Plagiarism Detection Methods. *Proceedings 10th Workshop on Building and Using Comparable Corpora (BUCC)* (S. 6–15). ACL. doi: 10.18653/v1/w17-2502.

Fishman, T. (2009). „We Know It When We See It" Is Not Good Enough: Toward a Standard Definition of Plagiarism That Transcends Theft, Fraud, and Copyright. *Proceedings 4th Asia Pacific Conference on Educational Integrity* (S. 1–5). University of Wollongong. https://ro.uow.edu.au/apcei/09/papers/37/.

Foltýnek, T., Meuschke, N. & Gipp, B. (2019). Academic Plagiarism Detection: A Systematic Literature Review. *ACM Computing Surveys*, 52(6), 112:1–112:42. doi: 10.1145/3345317.

Franco-Salvador, M., Gupta, P., Rosso, P. & Banchs, R. E. (2016). Cross-Language Plagiarism Detection Over Continuous-Space- and Knowledge Graph-Based Representations of Language. *Knowledge-Based Systems*, 111, 87–99. doi: 10.1016/j.knosys.2016.08.004.

Franzky, T., Hätscher, P., Kohl, K. E., Krämer, S., Nunnenmacher, U., Münzinger, J. & Trevisiol, O. (2016). *Plagiate verhindern: Ursachen kennen, Lehre gestalten, mit Fällen umgehen*. Präsentation Tagesworkshop für Lehrende. https://www.plagiatspraevention.uni-konstanz.de/lehrmaterial/dozentenmaterial/.

Gipp, B., Meuschke, N. & Beel, J. (2011). Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag. *Proceedings ACM/IEEE Joint Conference on Digital Libraries* (S. 255–258). doi: 10.1145/1998076.1998124.

Gipp, B., Meuschke, N. & Breitinger, C. (2014). Citation-based Plagiarism Detection: Practicability on a Large-Scale Scientific Corpus. *Journal of the Association for Information Science and Technology*, 65(8), 1527–1540. doi: 10.1002/asi.23228.

Glatzmeier, A. (2019). Öffentlicher Diskurs – Ursachen – Strategien: Gute wissenschaftliche Praxis als Herausforderung – nicht nur für die Fachöffentlichkeiten. In A. Geukes (Hrsg.), *Konferenzband uni.digital 2019: Teaching, assessment, learning* (S. 26–39). Freie Universität Berlin. doi: 10.17169/refubium-26641.

Gomaa, W. H. & Fahmy, A. A. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, 68(13), 13–18. doi: 10.5120/11638-7118.

Guidi, F. & Sacerdoti Coen, C. (2016). A Survey on Retrieval of Mathematical Knowledge. *Mathematics in Computer Science*, 10(4), 409–427. doi: 10.1007/s11786-016-0274-0.

Gupta, D., Kanjirangat, V. & Leema L., M. (2016). Plagiarism Detection in Text Documents Using Sentence Bounded Stop Word N-Grams. *Journal of Engineering Science and Technology*, 11(10), 1403–1420. http://jestec.taylors.edu.my/Vol 11 issue 10 October 2016/11_10_4.pdf.

Hussein, A. S. (2015). A Plagiarism Detection System for Arabic Documents. *Proceedings International Conference on Intelligent System. AISC 323* (S. 541–552). Springer. doi: 10.1007/978-3-319-11310-4_47.

Ison, D. (2020). Detection of Online Contract Cheating Through Stylometry: A Pilot Study. *Online Learning*, 24(2). doi: 10.24059/olj.v24i2.2096.

Iwanowski, M., Cacko, A. & Sarwas, G. (2016). Comparing Images for Document Plagiarism Detection. In L. J. Chmielewski, A. Datta, R. Kozera & K. Wojciechowski (Hrsg.), *Computer Vision and Graphics* Bd. 9972 (S. 532–543). Springer. doi: 10.1007/978-3-319-46418-3_47.

Johannes Gutenberg-Universität Mainz (2021). Projekt „Akademische Integrität". https://www.akin.uni-mainz.de/.

Juola, P. (2017). Detecting Contract Cheating Via Stylometric Methods. In I. Glendinning, T. Foltýnek & J. Rybička (Hrsg.), *Proceedings Plagiarism Across Europe and Beyond Conference* (S. 187–198). Mendel University. http://academicintegrity.eu/conference/proceedings/2017/Juola_Detecting.pdf.

Meuschke, N., Gondek, C., Seebacher, D., Breitinger, C., Keim, D. & Gipp, B. (2018). An Adaptive Image-Based Plagiarism Detection Approach. *Proceedings ACM/IEEE Joint Conference on Digital Libraries* (S. 131–140). Fort Worth, USA. doi: 10.1145/3197026.3197042.

Meuschke, N., Schubotz, M., Hamborg, F., Skopal, T. & Gipp, B. (2017). Analyzing Mathematical Content to Detect Academic Plagiarism. *Proceedings ACM Conference on Information and Knowledge Management*. (S. 2211–2214). ACM. doi: 10.1145/3132847.3133144.

Meuschke, N., Stange, V., Schubotz, M. & Gipp, B. (2018). HyPlag: A Hybrid Approach to Academic Plagiarism Detection. *Proceedings 41st ACM SIGIR Conference* (S. 1321–1324). doi: 10.1145/3209978.3210177.

Meuschke, N., Stange, V., Schubotz, M., Kramer, M. & Gipp, B. (2019). Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations. *Proceedings ACM/IEEE Joint Conference on Digital Libraries* (S. 120–129). doi: 10.1109/jcdl.2019.00026.

Pertile, S. de L., Moreira, V. P. & Rosso, P. (2016). Comparing and Combining Content- and Citation-Based Approaches for Plagiarism Detection. *Journal of the Association for Information Science and Technology*, 67(10), 2511–2526. doi: 10.1002/asi.23593.

Potthast, M., Barrón-Cedeño, A., Stein, B. & Rosso, P. (2011). Cross-language Plagiarism Detection. *Language Resources and Evaluation*, 45(1), 45–62. doi: 10.1007/s10579-009-9114-z.

Retraction Watch (2021). *Tracking retractions as a window into the scientific process.* https://retractionwatch.com/.

Roostaee, M., Sadreddini, M. H. & Fakhrahmad, S. M. (2020). An effective approach to candidate retrieval for cross-language plagiarism detection: A fusion of conceptual and keyword-based schemes. *Information Processing & Management*, 57(2), 102150: 1-19. doi: 10.1016/j.ipm.2019.102150.

Rotzal, T. & Schuh, D. (2016). Grundlagenlehre: Bibliotheken als Vermittler wissenschaftlicher Arbeitstechniken, Werte und Normen. *o-bib. Das offene Bibliotheksjournal*, 61–74. doi: 10.5282/o-bib/2016h4s61-74.

Safin, K. & Kuznetsova, R. (2017). Style Breach Detection with Neural Sentence Embeddings. In L. Cappellato, N. Ferro, L. Goeuriot & T. Mandl (Hrsg.), *Working Notes of the Conference and Labs of the Evaluation Forum*. http://ceur-ws.org/Vol-1866/paper_69.pdf.

Sattler, S. (2007). *Plagiate in Hausarbeiten: Erklärungsmodelle mit Hilfe der Rational Choice Theorie*. Dr. Kovač.

Theisohn, P. (2009). *Plagiat: Eine unoriginelle Literaturgeschichte*. A. Kröner.

Universität Konstanz. (2021). Projekt Refairenz. https://www.plagiatspraevention.uni-konstanz.de.
VroniPlag Wiki. (2021). Übersicht. https://vroniplag.fandom.com/.
Weber-Wulff, D. (2010). Unter Schizophrenen: Plagiate bekämpfen mit Open Access. *Neue Gesellschaft / Frankfurter Hefte*, 12, 57–59. https://www.frankfurter-hefte.de/artikel/plagiate-bekaempfen-mit-open-access-788/.
Weber-Wulff, D. (2014). *False Feathers: A Perspective on Academic Plagiarism*. Springer. doi: 10.1007/978-3-642-39961-9.