h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

Hochschule Darmstadt

Fachbereiche Mathematik und
Naturwissenschaften & Informatik

Zurück in die Zukunft:
Named-Entity-Recognition für digitale (historisch-kritische) Editionen

Abschlussarbeit zur Erlangung des akademischen Grades

Master of Science (M. Sc.)
im Studiengang Data Science

vorgelegt von

Tamara Haeder

Referentin	:	Prof. Dr. Uta Störl
Korreferentin	:	Prof. Dr. Jutta Groos

Ausgabedatum	:	01.05.2020
Abgabedatum	:	13.11.2020

ERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Heidelberg, 13.11.2020

Tamara Haeder

ABSTRACT

Most of the research on Natural Language Processing in Digital Humanities has focused on difficulties caused by diachronic data. Other fields can cause significant drops in performance of Natural Language techniques, however. Techniques that model sequence labelling problems often build on probability distributions of words and corresponding labels. Problems e.g. for Named-Entity-Recognition Systems may occur if data is not annotated consistently. It is not possible anymore to unambiguously identify the correct entity for inconsistently annotated data. Besides of a consistent annotation scheme within a single digital (scholarly) edition, it is important to develop and maintain standards across editions to generate adaptable models.

ZUSAMMENFASSUNG

Die Mehrheit der Natural Language Processing Untersuchungen im Rahmen der Digital Humanities betonen die Schwierigkeiten, welche sich durch diachronische Daten ergeben. Dabei können andere Bereiche bedeutende Verluste der Performanz von Natural Language Processing Verfahren verursachen. Verfahren, die Sequenzlabelling Probleme modellieren, basieren oft auf Wahrscheinlichkeitsverteilungen der Worte und ihrer Label. Wenn diese nicht konsequent konsistent annotiert werden, können zum Beispiel Named-Entity-Recognition Verfahren keine eindeutige Zuordnung der Worte zu ihren korrespondierenden Entitäten vornehmen. Neben einem einheitlichen Annotationsschema innerhalb einer digitalen (historisch-kritischen) Edition, ist die Etablierung von allgemeinen Standards wichtig, damit adaptierbare Modelle generiert werden können.

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

Teil I

THESIS

1 EINLEITUNG

1.1 MOTIVATION

Digitale historisch-kritische Editionen werden immer populärer. Vor langer Zeit hätte die Bezeichnung noch für Verwunderung gesorgt, da zeitens die Verwendung von „digital“ und „historisch“ in einem Atemzug als Oxymoron, als zwei sich ausschließende Begriffe, galt. Mit zunehmender Digitalisierung wurde auch das Potential dieser in den Literatur- und Gesellschaftswissenschaften erkannt. Es entstand eine neue Disziplin, die der Digital Humanities, wodurch die Vereinbarkeit digitaler und historischer Konzepte geschaffen wurde. Die reine Transformation von Papier-Dokumenten in computerlesbare Form gilt als Digitalisierung. Moderne Technologien gehen einen Schritt weiter und erweitern die Optionen zur Datenverarbeitung,-analyse und -darstellung. John Smith erkannte schon vor über 40 Jahren die Möglichkeiten von computergestützten Applikationen für Sprach- und Literaturwissenschaften und unterteilte diese in zwei Gruppen: Textuelle Manipulation (Indexierung, Bilden von Konkordanzen, etc.) und tatsächliche thematische und stilistische Analysen.[30] Machan führte 1991 die Begriffe „lower criticism“ und „higher criticism“ im Rahmen der kritischen Auseinandersetzung mit Literatur ein. Die Konzepte von Smith und Machan ähneln sich sehr stark.[24] Zu Beginn wurden sowohl Aufgaben, die unter „lower criticism“ als auch Aufgaben, die unter „higher criticism“ fallen, manuell durchgeführt. Durch spezielle Algorithmen ist nun eine effizientere Ausführung von Smiths Vision der computergestützten textuellen und thematischen Analyse von Literatur möglich.

Die reine Digitalisierung von Texten erweitert das Potential im Umgang mit Texten. Sie erleichtert den Zugang zu diesen und erreicht somit eine breite Masse an Nutzer*innen. Dies erhöht einerseits den Anspruch an die Qualität der Textdarstellung, da verschiedene Nutzer*innen unterschiedliche Ansprüche stellen. Andererseits können kritische, vielfältige Blickwinkel die Entwickler*innen digitaler (historischer-kritischer) Editionen antreiben, kreative und fortschrittliche Ansätze zu verfolgen, um die Erwartungen bestmöglich zu erfüllen.

Neben Nutzer*innen gibt es eine weitere Personengruppe, die mit digitalen (historisch-kritischen) Editionen in Verbindung gebracht wird: Editor*innen. Sie sind maßgebend für die Qualität der Editionen und bilden die Grundlage für die Transformation von einer digitalisierten Druckedition in eine digitale (historisch-kritische) Edition. Erst durch ihre kritische Auseinandersetzung mit den Dokumenten wird eine digitale Edition zu einer digitalen historisch-kritischen Edition. Sie schlagen die Brücke zwischen Vergangenheit und Gegenwart.

Um die digitalen (historisch-kritischen) Editionen der Zukunft aufzubauen, werden Computerwissenschaftler*innen im Aufbau hinzugezogen. Durch Machine Learning Verfahren, insbesondere im Bereich Data Mining und Natural Language Processing, kann der Arbeitsprozess von Editor*innen effizienter gestaltet werden. Obwohl durch die Transkription von historischen Dokumenten *„Big Data of the Past"*[22] entsteht, wurden Machine Learning Algorithmen bisher selten in den Editionsprozess eingebaut.[31] Vor allem Data Mining und Natural Langauge Processing Techniken können die Editor*innen effizient bei ihrer Arbeit unterstützen.

Die im Rahmen dieser Arbeit durchgeführt Bedarfserhebung zeigt die Nachfrage nach Unterstützungstools. Editor*innen sehen insbesondere bei der Annotation von Entitäten Potential, diesen Prozess durch Named-Entity-Recognition Systeme zu verbessern.

Named-Entity-Recognition Verfahren wurden bereits erfolgreich auf zeitgenössischen Zeitungsartikeln trainiert.[25] Allerdings lassen diese sich nicht ohne Probleme auf historische Texte übertragen.[35] Hierbei erhöhen OCR-Fehler, nicht-standardisierte Rechtschreibung und domänenspezifische Textarten den Fehler bestehender Systeme. Es gibt zwei Ansätze, wie damit umgegangen werden kann: Entweder werden bestehende Systeme erweitert, sodass sie adaptierbar werden oder es werden neue Systeme für historische Daten from scratch erstellt.

Die vorliegende Arbeit verfolgt letzteren Ansatz. Anhand der Dokumente der Frank Wedekind Briefedition wird ein Named-Entity-Recognition Modell zur automatisierten Annotation von Personen, Orten, Örtlichkeiten, Ereignissen und Werken erstellt. In vielen Editionen, so auch in der Frank Wedekind Edition, dienen die Annotationen zur gezielten Suche nach bestimmten Entitätsreferenzen. Zusätzlich können sie Teil einer Facetten-Suche sein, welche die Ergebnisse nach gewünschten Entitäten und spezifischen Entitätsreferenzen filtert.

Zusammenfassend unterstützt ein Named-Entity-Recognition System nicht nur Editor*innen im Editionsprozess, sondern erweitert zusätzlich die Suchfunktionalitäten für Nutzer*innen.

1.2 AUFBAU

Zu Beginn der Arbeit werden in Kapitel 2 die theoretischen Grundlagen geschaffen. Hierbei wird zuerst in Abschnitt 2.1 eine Definition zu digitalen historisch-kritischen Editionen vorgenommen, indem der Begriff in seinen einzelnen Bestandteilen erklärt wird. Abschnitt 2.2 ordnet die Rolle von Editor*innen und Computerwissenschaftler*innen in den Kontext ein. Danach wird in Abschnitt 2.3 die Sichtweise der Computerwissenschaftler*innen auf digitale (historisch-kritische) Editionen in den Fokus gerückt und eine Analogie zu Information Retrieval Systemen gezogen. Eng damit verbunden ist der Begriff der Information Extraction, durch den eine Überleitung zur Beschreibung von Named-Entity-Recognition in Abschnitt 2.4 gefunden wird. Nach einer kurzen Einleitung werden verschiedene Verfahren für Named-

Entity-Recognition Aufgaben vorgestellt. Da sich dafür im Laufe der Jahre Machine Learning Ansätze gegenüber manuellen Ansätzen bewährt haben, wird der Fokus auf Erstere gelegt. Abschließend werden am Ende des Kapitels die gängigen Evaluationsmaße definiert.

Kapitel 3 gibt einen Überblick des aktuellen Forschungsstandes. Einerseits werden Arbeiten zu allgemeinen Natural Language Processing Themen in Bezug auf digitale (historsiche-Editionen) vorgestellt und andererseits werden Arbeiten speziell zu Named-Entity-Recognition Aufgaben diskutiert.

Um einen fundierte Basis für die Begründung der Implementierung von Machine Learning Verfahren zu schaffen, wurden Editor*innen und Nutzer*innen zu ihren Anforderungen und Bedürfnissen befragt. Die Ergebnisse werden in Kapitel 4 präsentiert.

In Kapitel 5 wird der praktische Teil der Arbeit vorgestellt. Zuerst werden in den Abschnitten 5.1 und 5.2 der Datensatz und die notwendigen Datenaufbereitungsschritte erklärt. Danach wird in Abschnitt 5.3 begründet, wieso das Conditional Random Field Verfahren ausgewählt wurde, um ein Named-Entity-Recognition Modell zu erstellen und welche Feature dafür verwendet wurden. In Abschnitt 5.4 werden die Ergebnisse des Modells präsentiert und kritisch diskutiert.

Im abschließenden Kapitel 6 werden in Abschnitt 6.1 die wichtigsten Punkte und Erkenntnisse zusammengefasst und in Abschnitt **??** werden Anreize für weitere Natural Language Processing Verfahren in digitalen (historisch-kritischen) Editionen gesetzt.

2

GRUNDLAGEN

Das Grundlagenkapitel beginnt mit einer Definition von digitalen (historisch-kritischen) Editionen (Abschnitt 2.1). Da der Fokus auf der Arbeit von Editor*innen digitaler (historisch-kritischer) Editionen liegt, wird in Abschnitt 2.2 Bezug zu ihrem Arbeitsprozess genommen. Abschnitt 2.3 liefert Definitionen zu Information Retrieval- und Information Extraction- Systemen. Nach den Definitionen wird eine Verbindung zwischen den Konzepten von digitalen (historisch-kritischen) Editionen und Information Retrieval- und Information Extraction- Systemen hergestellt. In Abschnitt 2.4 wird eine kurze Einleitung zu Named Entity Recognition Aufgaben gegeben und nachfolgend werden verschiedene Methoden für solche Aufgaben diskutiert. In diesem Abschnitt wird der Fokus auf Machine Learning Verfahren gesetzt (Abschnitt 2.4.2), da im Rahmen dieser Arbeit ein Conditional Random Field implementiert wird (Kapitel 5, Abschnitt 5.3). Am Ende des Kapitels (Abschnitt 2.4.3) werden die gängigen Evaluationsmaße von IR-Systemen erklärt.

2.1 DEFINITION DIGITALER HISTORISCH-KRITISCHER EDITIONEN

Als eines der Gründungsmitglieder des Instituts für Dokumentologie und Editorik prägt Patrick Sahle die formale Definition von digitalen historisch-kritischen Editionen maßgebend:

> „Scholarly digital editions are scholarly editions that are guided by a digital paradigm in their theory, method and pratice.“[28]

Um diese Definition im Ganzen zu verstehen, muss sie in ihre einzelnen Komponenten zerlegt werden. Digitale historisch-kritische Editionen bauen auf historisch-kritischen Editionen auf. Deshalb ist zuvor zu klären, was historisch-kritische Editionen sind. Sahle beschreibt diese als „kritische Wiedergabe historischer Dokumente“. Unter Dokumenten werden sehr allgemein gehalten nicht-abstrakte Objekte verstanden, welche die Edition zum Thema haben. Diese können z. B. in Form von Texten, Arbeiten, Bilder u.a. vorliegen. Die originalgetreue Wiedergabe eines Dokumentes findet entweder in Form von Bildern, sog. Faksimiles, oder durch Transkriptionen statt. Eine reine Wiedergabe ist nicht ausreichend, sondern sie muss in einer *kritischen* Repräsentation passieren. Je nach Art der Edition variiert die Auslegung des Begriffs „kritisch“. Allen Editionen gemein ist ein differenzierter, reflektierender Prozess, mit dessen Hilfe das Verständnis des Dokumentes an Zugänglichkeit gewinnen soll. Dieser reicht vom Aufstellen von Regeln zur Transkription bis hin zur Identifikation und Benennung von Named Entities. Da es sich um historische Dokumente handelt, sind diese Leser*innen/ Nutzer*innen nicht intuitiv. Um die zeitliche Distanz zwischen historischen

Dokumenten und modernen Nutzer*innen zu überwinden, ist eine kritische Auseinandersetzung essentiell.[28]

Bis dahin sind die Kriterien für eine historisch-kritische Edition definiert. Was aber macht eine historisch-kritischen Edition zu einer *digitalen* historisch-kritischen Edition? Im Laufe der Jahre hat sich neben der bis dahin nur als „Edition" bezeichneten gedruckten Edition auch die digitale Edition entwickelt. Der Unterschied liegt im digitalen Paradigma. Zusätzliche Funktionalitäten wie Zugänglichkeit, Suchbarkeit und Nutzbarkeit erweitern die Edition und machen sie zur digitalen Edition. Hauptcharakteristik einer digitalen Edition ist die Repräsentation einer theoretisch großen Anzahl an Dokumenten durch eine theoretisch unendliche Anzahl verschiedener Ansichten wie z. B. Faksimiles, diplomatischer Transkriptionen oder Leseansichten. Hinzu kommt die Möglichkeit zur Interaktion mit und Kontrolle über die Edition durch Suchfunktionen, Hyperlinks und anderen technischen Tools.[28] Diese Funktionen gilt es zu optimieren, da eine schlechte Funktionsweise dieser keinen Vorteil ggü. Druckversionen mehr liefert. Wenn die Funktionen einer digitalen Edition nicht gewinnbringend sind, ist die digitale Edition vergleichbar mit einer digitalisierten Edition, welche das Ergebnis einer Digitalisierung einer Druckversion ist.[31]

Durch digitale Editionen wird nicht nur das Originaldokument betrachtet, sondern es wird zusätzlich mit Wissen und Zusatzmaterial durch die Editor*innen bereichert. Editor*innen nehmen also eine essentielle Rolle in der Erstellung einer digitalen Edition ein und tragen zur Abgrenzung von digitalisierten Druckeditionen bei.

2.2 EDITIONSPROZESS

In der Entwicklung von digitalen (historisch-kritischen Editionen) können Editor*innen durch Editierungstools unterstützt werden. Spadini bietet hierfür eine Vergleichsanalyse verschiedener Tools.[31] Die Vorauswahl basiert auf einem hohen Maß an Nutzerfreundlichkeit. Die Tools werden mit Hinblick auf die Integration von Editionsprozessen, Annotationen, Markups, Import und Export, Encodierung des Textes, Suchfunktionalitäten, Bildmanagement und Kollaborationsmöglichkeiten analysiert.

Spadini separiert diese Charakteristiken von Natural Language Processing und Data Mining Techniken. Möglicherweise liegt der Grund dafür darin, dass nur wenige Editionstools solche Techniken bisher implementiert haben. Spadini geht einen Schritt weiter und sagt, dass sogar die wenigsten digitalen (historisch-kritischen) Editionen von Natural Language Processing Methoden Gebrauch machen. Gleichzeitig hebt sie den Nutzen solcher Methoden hervor. Durch die Digitalisierung von Text werden neue Analyseformen, wie statistische Analysen, Visualisierungen und linguistische Annotationen möglich. Gründe für die fehlende Integration sieht sie einerseits darin, dass digitale (historisch-kritische) Editionen lange Zeit nur als Nebenprojekt in historischen und literarischen Studien angesehen wurden. Der Hauptgrund

liegt ihrer Ansicht nach allerdings im fehlenden Bewusstsein für das Potential, das in Daten steckt.

Editor*innen leisten einen bedeutenden Beitrag zur Qualität von digitalen (historisch-kritischen) Editionen. Ihre Arbeit ist eine der traditionsreichsten in den Digital Humanities. Ziel der vorliegenden Arbeit ist es, einen neuen Blickwinkel zu geben, das Potential von Daten aufzuzeigen und praktische Einblicke in die Möglichkeiten zu gewähren, wie Editor*innen bei ihrer Arbeit unterstützt werden können.

Tradition und Moderne scheinen Gegenpole zu sein. Um diese Kompetenzen miteinander zu vereinbaren, wurde DiXiT als eines der ersten Ausbildungsnetzwerke für digitales (historisch-kritisches) Editieren gegründet: „*The production of the digital edition has seen editors transforming into encoders and, even programmers. As a consequence, we have to rethink the training, skills and knowledge needed for new editors.*“[8] Bis allerdings die Generation der „new editors“, die Kenntnisse der Digital Humanities und der Computerwissenschaften in sich vereinen, ausgebildet ist, muss eine Zusammenarbeit von „alten Editor*innen“ und Computerwissenschaftler*innen stattfinden. Die drei Haupthemen, mit denen sich DiXiT beschäftigt, sind: Theorie, Praxis, Methoden (WP1), Technologie, Standardsoftware (WP2) und Academia, Kulturelles Erbe, Gesellschaft (WP3).[4] Durch die stetig wachsende Anzahl neuer Technologien, ist WP2 ein sich schnell verändernder Bereich, der ein hohes Maß an Adaptivität verlangt. DiXiT legt die Prioritäten auf die Integration von webbasierten Tools in die TEI-Umgebung und auf die Integration von Tools zur Textanalyse, zu Stylometrie und zu Visualisierungen.[4] Die vorliegende Arbeit ist aufgrund des Fokus auf analytische Tools zur Textanalyse in WP2, DiXiT einzuordnen.

2.3 INFORMATION RETRIEVAL AND EXTRACTION

Aus der Sicht von Computerwissenschaftler*innen stellen digitale (historisch-kritische) Editionen Information Retrieval Systeme dar. Information Retrieval (IR) beschreibt den Prozess der Generation von Informationen aus Daten. *Daten* sind in ihrer reinen Form maschinenlesbare Zahlen. Wie sie abgespeichert werden, wird durch eine Syntax festgelegt. Wenn Daten eine semantische Bedeutung erhalten, wird von *Wissen* gesprochen. Trägt das Wissen zur Beantwortung spezifischer Fragen oder zur Lösung spezifischer Probleme bei, wird die verwendete Teilmenge des Wissens als *Information* bezeichnet. Der Begriff Information ist also sehr nutzer*innenzentriert, da die Ausgangsbasis eine Frage eines Nutzers/ einer Nutzerin ist. Aufgabe von IR-Systemen ist also ein optimaler Transfer von Daten in Wissen und darauf folgend die Extraktion von Informationen aus dem Wissen.[18]

Anhand der Definition von digitalen historisch-kritischen Editionen aus Abschnitt 2.1 lässt sich ableiten, dass digitale historisch-kritische Editionen Information Retrieval Systeme sind. Die Abspeicherung der Urdokumente in maschinenlesbarer Form repräsentieren Daten. Dieser Prozess ist Teil der Transkription. Die Darstellung der Daten und somit die Generierung von

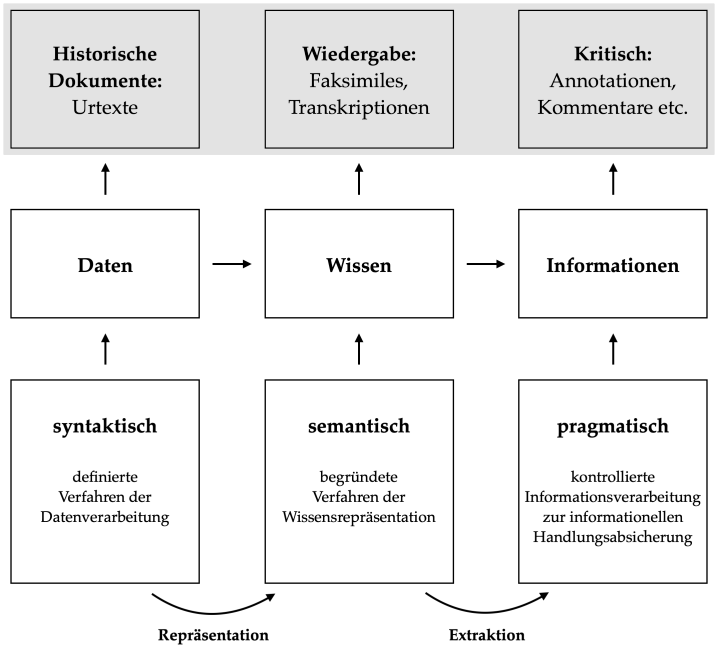


Abbildung 2.1: Digitale (historisch-kritische) Editionen als IR-System (Quelle:[18], S.20). Der grau hinterlegte Abschnitt, stellt eine Erweiterung dieser Graphik dar.

Wissen findet in digitalen historisch-kritischen Editionen als Faksimiles oder durch verschiedene Textansichten statt. Im letzten Schritt wird aus dem Wissen Information generiert. Dies geschieht durch die kritische Arbeit der Editor*innen z. B. in Form von Annotationen. In Abbildung 2.1 findet sich eine Darstellung des Zusammenhangs von digitalen historisch-kritischen Editionen. Sie basiert auf der graphischen Darstellung der festgelegten Typologie von IR-Systemen durch die deutsche Informationswissenschaft aus dem Buch Information Retrieval 1.[18] Der grau hinterlegte Bereich ist die Erweiterung der Graphik.

Praktischer gesprochen liefern IR-Systeme gegeben einer Suchanfrage ein relevantes Subset an Dokumenten aus einer Dokumentenkollektion. Weiterhin kann das System unterstützende Funktionen (Relevanzsortierung, Filterung etc.) einbauen, die dem/der Nutzer*in hilft, die von ihm/ihr gewünschte Information zu finden.[14] Neben Information Retrieval gibt es noch einen weiteren Begriff, der oft damit in Zusammenhang gebracht wird: Information Extraction. Information Extraction beschreibt die Identifikation und Benennung vordefinierter Informationen aus Texten. Allgemeiner gesprochen erstellen IE-Systeme strukturierte Informationen aus unstrukturierten Quellen.[14] IR-Systeme liefern relevante Dokumente aus einer Dokumentenkollektion und IE-Systeme extrahieren relevante Informationen aus einem Dokument. Diese Informationen können wiederum dem IR-System helfen, relevante Dokumente zu finden.[14] Folglich sind IE Methoden auch für digitale (historisch-kritische) Editionen von Nutzen.

tag	enamex (Name)	timex (Zeit)	numex (Zahlen)
	Ort	Zeit	Geld
type	Person	Datum	Prozent
	Organisation		

Tabelle 2.1: Konventionelle Entitäten

2.4 NAMED ENTITY RECOGNITION

Named Entity Recognition (auch: Entity Chunking, Entity Extraction oder Entity Identification) ist ein Teilgebiet der Information Extraction (IE). NER-Algorithmen erkennen und benennen vordefinierte Informationseinheiten (sog. Entitäten). Je nach Anwendungsfall gibt es unterschiedliche Auslegungen, welche Einheiten und auf welchem Granularitätniveau, Einheiten als Entitäten definiert werden. Es gibt jedoch konventionelle Entitäten, welche unumstritten als solche anerkannt sind. Entitäten werden für gewöhnlich gemäß des XML-Stilformats, beschrieben in der Message Understanding Conference (MUC)[5], definiert:

<**tag** TYPE="**type**">*Entitätsreferenz*</**tag**>

Je nach Projekt und Bereich werden nicht alle diese Entitäten verwendet oder es werden domänen-spezifische Entitäten definiert und hinzugefügt.
In dieser Arbeit wird von *Entität* gesprochen, wenn wie in Tabelle 2.1 dargestellt, die Ausprägungen von „type“ (Ort, Personen, etc.) gemeint sind. Eine *Entitätsreferenz* ist eine konkrete Ausprägung einer *Entität* (Amerika, Christoph Kolumbus, etc.). Die dargestellten Beispiele wären dementsprechend:

Enitätsreferenz: Amerika → *Entität*: Ort
Entitätsreferenz: Christoph Kolumbus → *Entität*: Person

Entitätsreferenzen sind Worte bzw. Token in einem Dokument. Sind Token keiner Entität zuzuordnen, werden sie mit dem Label „Other“ versehen.

2.4.1 *Manuelle Verfahren für Named Entitiy Recognition*

Zu Beginn des Aufkommens von NER Aufgaben, setzten die Systeme auf manuell aufgestellte, regelbasierte Algorithmen, um Entitätsreferenzen in Texten zu identifizieren und sie mit der richtigen Entität zu benennen.[20]

2.4.1.1 *Wörterbuch Lookup*

Der simpelste Ansatz ist ein Wörterbuch Lookup, bei dem die Entitätsreferenzen aus dem Trainingsdatensatz mit den jeweiligen Entitäten abgespeichert werden (Wörterbuch). Der Testdatensatz vergleicht die Token mit dem Wörterbuch und labelt diese dementsprechend. Falls der Token des Testdatensatzes nicht im Wörterbuch existiert, wird er mit „Other“ gelabelt. Einerseits ist der Ansatz sehr simpel, leicht zu implementieren und je nach

Speicherung des Wörterbuchs besitzen sie eine kurze Laufzeit. Andererseits werden keine neuen Entitätsreferenzen erkannt, da nur Entitätsreferenzen erkannt werden können, die im Trainingsdatensatz existieren.[20] Das führt zu Problemen, wenn eine unglückliche Aufteilung des Trainings- und Testdatensatzes vorgenommen wurde oder wenn über die Zeit neue Entitätsreferenzen hinzukommen z. B. wenn in der Biomedizin ein neues Medikament entwickelt wird. Hyponyme stellen bei einem reinen Wörterbuch Lookup Probleme dar. So kann z. B. ein „Ball“ ein Spielgegenstand aber auch eine Tanzveranstaltung meinen. Zur Disambiguierung müssen Begriffe in ihrem Kontext betrachtet werden.

2.4.1.2 *Manuelle Regeln*

Ein weiterer statischer Ansatz ist das manuelle Erstellen von linguistischen, grammatikalisch-basierten Regeln. Für jede Entität wird auf Basis der dazugehörigen Entitätreferenzen Muster erkannt, die viele Entitätsreferenzen innerhalb einer Entität gemeinsam haben.[20] Eine Regel für die Identifikation von Personen im isländischen Sprachraum ist z. B. die Endung eines Wortes auf „-son“ oder „-dottir“.[1] Ein generischerer Ansatz ist die Beschreibung der Regeln durch Wortstrukturen, wie Kapitalisierung, Länge, interne Zeichensetzung, u.a..[20] Das Erstellen von Regeln löst zwar das Problem des Wörterbuch Lookups, sodass auch neue Entitätsreferenzen erkannt werden können, ist allerdings sehr zeitintensiv und erfordert Expert*innen-Wissen.[21]

2.4.2 *Machine Learning Verfahren für Named Entitiy Recognition*

Mit der Weiterentwicklung von Technologien verwenden heutige Systeme Machine Learning Ansätze, die den Prozess automatisieren und den manuellen Aufwand auf ein Minimum reduzieren.[26] Machine Learning Ansätze lassen sich in drei Kategorien spezifizieren: supervised, semi-supervised und unsupervised. Supervised Machine Learning nutzt das Vorhandensein von annotierten Daten, um Muster der einzelnen Kategorien zu erlernen. Bei unsupervised Learning Algorithmen gibt es den Vorteil von annotierten Daten nicht. Ein Algorithmus muss anhand der Beobachtungen Muster erkennen und basierend auf den vorhandenen Informationen gleiche Beobachtungen gruppieren. Semi-supervised Learning Methoden nutzen einen kleinen Menge an annotierten Daten gemischt mit Daten ohne Label. Da für diese Arbeit annotierte Daten vorliegen, wird der Ansatz des supervised Learnings gewählt.

Named Entitiy Recognition Aufgaben werden als Sequenzlabelling Problem modelliert. Dabei wird gegeben einer beobachteten Inputsequenz, jedem Element der Sequenz ein Label zugewiesen.[33] Im Falle von Named Entity Recognition Aufgaben besteht die Inputsequenz aus Wörtern eines Satzes und die zuzuweisenden Labeln entsprechen Entitäten.

1 Im Isländischen setzt sich der Nachname eines Kindes aus dem Vorname des Vaters und dem Suffix „-son“ (deutsch: Sohn) für Jungs und „-dottir“ (deutsch: Tochter) zusammen.

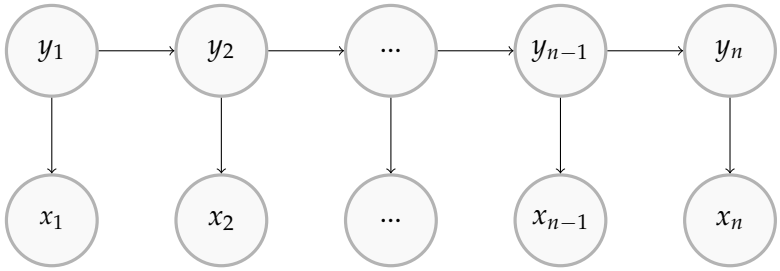


Abbildung 2.2: Graph Hidden Markov Modelle

In einem stochastischen Framework, ist die beste Outputsequenz $Y = (Y_1, Y_2, ..., Y_{n-1}, Y_n)$ für eine beobachtete Inputsequenz $X = (X_1, X_2, ..., X_{n-1}, X_n)$, diejenige welche die bedingte Wahrscheinlichkeit $P(Y|X)$ maximiert.[20] Im Folgenden seien $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_n)$ Realisierungen von X und Y. Angenommen es gibt K mögliche Label, ergibt sich für die Realisierung y_i der Zufallsvariable Y_i, dass $y_i \in \{1, \ldots, K\}$, $i = 1, \ldots, n$ gilt.

Nachfolgend werden drei aufeinander aufbauende Modelle (Hidden Markov Modell (HMM), Maximum Entropy Markov Modell (MEMM) und Conditional Random Fields (CRF)) miteinander verglichen. Am Ende der Beschreibung dieser Modelle findet sich in Abbildung 2.5 eine graphische Darstellung der Modellzusammenhänge und in Tabelle 2.2 wird eine tabellarische Zusammenfassung der Ergebnisse gegeben.

2.4.2.1 *Hidden Markov Modelle (HMM)*

In zusammenhängenden Texten besteht eine Abhängigkeit der Entitäten, d.h. sie können nicht einzeln also unabhängig voneinander betrachtet werden, sondern müssen in den Kontext gesetzt werden. Hidden Markov Modelle modellieren eine Sequenz an Beobachtungen X, indem sie eine den Beobachtungen zugrundeliegende unbeobachtete (hidden) Sequenz an Zuständen Y annehmen. HMM besitzen zwei essentielle Eigenschaften: 1. Jeder Zustand y_t hängt ausschließlich vom direkten vorangegangenen Zustand y_{t-1} ab. 2. Jede Beobachtung x_t hängt ausschließlich vom aktuellen Zustand y_t ab.[20] Abbildung 2.2 verdeutlicht die Abhängigkeiten graphisch. In HMM ist die gemeinsame Verteilung $P(Y, X)$ definiert als[33]:

$$P(Y, X) = P(X)P(Y|X). \tag{2.1}$$

Gegeben der Eigenschaften von HMM ist $P(Y|X)$ unbekannt und soll bestimmt werden (2. Eigenschaft: x_t hängt von y_t ab). Unter Anwendung des Bayes Theorems[11], lässt sich $P(Y, X)$ umformulieren. Das Bayes Theorem besagt Folgendes:

$$P(X, Y) = P(X)P(Y|X) \iff P(Y, X) = P(Y)P(X|Y). \tag{2.2}$$

Daraus folgt:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}. \tag{2.3}$$

In Hidden Markov Modelle wird mit zwei Matrizen gerechnet. Zum Einen wird die Übergangswahrscheinlichkeiten der Zustände in der so genannten *Transition-Matrix* abgebildet. Zum Anderen wird die Wahrscheinlichkeit, dass eine Beobachtung unter einem bestimmten Zustand auftritt in der so genannten *Emission-Matrix* abgebildet.[2] Unter Einbezug der 1. Eigenschaft von HMM (1. Eigenschaft: y_t hängt von y_{t-1} ab), lässt sich die Gleichung 2.3 für ein Wort und ein Label wie folgt notieren:

$$P(Y_i = y_i|X_i = x_i) = \frac{P(X_i = x_i|Y_i = y_i)P(Y_i = y_i|Y_{i-1} = y_{i-1})}{P(X_i = x_i)} \tag{2.4}$$

Zur vereinfachten Notation wird die Gleichung 2.4 (und alle Weiteren) wie folgt notiert:

$$P(Y_i|X_i) = \frac{P(X_i|Y_i)P(Y_i|Y_{i-1})}{P(X_i)} \tag{2.5}$$

Die optimale, geschätzte Outputsequenz der Zustände $\hat{Y}$ gegeben der Beobachtungssequenz X, ist diejenige mit der größten Wahrscheinlichkeit $P(\hat{Y}|X)$. Da die Beobachtungen X im Maximierungsprozess für alle möglichen Y gleich ist, kann der Bruch vernachlässigt werden. Das Problem lässt sich dann wie folgt beschreiben:

$$\hat{Y} = \arg\max_Y P(Y|X) \tag{2.6}$$

$$= \arg\max_Y P(X|Y)P(Y) \tag{2.7}$$

$$= \arg\max_Y \prod_{i=1}^{n} P(x_i|y_i)P(y_i|y_{i-1}) \tag{2.8}$$

HMM nutzen nur ein einziges Feature zur Bestimmung von $\hat{Y}$, nämlich das Wort an sich. Das führt allerdings zu Problemen, wenn im Testdatensatz Wörter enthalten sind, die nicht im Trainingsdatensatz vorhanden sind.[33] Das Modell kennt diese neue Information nicht und weiß nicht, wie es mit ihr umgehen soll.

2.4.2.2 *Maximum Entropy Markov Modelle (MEMM)*

Maximum Entropy Markov Modelle sind eine Kombination aus Hidden Markov Modellen (HMM) und Maximum Entropy Modellen (MEM). Genau wie logistische Regressionen sind MEM Log-Lineare Modelle. Im Gegensatz zur logistischen Regression können Maximum Entropy Modelle mit multinomialen und nicht nur mit binomialen Verteilungen umgehen. Modelle

treffen ihre Vorhersage für ein Label auf Basis mehrerer Feature für einen gegebenen Datenpunkt. Sie ziehen jedoch *nicht* die sequentielle Struktur der Daten in Betracht.

Um die Abhängigkeiten zwischen den Labeln zu modellieren, werden Maximum Entropy Markov Modelle verwendet. Diese beziehen die sequentiellen Inputwerte durch HMM ein und modellieren zusätzliche Feature durch MEM.

Im Gegensatz zu HMM bestehen MEMM nicht mehr aus zwei Matrizen (Transition- und Emission-Matrix), sondern nur noch aus einer. Diese Matrix bildet die Übergangswahrscheinlichkeit für jedes mögliche vergangene Label y_{i-1} kombiniert mit dem aktuellen Beobachtungswert x_i zum aktuellen Label y_i ab. MEMM sind diskriminative Modelle und berechnen direkt die bedingte Verteilung $P(Y|X)$:

$$P(Y|X) = \prod_{i=1}^{n} P(y_i|y_{i-1}, x) \tag{2.9}$$

In Abb. 2.3 findet sich eine graphische Darstellung der Zusammenhänge.

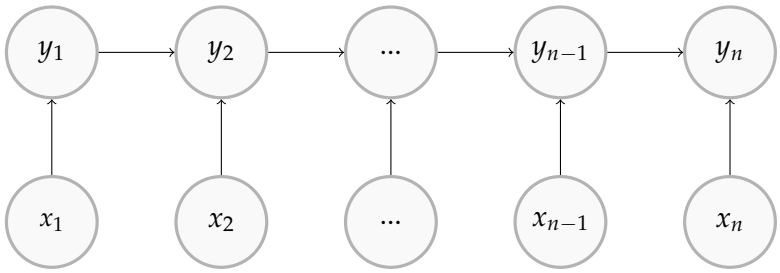


Abbildung 2.3: Graph Maximum Entropy Markov Modelle

Log-Linear Modelle nutzen eine Linearkombination aus Gewichten w_l und Feature, die durch eine Feature-Funktion f_l, $l = 1, \ldots, J$ ausgedrückt werden. Mithilfe der Feature-Funktion können zusätzlich generierte Feature zu einer Beobachtung wie z. B. Wortlänge, POS-Tag, Kapitalisierung etc. definiert werden.[33] Die Gewichte werden mittels Gradientenverfahren bestimmt und beschreiben wie relevant eine Feature-Funktion für die Ermittlung eines Labels ist. Für MEMM sieht die bedingte Wahrscheinlichkeitsverteilung wie folgt aus[33]:

$$P(y_i|y_{i-1}, x) = \frac{1}{Z_i(y_{i-1}, x)} \exp\left(\sum_{l=1}^{J} w_l f_l(y_i, y_{i-1}, x)\right), \tag{2.10}$$

wobei der Nenner $Z_i(y_{i-1}, x)$ zur Normalisierung über alle möglichen Labelsequenzen dient:

$$Z_i(y_{i-1}, x) = \sum_{y'} \exp\left(\sum_{l=1}^{J} w_l f_l(y'_i, y_{i-1}, x)\right), \tag{2.11}$$

wobei $y' = (y'_1, \ldots, y'_n)$, wobei $y'_i \in \{1, \ldots, K\}$.

Es gibt K^n mögliche Labelsequenzen. Für große n ist die Berechnung über alle möglichen Labelsequenzen zu kostenintensiv. Durch Anwendung des Viterbi-Algorithmus kann die Laufzeit auf $\mathcal{O}(nK^2)$ reduziert werden. Die optimale Labelsequenz $\hat{Y}$ für Maximum Entropy Markov Modelle ist diejenige mit der höchsten bedingten Wahrscheinlichkeit $P(\hat{Y}|X)$:

$$\hat{Y} = \arg\max_Y P(Y|X) \tag{2.12}$$

$$= \arg\max_Y \prod_{i=1}^{n} P(y_i|y_{i-1}, x) \tag{2.13}$$

$$= \arg\max_Y \prod_{i}^{n} \frac{1}{Z_i(y_{i-1}, X)} \exp(\sum_{l=1}^{J} w_j f_j(y_i, y_{i-1}, x)) \tag{2.14}$$

MEMM lösen das Problem von HMM, indem direkt die bedingte Verteilung anstelle der gemeinsamen Verteilung modelliert wird. Eine bedeutende Schwierigkeit von MEMM stellt jedoch das *Label Bias Problem* dar, verursacht durch die lokale Normalisierung $Z(X)$. Bei dem Übergang von einem Zustand in einen anderen, konkurrieren die verschiedenen Zustände miteinander (Summe der Übergangswahrscheinlichkeiten ergibt 1). Es werden die Übergänge gewählt, welche die Wahrscheinlichkeit der Sequenz maximieren. Die Übergangswahrscheinlichkeit zu Zuständen, die nur aus wenigen anderen Zuständen entstehen können, haben nur wenige Kanten. Da sich die Kantensummen 1 ergibt, haben die wenigen Kanten jeweils eine hohe Übergangswahrscheinlichkeit. Das führt dazu, dass diese Kanten im Maximierungsprozess bevorzugt werden. Folglich werden die Zustände bevorzugt, zu denen nur wenige Kanten führen.[33]

2.4.2.3 *Conditional Random Fields (CRF)*

Conditional Random Fields ähneln in großen Teilen MEMM: Sie sind Log-Lineare Modelle, fallen unter diskriminative Modelle und modellieren somit direkt die bedingte Wahrscheinlichkeit $P(Y|X)$[36]:

$$P(Y|X) = \prod_{i=1}^{n} P(y_i|y_{i-1}, x) \tag{2.15}$$

Linear-Chain CRF sind spezielle CRF, welche die Outputvariablen als Sequenz modellieren und somit für NER Tasks geeignet sind. Im Folgenden sind Linear-Chain CRF gemeint, wenn von CRF gesprochen wird.

$$P(y_i|y_{i-1}, x) = \frac{1}{Z(x)} \exp(\sum_{l=1}^{J} \sum_{i=1}^{n} w_l f_l(y_{i-1}, y_i, x, i)), \tag{2.16}$$

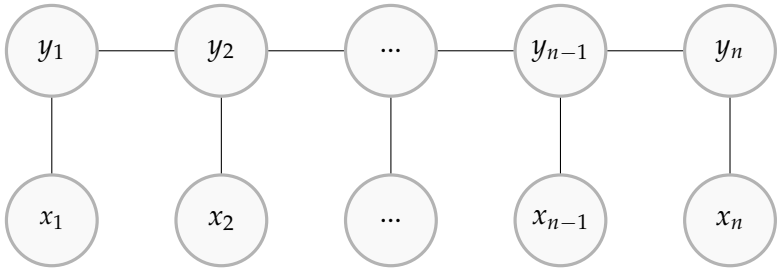


Abbildung 2.4: Graph Conditional Random Fields

wobei der Nenner $Z(x)$ zur Normalisierung über alle möglichen Labelsequenzen dient:

$$Z(x) = \sum_{y} \exp\left(\sum_{l=1}^{J} \sum_{i=1}^{n} w_l f_l(y_{i-1}, y_i, x, i)\right) \tag{2.17}$$

MEMM und CRF ähneln sich sehr in ihrer Form. Es gibt jedoch einen entscheidenden Unterschied, der in der Normalisierung liegt. Bei den Maximum Entropy Markov Modellen wird eine lokale Normalisierung über $Z_i(y_{i-1}, x)$ vorgenommen und bei Conditional Random Fields wird eine globale Normalisierung über $Z(x)$ durchgeführt.[36] Dadurch löst sich das Label Bias Problem und die optimale, geschätzte Labelsequenz $\hat{Y}$ bestimmt sich wie folgt:

$$\hat{Y} = \arg\max_{Y} P(Y|X) \tag{2.18}$$

$$= \arg\max_{Y} \sum_{l=1}^{K} \sum_{i=1}^{n} w_i f_i(y_{i-1}, y_i, x, i) \tag{2.19}$$

Da die Exponentialfunktion streng monoton steigend und $Z(X)$ für alle möglichen Labelsequenzen konstant ist (Gleichung 2.16), kann dies zur Vereinfachung in der Gleichung 2.19 vernachlässigt werden.

2.4.3 *Evaluation*

Bei der Evaluation eines NER Systems können zwei Ausgänge auftreten: Es findet eine richtige oder eine falsche Klassifizierung statt. Das optimale Szenario ist, dass das System alle Entitäten richtig erkennt: Richtig-Positiv und Richtig-Negativ. Es ist sehr unwahrscheinlich, dass das System alle Token richtig zuordnet. Es können zwei Fehler auftreten: Falsch-Positiv und Falsch-Negativ.

Es gibt je nach Evaluationsmethode für Named-Entity-Recognition unterschiedliche Definitionen der genannten Kategorien. Entweder es wird auf Token-Level oder auf Entitäts-Level evaluiert. Token-Level bedeutet, dass immer nur einzelne Token betrachtet werden und eine Prüfung stattfindet, ob der Token der korrekten Entität zugeordnet wird. Hier bedeutet, dass B-XXX

	Hidden Markov Modelle (HMM)	Maximum Entropy Markov Modelle (MEMM)	Conditional Random Fields (CRF)
Modellart	generative	diskriminative	diskriminative
Struktur	gerichtet	gerichtet	ungerichtet
Wktsverteilung	$P(Y,X) = \prod_{i=1}^{n} P(x_i\|y_i)P(y_i\|y_{i-1})$	$P(Y\|X) = \prod_{i=1}^{n} P(y_i\|y_{i-1},x)$	$P(Y\|X) = \prod_{i=1}^{n} P(y_{i-1}\|y_i,x)$
Funktion		$P(y_i\|y_{i-1},x) = \frac{1}{Z_i(y_{i-1},x)} \sum_{y'} \exp(\sum_{l=1}^{J} w_l f_l(y_i,y_{i-1},x))$	$P(y_{i-1}\|y_i,x) = \frac{1}{Z(x)} \sum_{y} \exp(\sum_{l=1}^{J}\sum_{i=1}^{n} w_l, f_l(y_{i-1},y_i,x,i))$
Normalisierung		lokal	global
Normalisierungs-funktion		$Z_i(y_{i-1},x) = \sum_{y'} \exp(\sum_{l=1}^{J} w_l f_l(y'_i,y_{i-1},x)$	$Z(x) = \sum_{y} \exp(\sum_{l=1}^{J}\sum_{i=1}^{n} w_i f_i(y_{i-1},y_i,x,i))$
Vorteile ggü. voran-gehendem Modell	Modelliert Abhängig-keiten der Label.	Es können weitere Feature durch die Feature Funktion integriert werden.	Löst das Label Bias Problem.

Tabelle 2.2: Modellübersicht

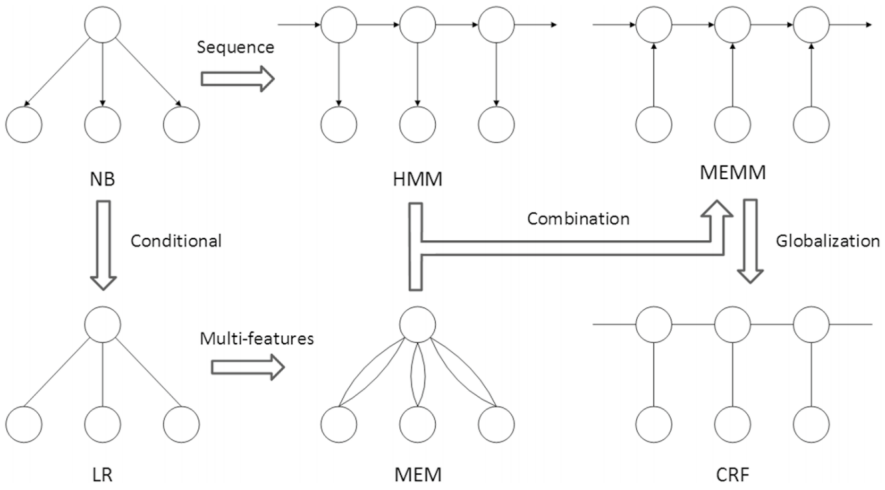


Abbildung 2.5: Übersicht der Modellzusammenhänge (Quelle: [36] S.4291)

und I-XXX *jeweils* eine Entität darstellen. Eine Evaluation auf Entitäts-Level bedeutet, dass B-XXX und I-XXX *zusammen* eine Entität darstellen und diese nur richtig erkannt wird, wenn *alle* Token, die zu dieser Entität gehören, richtig erkannt werden.[23]

Text	Johann	Wolfgang	von	Goethe
Goldstandard	B-PER	I-PER	I-PER	I-PER
Vorhersage	B-PER	I-PER	O	O

Tabelle 2.3: Beispiel: Evaluation

Wird Beispiel 2.3 auf Token-Level evaluiert, gibt es zwei richtige Zuordnungen („Johann“, „Wolfgang“) und zwei falsche Zuordnungen („von“, „Goethe“). Wird das Beispiel allerdings auf Entitäts-Level evaluiert, gibt es keine richtige Zuordnung, da „Johann Wolfgang von Goethe“ eine Entität darstellt, deren Token alle richtig zugeordnet werden müssten. Vielmehr werden hier sogar zwei Fehler gezählt: 1 Falsch-Positiv („Johann Wolfgang“) und 1 Falsch-Negativ („Johann Wolfgang von Goethe“). Nachfolgend werden die vier verschiedenen Klassifikationen, die ein System machen kann aufgelistet und generisch erklärt: Das System ...

RICHTIG-POSITIV	erkennt eine tatsächliche Entität.
RICHTIG-NEGATIV	erkennt, wenn ein Token nicht zu einer Entität gehört.
FALSCH-POSITIV	vermutet eine Entität, die keine ist.
	vermutet eine falsche Entität.
	setzt die Grenzen einer Entität falsch.
FALSCH-NEGATIV	übersieht eine Entität.
	setzt die Grenzen einer Entität falsch.

Die Klassifikationen gelten sowohl für eine Evaluation auf Token-Level als auch auf Entitäts-Level, wobei die Ausprägung „setzt die Grenzen einer Entität falsch“ nur für letztere gilt. Bei der Auflistung ist deutlich zu sehen, dass das Setzen von falschen Grenzen zu zwei Fehlern führt. In diesen Fällen wäre es „besser“ gewesen, wenn anstelle einer Entität mit falschen Grenzen, keine Entität vorhergesagt würde, da daraus nur ein Fehler (Falsch-Negativ) resultiert. Zusätzlich wird klar, dass die Evaluation auf Entitäts-Level strenger ist, da mehr Fehler gemacht werden können. Eine allgemeine, vereinfachte Darstellung der Klassifikationen wird in einer sogenannten Confusion Matrix präsentiert (Abb. 2.4). Die Abbildung hilft, um nachfolgend die Erklärung der Definitionen von Precision, Recall und dem F-Score zu verdeutlichen. Y steht für das tatsächliche Label und $\hat{Y}$ für das vom Modell vorhergesagte Label. Durch ihre Eingängigkeit und den regelmäßigen Gebrauch werden auch für die Klassifikationstypen die englischen Ausdrücke bzw. Abkürzungen verwendet, sodass sich Folgendes ergibt: Richtig-Positive: TP (True-Positiv), Richtig-Negativ: TN (True-Negative), Falsch-Positive: FP (False-Positive) und Falsch-Negative: FN (False-Negative).

		Y	
		Positive	Negative
$\hat{Y}$	Positive	TP	FP
	Negative	FN	TN

Tabelle 2.4: Confusion Matrix

2.4.3.1 *Precision, Recall und F-Score*

Im Information Retrieval sind Precision, Recall und der F-Score bewährte Evaluationsmaße für Klassifikationsprobleme.[15] Precision beschreibt den Anteil der richtig vorhergesagten Entitäten an alles vorhergesagten Entitäten:

$$Precision = \frac{TP}{TP + FP}. \tag{2.20}$$

Recall bestimmt den Anteil der richtig vorhergesagten Entitäten an den tatsächlichen Entitäten:

$$Recall = \frac{TP}{TP + FN}. \tag{2.21}$$

Der F-Score ist das gewichtete harmonische Mittel von Precision und Recall und ist wie folgt definiert:

$$F_\beta - Score = (1 + \beta^2)\frac{Precision * Recall}{\beta^2 * Precision + Recall} \tag{2.22}$$

$$= \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP}. \tag{2.23}$$

Für gewöhnlich werden Precision und Recall gleich gewichtet:

$$F_1 - Score = 2\frac{Precision * Recall}{Precision + Recall} \tag{2.24}$$

$$= \frac{TP}{TP + \frac{1}{2}(FN + FP)}. \tag{2.25}$$

Der Fall einer hohen Precision bedeutet dementsprechend, dass die Inputwerte, welche das Modell als Entitäten identifiziert hat, mit hoher Wahrscheinlichkeit eine tatsächliche Entität sind und somit richtig erkannt wurden. Das Maß trifft allerdings keine Annahmen darüber, wie viele von den eigentlichen Entitäten detektiert wurden. Er gibt lediglich Auskunft über die „Glaubwürdigkeit" eines Labels. Komplementär dazu dient der Recall, welcher angibt, wie viele der tatsächlichen Entitäten gefunden wurden. Es mangelt dem Recall jedoch daran Aussagen zur Güte eines Labels zu treffen.

Im Optimalfall sind sowohl Precision als auch Recall hoch, sodass alle Entitäten gefunden werden und die Entitäten, die als solche gelabelt werden, auch tatsächlich Entitäten sind.

3 LITERATURÜBERBLICK

In diesem Kapitel werden zu Beginn (Abschnitt 3.1) Herausforderungen im Umgang mit Texten, die auf nicht-standardisierten Rechtschreibungen basieren, aufgezeigt und mögliche Lösungsvorschläge präsentiert. Anschließend (Abschnitt 3.2) werden Paper vorgestellt, die verschiedene Natural Language Processing Techniken auf historische Texten anwenden, um entweder strukturelle und narrative Analysen durchzuführen oder mit dem Ziel ein übersichtliches Datenmanagement-System aufzubauen. Danach (Abschnitt 3.3) liegt der Fokus auf Named-Entity-Recoginition Aufgaben, indem zuerst ein manueller Ansatz und folgend (Abschnitt 3.4) verschiedene Machine Learning Ansätze und deren Ergebnisse aufgezeigt werden.

3.1 HERAUSFORDERUNGEN DER SPRACHVARIANZ

In den Digital Humanities stellt sich die Frage wie moderne Technologien einen gewinnbringenden Beitrag zu Humanwissenschaften leisten können. Wie speziell computerlinguistische Ansätze einen Mehrwert schaffen können, wurden von der COLING (Conference on Computational Linguistic) 2016 in Osaka, Japan, im Rahmen des Workshops LT4DH (Language Technology for Digital Humanities) untersucht. Neben anderen Fragestellungen, wurde die Herausforderungen diachronischer Daten und die damit verbundenen Sprachvariationen diskutiert. Weiterführend stellt sich die Frage wie regelbasierte, statistische und Machine Learning Methoden im Bereich der Digital Humanities adaptiert werden können. Hinrichs u. a. [19]

Die Normierung der deutschen Rechtschreibung fand im Laufe des 19. Jahrhunderts statt. Der erste Duden erschien 1880 und wurde 1901 zur amtlichen Rechtschreibung erklärt. Die jüngste und seit 1650 bestehende Sprachstufe ist Neuhochdeutsch, vorausgehend von Frühneuhochdeutsch (1350-1650), Mittelhochdeutsch (1050-1350) und Althochdeutsch (750-1050). Frank Wedekind lebte von 1864-1918. Er ist zwar in die modernste Sprachstufe einzuordnen aber war bei der amtlichen Standardisierung der Rechtschreibung bereits 37 Jahre alt. Die Briefe in der Frank Wedekind Briefedition wurden zwischen 1882 und 1918, wovon mehr als $\frac{1}{3}$ vor 1901 geschrieben wurden. Zusätzlich ist davon auszugehen, dass auch Briefe nach 1901 immer noch mit nicht-standardisierter Rechtschreibung verfasst wurden.

Wenn Dokumente in digitalen (historisch-kritischen) Editionen auf unstandardisierter Sprache basieren, kann das unterschiedliche Probleme verursachen. Einerseits erschwert es die Suche, wenn keine Techniken implementiert sind, die mit unterschiedlichen Schreibvarianten zwischen Query und Dokument umgehen können. Andererseits hat es Auswirkungen auf das Training als auch das Testen von Natural Language Processing Methoden:

Wörter, welche dieselbe Information liefern werden im Training durch ihre unterschiedliche Schreibweise getrennt voneinander verarbeitet. Während des Testens ist die Wahrscheinlichkeit unbekannter Wörter (Wörter, die in anderer Schreibweise im Trainingsdatensatz sind) größer.

Barteld, Biemann und Zinsmeister [1] beschäftigen sich anhand des Mittelhochdeutschen mit der Frage wie Variationen in der Rechtschreibung eines Wortes erkannt werden können. Beispielsweise können „in“ und „ihm“ im Mittelhochdeutsch das gleiche Wort repräsentieren. Der populärste Ansatz ist eine Normalisierung/ Kanonisierung beider zu modernem Deutsch „ihn“. Barteld, Biemann und Zinsmeister folgen nicht dem klassischen Ansatz der Kanonisierung zu einer standardisierten Schreibweise, sondern erstellen eine „*historical variant detection*“. Dabei wird ein Set aus allen möglichen Schreibvariationen eines Wortes generiert und ein unbekanntes Wort bzw. eine unbekannte Schreibweise eines Wortes wird unter Einbezug des linguistischen Kontexts zur passenden historischen Variante zugeordnet. Dieser Ansatz hat den Vorteil, dass auch Wörter verarbeitet werden, die kein ähnliches standardisiertes Pendant besitzen. Da die Sprachstufe von Frank Wedekind zeitlich sehr nah an dem Zeitpunkt der Standardisierung liegt, ist davon auszugehen, dass nur wenig Varianz in der Sprache besteht und moderne Lemmatizierer die Mehrzahl der Wörter richtig kanonisieren.

Viele Machine Learning Modelle verwenden nicht nur das Wort als Feature, sondern auch Merkmale, die vom Wort abgeleitet werden können. Oft verwendete linguistische Merkmale eines Wortes sind das eben erwähnte Lemma also der Wortstamm und der Part-of-Speech-Tag (PoS-Tag), welches eine Wortartzuordnung durchführt. Die meisten Lemmatizer und PoS-Tagger sind für die englische Sprache erstellt. Ähnlich viele gibt es für das Neuhochdeutsche. Einer dieser ist der TreeTagger, welcher eine 97.5% Genauigkeit auf Zeitungsartikeln erzielt.[29] Die Performanz solcher Tagger lässt auf früheren Sprachstufen aufgrund der Variation der Rechtschreibung jedoch nach. Der TreeTagger ist seit 2017 auch auf mittelhochdeutsche Texte anwendbar.[9] Es ist wie bereits erwähnt davon auszugehen, dass diese Tools auf der Sprache von Frank Wedekind ausreichend gut performen.

3.2 NATURAL LANGUAGE PROCESSING

Wenn die Schwierigkeiten, welche sich durch Varianzen in der Orthographie ergeben, überwunden sind, gibt es viele Möglichkeiten, wie Natural Language Processing Verfahren digitale (historisch-kritische) Editionen verbessern können.

Moderne Technologien können an verschiedene Aufgaben von Editor*innen anknüpfen können. Ausgewählte Bereiche, die durch computerbasierte Tools und Methoden unterstützt werden können, sind das Transkribieren von Texten, die Analyse von Sprache, inhaltliche, thematische Analysen und das Anreichern der Texte durch externe Informationen.

Durch die Digitalisierung von nicht-abstrakten Gegenständen, d.h. durch die Generierung von Daten, entsteht ein neues Potential der kritischen Aus-

einandersetzung mit Dokumenten. Dieses Potential entfaltet sich insbesondere im Natural Language Processing. Das Buch „Language Technology for Cultural Heritage“ [32] gibt einen sehr guten Überblick über die Möglichkeiten, wie Technologien zur Sprachverarbeitung für historische Dokumente Anwendung finden können. Es werden hierbei sechs verschiedene Themenbereiche präsentiert: 1. Technologien zu Vorverarbeitung historischer Überlieferungen, 2. Adaptivität existierender Language Processing Tools an ältere Sprachstufen, 3. linguistische Ressourcen, 4. Personalisierung, 5. strukturelle und narrative Analysen und 6. Datenmanagement, Visualisierung und Information Retrieval. Aufgrund des stärkeren Bezugs zu Natural Language Processing, werden nachfolgend die Punkte 5. und 6. genauer betrachtet. Um einen greifbaren Einblick im Zusammenhang dieser Themen mit digitalen (historisch-kritischen) Editionen zu ermöglichen, werden jeweils exemplarische Paper vorgestellt, die einen praktischen Bezug herstellen.

Named-Entity-Recognition Verfahren sind thematisch in den aktuellen Abschnitt „Natural Language Processing“ einzuordnen. Aufgrund der Relevanz und des Umfangs wird der aktuelle Forschungsstand dazu in einem eigenständigen Abschnitt vorgestellt.

3.2.1 *Strukturelle und narrative Analysen*

Im Natural Language Processing können strukturelle und narrative Analysen des Inhaltes automatisiert werden. Je nach Analyseschwerpunkt vereinfacht es die Recherchen im Bereich der Linguistik, Literatur, Historik, Soziologie und Ethnologie.[32] Nachfolgend werden drei Paper präsentiert, die zum genannten Themenbereich Methoden und Tools auf historische Dokumente angewendet haben. Sie beschäftigen sich mit Textsegmentierung, Annotationen und der Identifikation von wiederkehrenden Elementen über einzelne Dokumente hinaus.

In privaten, porugiesischen Briefen dem 16. - 19. Jahrhundert findet eine Diskursanalyse der Gesellschaft statt. Dafür teilen Hendrickx, Généreux und Marquilhas[17] die Briefe in verschiedene Segmente hinsichtlich ihrer Diskursfunktion (*opener* (Name, Datum, Ort, Begrüßung), *harengue* (formeller Einleitungsteil), *peroration* (Schlussfolgerung), *closer* (Name, Signatur, Ort)). Die Zuordnung der einzelnen Briefsegmente zu den Diskursfunktionen findet automatisch durch n-Gramme statt. Durch Machine Learning Algorithmen erhalten die Abschnitte zusätzlich semantische Kategorien (z. B. Körper (Gesundheit, Krankheit), Sozial (Religion, Beziehung)), die den Inhalt thematisch beschreiben.

Eine weitere Möglichkeit für narrative Analysen bieten Declerck, Scheidel und Lendvai [6], die ein Annotationssystem, das automatisch Märchen analysiert und mit den von Propp definierten narrativen Motiven annotiert, aufstellen.

Um ein größeres Verständnis für indische Rituale zu erhalten, stellen Reiter u. a.[27] ein Framework auf, das wiederkehrende Elemente über verschie-

dene Texte hinweg erkennt. Dadurch ist es ihnen möglich zu identifizieren, welche Bräuche in indischen Ritualen gängig sind.

3.2.2 *Datenmanagement, Visualisierung und Information Retrieval*

Natural Language Processing Techniken können nicht nur für die Analyse von Strukturen und linguistischen Einheiten verwendet werden, sondern bieten auch Möglichkeiten, Daten übersichtlich und eingängig zu verwalten und zu präsentieren.

Berzak u. a.[3] verwenden Natural Language Techniken, um eine graphische Visualisierung einer große Dokumentensammlung mit Reden von Fidel Castro zu erstellen. Die einzelnen sog. Knoten im Graphen stehen für die Dokumente und die semantischen Verbindungen werden mit sog. Kanten abgebildet. Je ähnlicher sich einzelne Dokumente sind, desto dicker werden die verbindenden Kanten gezeichnet. Die Ähnlichkeit von zwei Dokumenten wird auf zwei Arten gemessen. Einerseits auf Basis der Überlappenden Named Entities (Namen, Orte, Daten, etc.) und andererseits aufgrund der Ähnlichkeit ihrer Vokabulare, die mittels Cosinus-Ähnlichkeit bestimmt wird. Aufgrund der Übersichtlichkeit und der Relevanz, entsteht eine Verbindung zwischen zwei Dokumenten dann, wenn die Ähnlichkeit einen bestimmten Schwellenwert überschreitet. Zusätzlich wird ein Clusteralgorithmus angewendet, der Dokumentengruppen identifiziert und diese im Graphen nah beieinander abbildet. Die Interpretation der Cluster wurde nicht genauer erörtert, vorzustellen wäre aber z. B. ein Cluster auf Basis der inhaltlichen Themen der Dokumente. Dieser Ansatz bringt auf mehreren Ebenen Vorteile. Zum einen erlaubt es einen schnellen Überblick und ein leichtes Navigieren in der Dokumentenkollektion. Zum anderen dient es unterstützend zur Findung der gewünschten Ergebnisse bei einer Suchanfrage, da der/die Nutzer*in im Graphen nach ähnlichen Dokumenten zu den Ergebnisdokumenten suchen kann.

Coeckelbergs, van Hooland und van Hecke schlagen den Ansatz des Topic Modellings zur Analyse des Inhaltes vor. Mittels LDA (Latent Dirichlet Allocation) erstellen sie Themenkategorien zu Texten aus der Bibel. Dadurch werden sogenannte Topics generiert. Für jedes Wort im Corpus und jedes Dokument kann der Zusammenhang zu den einzelnen Topics ermittelt werden. Dieser Ansatz ermöglicht zusätzlich die Gruppierung thematisch ähnlicher Dokumente basierend auf ihrer Topicverteilung. Topic Modelling bringt auch Vorteile für Nutzer*innen digitaler (historisch-kritischer) Editionen, da sie einerseits die Suche unterstützen können und andererseits einen schnellen Überblick über die Inhalte der Edition bereitstellen. Die Autoren gehen noch einen Schritt weiter und verlinken die Bibeltexte zu Sekundärliteratur. Dabei werden Sätze aus der den Bibeltexten mittels Fuzzymatching-Techniken zu wissenschaftlichen Artikeln zugeordnet. Die Genauigkeit der Zuordnung wird mit zusätzlichen Einbezug von Topic Modelling erhöht.

Einen anderen Ansatz zur Wissensrepräsentation verfolgen Witte u. a.[34], die automatisch eine Ontologie auf einer Architekturenzyklopädie generieren, die ebenso eine Navigation und Suche in der Enzyklopädie erleichtert.

3.3 NAMED ENTITY RECOGNITION MIT REGELBASIERTEN VERFAHREN

Grover u. a. [16] untersuchten Aufzeichnungen von britischen Parlamentsberichten aus dem 17. Jhd. und dem frühen 19. Jhd.. Ein Named-Entity-Recognition Modell soll Personen und Orte aus den Texten extrahieren.

Es wurde sich für einen regelbasierten Ansatz mit zusätzlichem Lexikon-Lookup entschieden. Einerseits aus dem Grund, dass keine bereits annotierten Daten vorlagen und eine manuelle Annotation sehr zeitaufwendig ist. Da die Texte viele OCR-Fehler enthalten und die Verwendung von Groß- und Kleinschreibung sehr uneinheitlich ist, war die Erfolgswahrscheinlichkeit von Machine Learning Ansätzen zu gering, um das Risiko und den Zeitaufwand rechtfertigen zu können.

Es wurden sowohl für Personen als auch für Ortsnamen Regeln erstellt und Lexikon Look-Ups vorgenommen. Ein inkrementeller Ansatz, der zuerst Personen labelt und im zweiten Schritt Ortsnamen, verhindert, dass Orte als Namen gelabelt werden (z. B. Earl of Warwick). Die Ergebnisse zeigen, dass Personennamen eher erkannt werden, da sie gut durch Regeln beschrieben werden können und resistenter ggü. OCR-Fehlern sind. Ortsnamen hängen stärker von guten Lexika ab und sind nicht so resistent ggü. OCR-Fehlern. Bei einem Wörtbuch-Look-Up müssen zusätzlich Fuzzymatching-Techniken eingebaut werden.

3.4 NAMED ENTITY RECOGNITION MIT MACHINE LEARNING VERFAHREN

3.4.1 *Open-Source NER-Systeme*

Won, Murrieta-Flores und Martins untersuchen inwieweit moderne Tools geographische Informationen (Orte) in historischen Dokumenten extrahieren können. Dafür werden die Mary Hamilton Papers (1756–1816, Modern English , 161 Briefe) und die Samuel Hartlib Papers (1600-1662, Early Modern English aber auch deutsch, franz., niederl. und lateinisch, 54 Briefe) aus dem Republic of Letters (RofL) Datensatz verwendet. Die Autoren vergleichen fünf verschiedene Open-Source NER-Systeme und generieren zusätzlich ein Ensemble Modell, das alle Systeme miteinander kombiniert. Folgende NER-Systeme wurden verwendet (in Klammern sind die zugrundeliegenden Modelle angegeben): Stanford NER (CRF), NER-Tagger (LSTM-CRF), Edinburgh Geoparser (regelbasiert, Ortslexikon), spaCy (nicht beschrieben, was spaCy genau implementiert) und Polyglot-NER (einfaches NN). Die Ensemble Methode kombiniert alle Systeme mittels Voting miteinander. Für Briefe, geschrieben in Modern Englisch erzielt Polyglot-NER mit einem F_1-Score von 61.6% die besten Ergebnisse der individuellen NER-

Systeme. Für Briefe auf Early Modern Englisch eignet sich am besten das Stanford NER-System, das ein F_1-Score von 70.7% erzielt. Die Ensemble Methode schneidet am besten für beide Datensätze ab und erreicht F_1-Scores von 71.5% (Modern Englisch) und 73.3% (Early Modern Englisch). Die Ergebnisse zeigen, dass die individuellen NER-System auf Datensätze mit unterschiedlichen Sprachstufen unterschiedlich gut performen. Das bedeutet, dass die Tools verschiedene Aspekten von Sprache abdecken. Ensemble Modelle reduzieren diese Variabilität und verbessern die Ergebnisse sogar. Das Ensemble Modell performt konstant über die Datensätze hinweg und überbietet die individuellen Systeme sogar. Won, Murrieta-Flores und Martins zeigen, dass es keine beachtlichen Veränderungen gibt, wenn die historische Schreibweise (in ihrem Untersuchungsgegenstand: Early Modern English) in eine moderne, standardisierte Schreibweise (Modern English) umgewandelt wird. Hierbei ist allerdings kritisch zu betrachten, dass eine Evaluation der Güte der Transformation nicht stattfand. Dadurch besteht die Möglichkeit, dass schon in diesem vorgelagerten Schritt Fehler auftreten, die das NER-System nicht ausgleichen kann.

Die Ergebnisse zeigen, dass die verwendeten modernen Tools nur bedingt auf historische Dokumente anwendbar sind. Die individuellen Tools performen sehr unterschiedlich auf unterschiedlichen Datensätze.

3.4.2 *CoNLL 2003 NER Shared Task*

In der von SIGNLL (ACL's Special Interest Group on Natural Language Learning) jährliche organisierten CoNLL (Conference on Computational Natural Language Learning) wurde 2003 eine Named-Entity-Recognition Shared Task gestellt.[**cite{sang2003}**] Die Konferenz stellte einen englischen und deutschen Datensatz bereit. Der englische Datensatz besteht aus Zeitungsartikeln der internationalen Nachrichtenagentur Reuters aus den Jahren 1996/97. Der deutsche Datensatz besteht aus Zeitungsartikeln der Frankfurter Rundschau aus dem Jahr 1992. Die zu annotierenden Entitäten waren: Person, Organisation und Orte. Die gängigen Vorverarbeitungsschritte der Daten bestanden aus Tokenizierung, Part-of-Speech Tagging, Lemmatizierung und Chunking. Insgesamt wurden 16 Modelle eingereicht. Die Mehrheit der Teilnehmer*innen entschied sich für die Verwendung von Maximum Entropy Modellen (3 ausschließlich MEM, 2 MEM in Kombination mit anderen Techniken) oder von Hidden Markov Modellen (4 und alle davon in Kombination mit anderen Techniken). Die Top 5 Feature, welche den größten Effekt für die Performanz der Modelle hatten sind: Lexikoneinträge, Part-of-Speech Tag, n-Gram, vorhergesagte Entität des vorangegangenen Wortes und orthographische Informationen. Alle Modelle haben externe Ressourcen wie Lexika oder Gazetteer verwendet, welche die Fehler der Modelle bis zu 20% reduziert haben.

Das beste Modell, sowohl für den deutschen als auch den englischen Datensatz, besteht aus einer Kombination aus Maximum Entropy Modell, Transformation-Based Learning, Hidden Markov Modell sowie einem ro-

busten Risikominimierer. Neben dem verfügbaren Datensatz für das Training des Modells wurde zusätzlich ein Gazetteer und der Output von zwei externen Named-Entity-Recognition Systemen verwendet. Der F_1-Score des englischen Datensatzes liegt bei 88.8% und der des deutschen Datensatzes bei 72.4%.

3.4.3 *HIPE 2020 NER Shared Task*

Ehrmann u. a. [10] geben einen Überblick über die erste Ausgabe des HIPE (Identifying Historical People, Places and other Entities) Shared Tasks. Die als Untersuchungsgegenstand zur Aufgabe gestellten Daten umfassen historische Zeitungsartikel der letzten 200 Jahre in deutscher, englischer und französischer Sprache. Die zu identifizierenden Entitäten sind Person, Ort und Organisation. Populäre Regeln, die diese Entitäten beschreiben können oft z. B. aufgrund von unterschiedlichen Nameskonventionen zwischen heute und früher und nicht-standardisierter Rechtschreibung nicht angewendet werden. Externe Resourcen müssen sorgfältig und passend ausgewählt werden. Unter anderem aus den genannten Problemen decken die meisten eingereichten Modelle State-of-the-Art Deep Learning Architekturen ab. Das Baseline Conditional Random Field Modell, mit Standardfeature (Casing des Anfangsbuchstaben des Wortes, Prefix, Suffix) erzielt einen F_1-Score von 47.6% auf dem deutschen Datensatz. Das beste Modell für die deutschen Daten besteht aus einem hierarchischen Transformer-Based Modell basierend auf BERT. On Top dieses Modells wurde noch ein CRF-Layer hinzugefügt. BERT ist eine noch sehr junge Methode (2018 von Google entwickelt[7]), welche WordEmbeddings, ähnlich zu Word2Vec verwendet, mit dem bedeutenden Unterschied, dass der Kontext der Wörter größere Beachtung findet. Das Modell erzielte einen F_1-Score von 79.7%. Das zweitbeste Modell liegt mit 12% Abstand hinter dem ersten Modell. Dieses Modell verwendet ein BiLSTM mit einem CRF-Layer on Top. In diesem Netz ist ein komplexer Embeddinglayer eingebaut, verschiedene Embeddings (wortbasiert, teilwortbasiert, zeichenbasiert, vortrainierte BERT-Embeddings und vortrainierte von HIPE bereitgestellt domänenspezifische Embeddings) miteinander kombiniert. Ein Modell hat versucht, das Baseline-CRF-Modell durch zusätzliche Feature das Modell zu verbessern. Dieser Ansatz hat zu einer Verbesserung des Baseline-Modells von 4% geführt.

3.5 EPOSTOLARIUM: CIRCULATION OF KNOWLEDGE AND LEARNED PRACTICES IN THE 17TH-CENTURY DUTCH REPUBLIC

Bevor dieses Kapitel abgeschlossen wird, wird noch auf ein sehr gutes Beispiel einer digitalen (historisch-kritischen) Briefedition aufmerksam gemacht, das erfolgreich Natural Language Processing Methoden für verschiedene Zwecke implementiert: „Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic“[1]. Die Edition beinhaltet ca. 20.000

1 http://ckcc.huygens.knaw.nl/epistolarium/

Briefe von Gelehrten aus den Niederlanden des 17. Jahrhundert. Dieses Jahrhundert ist geprägt von wissenschaftlichen Revolutionen. Die niederländische Gesellschaft spielte eine zentrale Rolle in der Verbreitung von Wissen und Informationen. Wie sich durch diesen Austausch ein verflochtenes Netzwerk aufgebaut hat, lässt sich in dieser Edition nachvollziehen. Das Projekt wird getragen von einer heterogenen Gruppe niederländischer Forscher verschiedener Universitäten, Forschungseinrichtungen und kulturellen Institutionen.

Die verwendeten Natural Language Processing Methoden reichen von der Suche nach Briefen bis hin zur Visualisierung von Korrespondenzen auf Google Maps. Positiv fällt zu Beginn schon die nutzerfreundliche Oberfläche auf. Die Startseite ist so aufgeteilt, dass auf dem linken drittel des Bildschirm die Suchfunktionalitäten (Volltextsuche, Facetten-Suche, Ähnlichkeitssuche) aufgezeigt werden. Zusätzlich zur Volltextsuche werden durch Topic Modelling Verfahren ähnliche Begriffe zu den Suchbegriffen angezeigt, die durch Anklicken der Suchanfrage hinzugefügt werden können. Die Facetten-Suche ermöglicht eine Sortierung und Filterung nach verschiedenen Metadaten wie Datum oder Sender*in. Die Ähnlichkeitssuche ermöglicht die freie Texteingabe. Basierend auf diesem Text werden Briefe angezeigt, die eine große Ähnlichkeit zur Texteingabe aufweisen. Rechter Hand werden die Ergebnisse der Suchanfragen aufgelistet. Diese Ergebnisse können in verschiedenster Weise visualisiert werden. Zum einen können die Wege der Briefe von Sender*in zu Empfänger*in auf Google Maps visualisiert werden.

Des Weiteren wird durch eine graphische Darstellung angezeigt, von wem und an wen die Briefe in der Ergebnisliste verschickt wurden (Correspondent Network). Neben dem Netzwerk der Korrespondenzpartner*innen kann auch ein Netzwerk für die erwähnten Personen in den Briefen angezeigt werden (Cocitation Graph). Topic Modelling Verfahren werden verwendet, um bei ausgewählten Briefen Vorschläge für weitere, ähnliche Briefe zu liefern. Keywordanalysen werden durchgeführt, sodass für jeden Brief, die relevanten Schlagworte angezgit werden können. Durch semiautomatisierte, regelbasierte Ansätze werden Person-Entitäten in den Briefen annotiert. Da die Briefe sehr viel Varianz in der Rechtschreibung und sogar in der Sprache aufweisen, werden Sprachidentifikationsverfahren und Rechtschreibnormalisierungen durchgeführt. Dies stellt sicher, dass die inhärente Varianz keine Probleme für die Natural Language Processing Verfahren verursacht. Eine Zusammenfassung der Methoden findet sich unter dem Reiter „Methodology“[2] der Webseite der Edition.

2 http://ckcc.huygens.knaw.nl/?page_id=13

4

UMFRAGE: EFFIZIENTES SUCHEN UND EDITIEREN VON DOKUMENTEN FÜR NUTZER*INNEN UND EDITOR*INNEN DIGITALER (HISTORISCH-KRITISCHER) EDITIONEN DURCH NATURAL LANGUAGE PROCESSING

In diesem Kapitel werden die Ergebnisse der Umfrage „Effizientes Suchen und Editieren von Dokumenten für Nutzer*innen und Editor*innen digitaler (historisch-kritischer) Editionen durch Natural Language Processing“ (siehe Anhang A) präsentiert. Zuerst wird der Vorgang beschrieben wie Kontaktdaten von Personen der Zielgruppe gesammelt wurden, in welchem Zeitraum die Umfrage stattfand und wie hoch die Partizipation und Abbrüche waren. Danach werden die Ergebnisse dargestellt. Die Ergebnisbesprechung folgt dem Aufbau der Umfrage, sodass zuerst die Antworten zu Fragen über Suche im Allgemeinen, zur Unterstützung des Editionsprozesses, zur Präsentation der Suchergebnisse und abschließend zu bisherigen Erfahrungen/Problemen mit digitalen (historisch-kritischen) Editionen vorgestellt werden.

Die Motivation zur Umfrage besteht darin, Anstöße und Möglichkeiten zu Natural Language Processing Methoden in digitalen (historisch-kritischen) Editionen zu geben. Zusätzlich soll durch eine offene und transparente Arbeitsweise der Computerwissenschaftler*innen Prävention gegen Ablehnung moderner Technologien geleistet werden. Zusätzlich soll verhindert werden, dass die Beiträge aus den Reihen der Computerwissenschaftler*innen an den Bedrüfnissen der Editor*innen vorbeigehen.

Die Ergebnisse werden zwischen Editor*innen und Nutzer*innen unterschieden, da einerseits der Frageblock „Unterstützung des Editionsprozesses“ nur für erstere relevant ist und andererseits untersucht wird, ob und inwieweit ein Unterschied im Suchverhalten der zwei Zielgruppen besteht. Als Editor*innen werden diejenigen verstanden, die am Aufbauprozess der digitalen (historischen-kritischen) Editionen beteiligt sind. Sie transkribieren, editieren und annotieren Dokumente. Auf der anderen Seite stehen die Nutzer*innen, die nur mit dem Interface der Editionen interagieren. Sie können Dokumente suchen und lesen. Editor*innen können auch Nutzer*innen sein. Die Umfrage ordnet jedoch alle Nutzer*innen als Editor*innen ein, wenn sie schon einmal als solche tätig waren.

4.1 KONTAKTPERSONEN

Da noch kein Verteiler für Kontaktpersonen besteht, wird „A Catalogue of Digital Editions“ von Greta Franzini [12] als Grundlage herangezogen. Der Katalog ist eine Sammlung von 259 digitalen (historisch-kritischen) Editionen, welche sich durch die Definition nach Patrick Sahle [28] auszeichnen.

Da keine Kontaktdaten der beitragenden Personen zu den jeweiligen digitalen (historisch-kritischen) Editionen hinterlegt sind, werden die URLs aufgerufen und manuell nach Kontaktpersonen und E-Mail-Adressen durchsucht. Um die Menge einzugrenzen, werden nur digitale (historisch-kritische) Editionen aufgerufen, die im Katalog von Franzini entweder „GER“ oder „ENG“ in *Language* und *Website Language* beinhalten. Diesem Schema folgend werden 158 E-Mail-Adressen zu 65 verschiedenen Projekten zusammengetragen. Die Umfrage wurde am 27.07.2020 aktiv geschaltet und am selbigen Tag wurden personalisierte E-Mails an die Personen in der Kontaktliste gesendet. 17 E-Mail-Adressen waren nicht mehr gültig. Nach 17 Tagen, am 12.08.2020, wurde die Umfrage deaktiviert.

In diesem Zeitraum haben 74 Personen an der Umfrage teilgenommen, wovon 40 den kompletten Fragebogen beantwortet haben und 34 teilweise. Dies entspricht einer Bruttorücklaufquote von 52.48%, einer Nettorücklaufquote von 28.37% und einer Abbruchquote von 45.95%. Die Abbruchquote resultiert unter anderem daraus, dass beim Zusammenstellen der Kontaktliste alle E-Mail-Adressen aufgenommen wurden, die in Verbindung mit dem jeweiligen Projekt standen. Daher besteht, die Möglichkeit, dass nicht alle Personen tatsächlich der Zielgruppe angehören. So wurde z. B. von einer Person rückgemeldet, dass sie nur bei der Auswahl der Quelltexte beteiligt war und sonst wenig Berührungspunkte mit digitalen (historisch-kritischen) Editionen besitzt. Für zukünftige Umfragen sollte die Kontaktliste mit Personen erweitert werden, welche einen stärkeren Bezug zu digitalen (historisch-kritischen) Editionen aufweisen. Hierbei können Netzwerke wie DiXiT[1] oder e-editiones[2] erste Anlaufstellen sein.

4.2 ERGEBNISSE

Im Folgenden werden die Ergebnisse der Umfrage betrachtet und auf Basis aller Fragebögen (sowohl vollständige als auch unvollständige) ausgewertet. Von den oben genannten 74 vollständigen Fragebögen haben 27 Personen allerdings noch nicht einmal die erste Frage beantwortet. 7 Personen haben im weiteren Verlauf die Umfrage vorzeitig beendet.

Zu Beginn der Umfrage wird eine Filterfrage gestellt, welche eine Unterscheidung der Teilnehmer*innen in Nutzer*innen vs. Editor*innen vornimmt. Der Fokus der nachfolgenden Ergebnisbesprechung liegt auf den Antworten der Editor*innen, da sich die vorliegende Arbeit die Unterstützung dieser Gruppe als Ziel gesetzt hat. Alle gezeigten Graphiken beziehen sich ausschließlich auf die Antworten der Editor*innen und zeigen (wenn nicht anders spezifiziert) die Ergebnisse zu Multiple Choice Fragen. Die Graphiken zu den Antworten der Nutzer*innen finden sich in Anhang A.

Unter den Personen, die den Fragebogen begonnen haben, befanden sich 36 Editor*innen, 11 Nutzer*innen und 27 Personen, die gleich zu Beginn den Fragebogen abbrachen, sodass keine Informationen über ihre Gruppenzuge-

1 https://dixit.uni-koeln.de/
2 https://e-editiones.org/

	Editor*innen	Nutzer*innen	Keine Info	Gesamt
Alle	36	11	27	74
Vollständig	32	8	0	40
Abbruchquote	11.11%	27.27%	100%	45.95%

Tabelle 4.1: Kreuztabelle Fragebogenstatus nach Gruppenzugehörigkeit

hörigkeit erhoben werden konnte. Nur 4 (ca. 11%) der Editor*innen haben den Fragebogen vorzeitig beendet. Eine Übersicht der Personengruppen und ihrer Abbruchquote ist in Tabelle 4.1 dargestellt.

Die Ergebnisse der 11 bzw. 8 Nutzer*innen sollen nicht unerwähnt bleiben. Aufgrund der geringen Fallzahl sind sie jedoch mit Vorsicht zu betrachten.

4.2.1 *Allgemeine Informationen zur Person*

In Tabelle 4.2 sind allgemeine Informationen über die Teilnehmer*innen aufgeteilt in Editor*innen und Nutzer*innen enthalten. Diese Angaben wurden am Ende des Fragebogens gestellt, wodurch sich die Teilnehmer*innenbasis auf die der vollständig ausgefüllten Fragebögen bezieht. (32 Editor*innen, 8 Nutzer*innen, 40 Gesamt). Neben den demographischen Daten (Geschlecht, Alter, höchster akademischer Abschluss/Titel und Sprache), wird angegeben wie viele Jahre an Erfahrung die Teilnehmer*innen bereits mit digitalen (historisch-kritischen) Editionen gesammelt haben. Zusätzlich werden die Dokumententypen, mit denen sie in Berührung kamen, aufgelistet.

Unter den Teilnehmer*innen befinden sich ungefähr gleich viele Männer wie Frauen mit einem Durchschnittsalter zwischen 31-50 Jahren. $\frac{2}{3}$ der Editor*innen haben einen Doktortitel. Der höchste Abschluss/Titel der restlichen Teilnehmer*innen ist entweder Master oder Professor. Die höchsten Abschlüsse/Titel der Nutzer*innen, die an der Umfrage teilnahmen, sind auch entweder Master, Doktor oder Professor. Circa $\frac{2}{3}$ der Teilnehmer*innen haben den Fragebogen auf deutsch und $\frac{1}{3}$ auf englisch beantwortet.

Die Mehrheit der Nutzer*innen haben 1-10 Jahre an Erfahrung mit digitalen (historisch-kritischen) Editionen. Unter den Editor*innen sind die Jahre an Erfahrung ungefähr gleichverteilt. Es gibt jedoch nur 3 Personen mit der meisten Erfahrung (21-25 Jahre). Da die Frage nach den bereits genutzten Dokumententypen Mehrfachantworten zuließ, summiert sich die Anzahl nicht auf die Teilnehmer*innenzahl auf. Bis auf Kodex und Tafel wird viel mit allen anderen Texttypen gearbeitet.

4.2.2 *Allgemeines*

Editor*innen haben angegeben, dass sie während der kritischen Auseinandersetzen mit digitalen, historischen Dokumenten am häufigsten Personen, Orte und Abkürzungen annotieren (Abb. 4.1). Die „Sonstiges“-Angabe lässt eine große Variation an Annotationen erkennen und reicht von Named En-

Variable	Ausprägungen	Editor (32)	Nutzer (8)	Gesamt (40)
Geschlecht	Männliche	14	5	19
	Weiblich	13	2	15
	Sonstiges	1	1	2
	Keine Angabe	4	0	4
Alter	$\leq$ 30 Jahre	0	1	1
	31 - 40 Jahre	10	4	14
	41 - 50 Jahre	15	3	18
	51 - 60 Jahre	4	0	4
	$>$ 60 Jahre	1	0	1
	Keine Angabe	2	0	2
Abschluss	Bachelor	0	0	0
	Master	7	4	11
	Doktor	19	2	21
	Professor	4	1	5
	Kein Abschluss	0	0	0
	Keine Angabe	2	1	3
Sprache	Deutsch	20	6	26
	Englisch	12	2	14
Erfahrung	1 - 5 Jahre	10	3	13
	6 - 10 Jahre	7	3	10
	11 - 15 Jahre	6	0	6
	16 - 20 Jahre	6	1	7
	21 - 25 Jahre	3	0	3
	$<$1 Jahr	0	0	0
	Keine	0	1	1
Texttyp	Manuskripte	27	6	33
	Briefe	23	6	27
	Drucke	25	4	29
	Bücher	20	6	26
	Kodex	6	0	6
	Tafel	0	0	0
	Tagebücher	16	4	20

Tabelle 4.2: Kreuztabelle: Allgemeine Informationen nach Gruppenzugehörigkeit

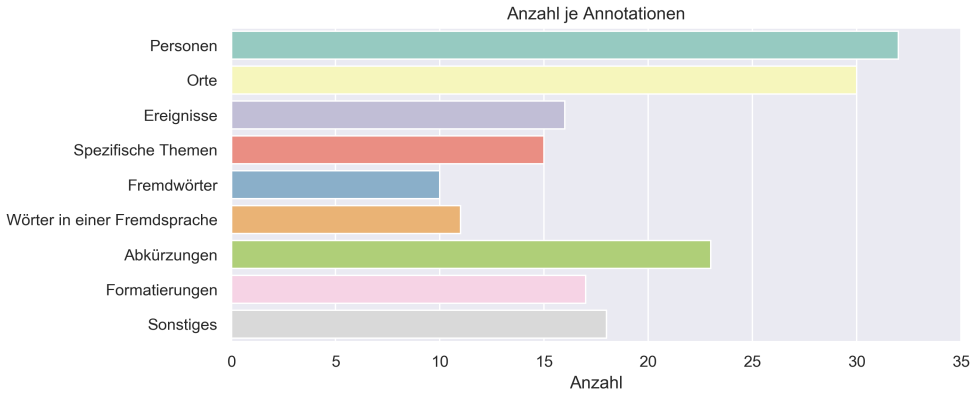


Abbildung 4.1: Annotationen von Editor*innen

tities wie Buchtiteln, Körperschaften, Organisationen über textuelle Eigenschaften wie Verschreibungen, Lücken im Text, Textvarianten bis hin zu übergreifenden verknüpfenden Informationen, einerseits von Beziehungen verschiedener Personen. Es werden aber auch Zusatzinformationen aus Primärliteratur oder geschichtliche Hintergründe genannt. Das deckt sich zu Teilen mit den Themen der Suchanfragen von Nutzer*innen, welche sich größtenteils für Personen bzw. die Beziehungen zwischen Personen interessieren. Die Mehrheit interessiert sich noch zusätzlich für spezifische Themen wie Literatur, Kunst und Politik.

Natürlichsprachliche Suche und Phrasensuche sind bei beiden Gruppen die populärste Art und Weise Suchanfragen zu formulieren. Zwar verfolgen

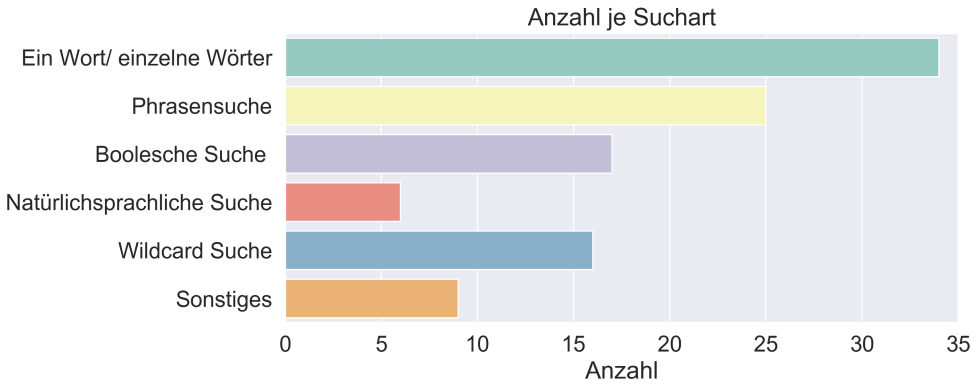


Abbildung 4.2: Formulierungsstrategien der Suchanfrage von Editor*innen

Nutzer*innen und Editor*innen die gleiche Strategie wie sie ihre Suchanfrage stellen, wünschen sich aber unterschiedliche Unterstützungsangebote. Wenn Editor*innen im Rahmen ihres Editionsprozesses Informationen in anderen Dokumenten nachlesen müssen und zum Auffinden der relevanten Dokumente die Suchfunktion nutzen, wünschen sie sich am häufigsten Autovervollständigung (Abb. 4.3). Nutzer*innen hingegen schätzen die Unterstützung durch Query Erweiterungen. Der Unterschied liegt wohl in den differenzierten Anforderungen an die Suchfunktion. Nutzer*innen er-

warten hauptsächlich, dass sie relevante Ergebnisse erhalten. Da digitalen

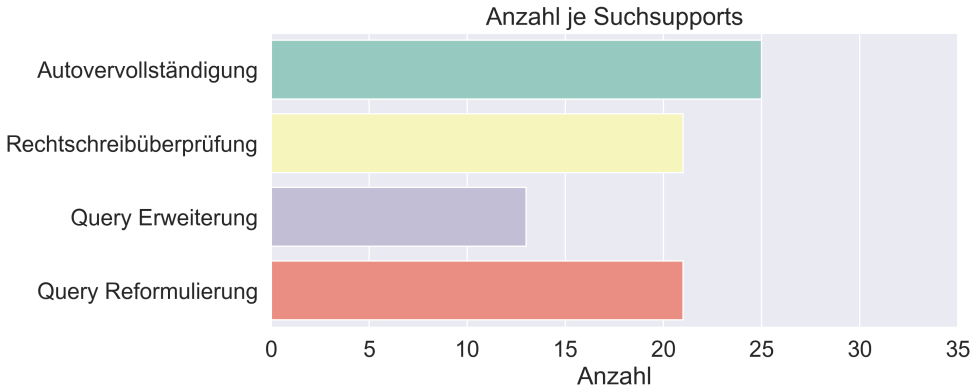


Abbildung 4.3: Unterstützungswünsche bei der Suchanfrage von Editor*innen

(historisch-kritischen) Editionen oft Texte mit spezieller (historischer oder fachlicher) Sprache zu Grunde liegen, mit der die Nutzer*innen nicht gänzlich vertraut sind, liegt eine Query Erweiterung z. B. durch Thesauri oder Querylogs nahe. Editor*innen weisen dagegen einen Effizienzanspruch auf, um sich schnellstmöglich wieder ihrer primären Arbeit, dem Annotieren und Kommentieren, widmen zu können.

Diejenigen Editor*innen mit dem Wunsch nach Query Reformulierung oder bzw. und Query Erweiterung gaben an, dass die Änderung ihrer Anfrage nicht im Hintergrund (automatisiert) ablaufen soll, sondern sie ein gewisses Maß an Kontrolle behalten wollen. Das bedeutet, dass sie manuell entscheiden wollen, wie ihre Anfrage verändert wird.

4.2.3 *Editionsprozess*

Da der Fokus der Arbeit auf den Editor*innen liegt, wurden explizit nur Editor*innen gefragt, wie sie im eigentlichen Editionsprozess unterstützt werden können. Hierbei haben die Teilnehmer*innen 5 bzw. 6 Vorschläge (ein Vorschlag ist eine Erweiterung eines anderen) erhalten (Abb. 4.4).[3]

Zum einen besteht die Möglichkeit Informationen optisch als Graph darzustellen (*Graph*). Dabei werden durch Knoten und Kanten Einheiten und Relationen miteinander in Verbindung gesetzt. Es können sowohl bereits annotierte als auch textuelle Informationen dargestellt werden. Somit erhalten Editor*innen einen schnellen und übersichtlichen Einblick in die digitale (historisch-kritische) Edition. Dieser Vorschlag zielt auf die Unterstützung bei der Suche nach Informationen aus anderen Dokumenten ab.
Genau wie bei Graphen findet bei *Topic Modelling* eine Zusammenfassung/ Kategorisierung von Informationen statt und dient der Unterstützung des Suchprozesses. Im Gegensatz zum graphischen Ansatz, werden hierbei kei-

3 Angaben in dieser Graphik sind in Prozent, da die Ergebnisse aus 6 Einzelfragen bestehen und die Fallzahl der Antworten unterschiedlich sein kann, da die Fragen nicht obligatorisch waren.

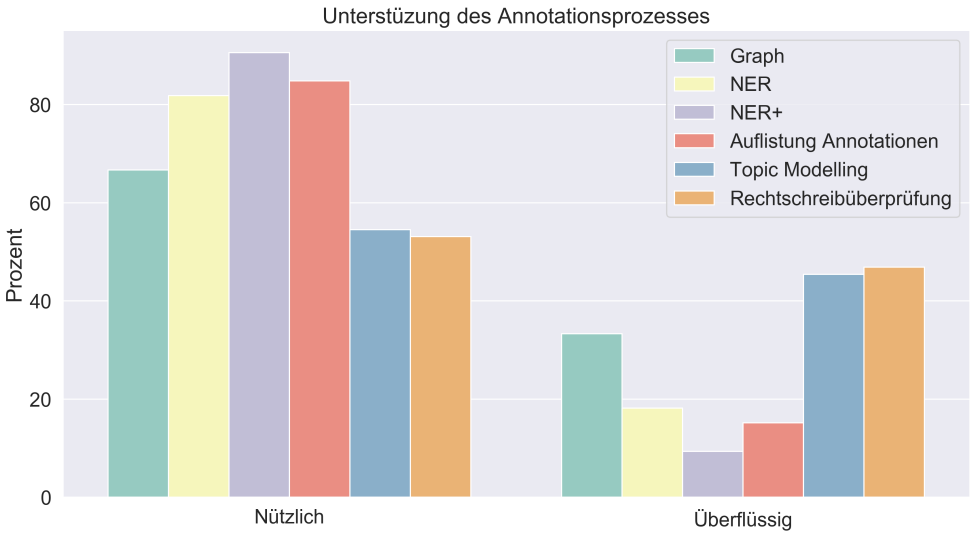


Abbildung 4.4: Relative Häufigkeit der Unterstützungswünsche des Annotationsprozesses

ne Relationen dargestellt, sondern thematische Überbegriffe aus den Dokumenten aufgestellt.

Speziell auf Briefeditionen abzielend, wurde die Option *Auflistung Annotation* aufgezeigt. Hierbei wird die Möglichkeit angeboten, alle Annotation aus den Korrespondenzstücken mit dem/der aktuellen Korrespondenzpartner/-partnerin aufzulisten. Zusätzlich können die Briefe durch Anklicken aufgerufen werden.

Editor*innen können nicht nur bei der Suche nach Informationen aus anderen Dokumenten unterstützt werden, sondern auch während der Transkription und der Annotation des aktuellen Dokuments. Viele Dokumente besitzen aufgrund ihres historischen Ursprungs eine andere Sprachform und somit eine andere Rechtschreibung als unsere Moderne. Deshalb muss bei der Transkription darauf geachtet werden, nicht aus Gewohnheit in die eigene Rechtschreibung zu verfallen und dadurch Rechtschreibfehler in der Transkription zu machen. Dies könnte verhindert werden, indem eine *Rechtschreibüberprüfung* basierend auf der Orthographie gegeben der zeitlichen Einordnung der Dokumente eingebaut wird.

Der Annotationsprozess kann mittels Named-Entity-Recognition (*NER*) unterstützt werden. Bei der Named-Entity-Recognition werden die zu annotierenden Entitäten automatisch erkannt und gelabelt. Zusätzlich besteht die Möglichkeit bei bereits abgespeicherten Annotationen, diese in neuen Dokumenten zu erkennen und gleichzeitig - wenn vorhanden - mit den dazugehörigen Zusatzinformationen/ Kommentaren anzureichern (*NER+*).

Wie in Abbildung 4.4 zu erkennen, ist für die meisten Editor*innen die Unterstützung des Annotationsprozesses durch Named-Entity-Recognition ein Anliegen. In Verbindung damit wünschen sie sich auch die Informationsanreicherung bereits hinterlegter Daten zu den annotierten Entitäten.

Die Auflistung aller Annotationen in Korrespondenzstücken mit dem/der Korrespondenzpartner*in ist der dominierende Wunsch zur Unterstützung der Suche nach aufschlussreichen Informationen aus anderen Dokumenten.

Für Editor*innen scheinen Informationen, die bereits von anderen Editor*innen validiert sind und klare Informationsangaben geben, von größerem Wert zu sein als Informationen, die nur in schwammiger Form vorliegen und noch eine Transferleistung benötigen. Das unterstreicht die These, dass Editor*innen einen Anspruch an Effizienz hegen.

Von geringerer Bedeutung scheint die Möglichkeit der Rechtschreibüberprüfung und die Topic Modellierung zu sein.

4.2.4 *Ergebnispräsentation*

Sowohl Editor*innen als auch Nutzer*innen möchten alle relevanten Ergebnisdokumente, die auf ihre Suchanfrage zutreffen, angezeigt bekommen (Abb. 4.5). Diese Fragestellung ist in Anbetracht der Ergebnisse subopti-

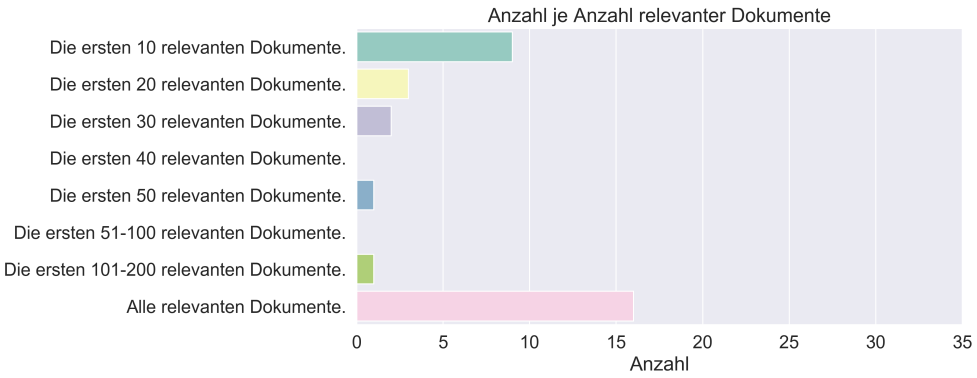


Abbildung 4.5: Anzahl gewünschter Ergebnisse (Single Choice Frage)

mal formuliert, da „relevant“ ein vager Begriff ist. In Information Retrieval-Systemen mit Booleschem Retrieval ist es möglich, die Dokumente binär in relevant und nicht-relevant zu klassifizieren. In Information Retrieval-Systemen mit sortiertem (ranked) Retrieval ist dies jedoch nicht eindeutig. Daher wäre eine Vorgehensweise, die Anzahl der Ergebnisdokumente abhängig von der Art der Suchanfrage festzulegen. Entsprechend der Realisierbarkeit sollten alle relevanten Dokumente angezeigt werden und bei sortierten Ergebnissen könnten die Editor*innen die ersten 10 relevanten Dokumente (2. häufigste Antwortkategorie) und die Nutzer*innen die ersten 50 relevanten Dokumente angezeigt bekommen.

Den beiden Gruppen ist es wichtig, dass es die Möglichkeit gibt, präzisere Einschränkungen und eigenständige Sortierung der Reihenfolge (Abb. 4.6) durch die Auswahl des/der Empfänger*in und/oder der/des Absenders/in und des Datums zu setzen (Abb.4.7 und Abb. 4.8).
Diejenigen, die sich eine Filtermöglichkeit nach Annotationen wünschen, erachten folgende Annotationen als wichtig: Person, Ort, Örtlichkeit, Bibliogra-

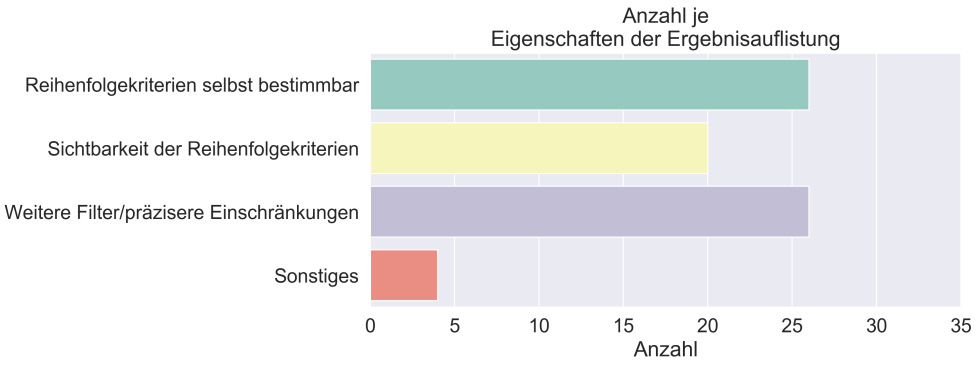


Abbildung 4.6: Eigenschaften der Ergebnisauflistung

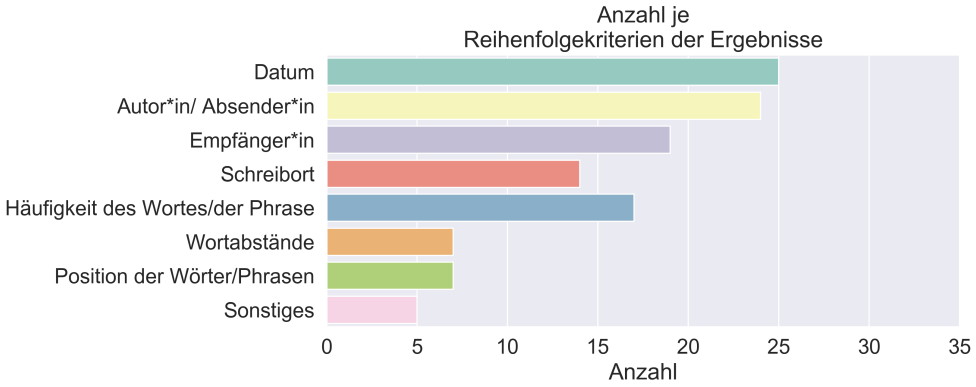


Abbildung 4.7: Attribute zur Bestimmung der Reihenfolge

phien, intertextuelle Bezüge und Hintergrundinformationen, die zum Textverständnis beitragen.

4.2.5 *Erfahrungen*

4.2.5.1 *Suchfunktion*

Positive Erfahrungen mit der Suchfunktion in bereits genutzten digitalen (historisch-kritischen) Editionen wurden bei Nutzer*innen durch eine einfache Handhabung, effiziente Stichwortsuchen, flexible und kombinierbare Suchoptionen und Volltextsuche gesammelt. Editor*innen ist es wichtig auf verschiedene Arten und Weisen suchen zu können: einzelne Wörter, Phrasensuche und Volltextsuche. Außerdem wird die Unterscheidung zwischen einer simplen und einer facettierten Suche positiv hervorgehoben. Nicht nur, die Unterscheidung der Suchfunktionen, sondern auch die Möglichkeit zur Unterscheidung zwischen der Suche im ganzen Katalog oder in einzelnen Dokumenten ist Editor*innen wichtig. Dabei ist auch eine Suche in XML-Dateien gefragt. Als hilfreiche Unterstützungsmöglichkeit heben Editor*innen Autocomplete hervor.

Probleme mit digitalen (historisch-kritischen) Editionen, welche die Befragten bisher genutzt haben, gab es sowohl bei Editor*innen als auch Nut-

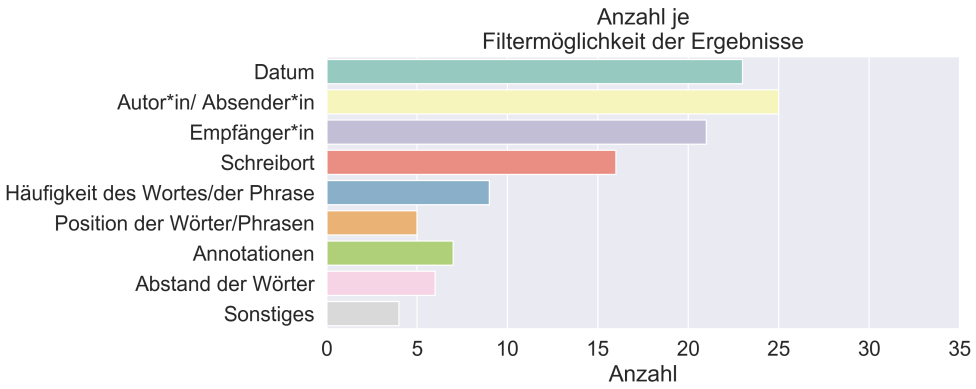


Abbildung 4.8: Attribute zur Filterung

zer*innen bei der Suche mit Sonderzeichen. Zusätzlich erwähnten manche Editor*innen, dass die zugrundliegenden Daten wie z. B. die TEI-Dateien nicht zugänglich waren.

4.2.5.2 *Ergebnispräsentation*

Nutzer*innen erwähnen hierbei im Besonderen das Vorhandensein von Faksimiles und damit zusammenhängend die Verfügbarkeit von Zoomoptionen und hybriden Volltext-Graphiken. Es ist ihnen wichtig, dass die Ergebnisse klar hervorgehoben werden. Bei dieser Antwort ist allerdings nicht ganz klar, ob die Ergebnisdokumente als Ganzes oder ob einzelne Textstellen innerhalb der Dokumente gemeint sind.

Eine übersichtliche Ergebnisdarstellung mit klaren Angaben zu Kriterien, der Reihenfolge und die Möglichkeit zur eigenständigen Sortierbarkeit ist Editor*innen wichtig. Die Ergebnisse sollen als Listen oder Tabellen ausgegeben werden. Es erwähnten auch einige, dass sie es sehr hilfreich finden, wenn die relevanten Textstellen ausschnittsweise bei der Ergebnisauflistung angezeigt werden.

4.2.6 *Bivariate Zusammenhänge*

Eine bivariate Betrachtung der Antworten wurde auf Basis der allgemeinen Informationen zur Person vorgenommen. Aufgrund der geringen Fallzahl in den einzelnen Kategorien können jedoch keine validen Aussagen getroffen werden.

4.3 FAZIT

Die Umfrage deckt verschiedene Themenkomplexe ab. Editor*innen annotieren am häufigsten Personen, Orte und Abkürzung. Die ersten zwei Entitäten decken sich mit den Interessen der Nutzer*innen.

Die gängigsten Formen der Suche sind die Phrasensuche und die Natürlichsprachliche Suche, wobei Editor*innen hierbei Autovervollständigung

und Nutzer*innen Vorschläge zu Query Erweiterung wünschen. Letztere erwarten vor allem eine einfache und intuitive Handhabung der Suchfunktion, wohingegen Editor*innen überwiegend ein hohes Ausmaß an Flexibilität, z. B. in Form einee Facettesuche, fordern.

Es ist beiden Gruppen wichtig, dass die Ergebnisse durch bestimmte Kriterien wie Empfänger*in, Absender*in oder Datum filterbar bzw. sortierbar sind. Es wird vor allem positiv hervorgehoben, wenn die Ergebnisse klar und übersichtlich dargestellt sind und Faksimiles zur Verfügung stehen.

Die Ergebnisse zum Themenkomplex des Editionsprozesses zeigen, dass Interesse und Bedarf an verschiedenen Unterstützungstools besteht. Einerseits kann Editor*innen schon geholfen werden, indem eine reine Auflistung der bereits annotierten Entitäten stattfindet, sodass das Auffinden von Zusatzinformationen erleichtert wird. Andererseits können gezielte Natural Language Processing Methoden die Editor*innen in ihrem Annotationsprozess unterstützen. Da über 80% der Teilnehmer*innen einen Mehrwehrt in Named-Entity-Recognition System sehen, wird dieser Wunsch umgesetzt und im nachfolgenden Kapitel beschrieben.

5

NAMED ENTITY RECOGNITION ZUR UNTERSTÜTZUNG DES ANNOTATIONSPROZESSES IN DER FRANK WEDEKIND BRIEFEDITION

5.1 DATENSATZ

Im Rahmen des Projekts „Edition der Korrespondenz Frank Wedekinds als Online-Volltextdatenbank“ werden Dokumente von und an Frank Wedekind analysiert. Es sind 1150 Dokumente in HTML und Plaintext (inkl. Kommentare der Editor*innen) verfügbar, wovon 650 Briefe und 316 Postkarten sind. Die Dokumente wurden im Zeitraum von 1882-1918 geschrieben. 597 Dokumente wurden von Wedekind selbst verfasst und 553 Dokumente sind an ihn adressiert.

5.2 VORVERARBEITUNG

Nach dem Import der *wedeojpa_20200628-011101.dump* Datei in eine PostgreSQL Datenbank werden die Spalten *documentid*, *documentcontentcleaned* und *documentcontenthtml* der Tabelle *document* zur weiteren Bearbeitung verwendet. Das Testdokument[1], weitere 22 Dokumente[2], die in *documentcontentcleaned* und *documentcontenthtml* None-Werte aufweisen, 2 Dokumente[3], deren Inhalt nur aus Zeilenumbrüchen besteht und noch 4 Dokumente[4], deren Inhalte in *documentcontentcleaned* und *documentcontenthtml* nicht übereinstimmen, werden ausgefiltert. Danach stehen 1150 Dokumente zur Analyse zur Verfügung. Der gesamte Workflow ist vereinfacht in Abbildung 5.1 dargestellt.

Die Daten liegen goldstandard annotiert in den HTML Dokumenten vor. Abbildung 5.2 zeigt beispielhaft den Code eines HTML Dokumentes, das als Plaintext wie in Abbildung 5.3 vorhanden ist und in der Frank Wedekind Briefedition wie in Abbildung 5.4 angezeigt wird.

Aus den HTML Dokumenten werden zwei Informationen gezogen. 1. Entitätsreferenzen und korrespondierende Entitäten und 2. Kommentare der Editor*innen.

1 *documentid*=350
2 *documentid*=[178, 175, 156, 147, 361, 459, 276, 286, 152, 148, 150, 157, 161, 269, 149, 155, 159, 151, 158, 154, 160, 153]
3 *documentid*= [146, 360]
4 *documentid*=[1099, 995, 0, 592]

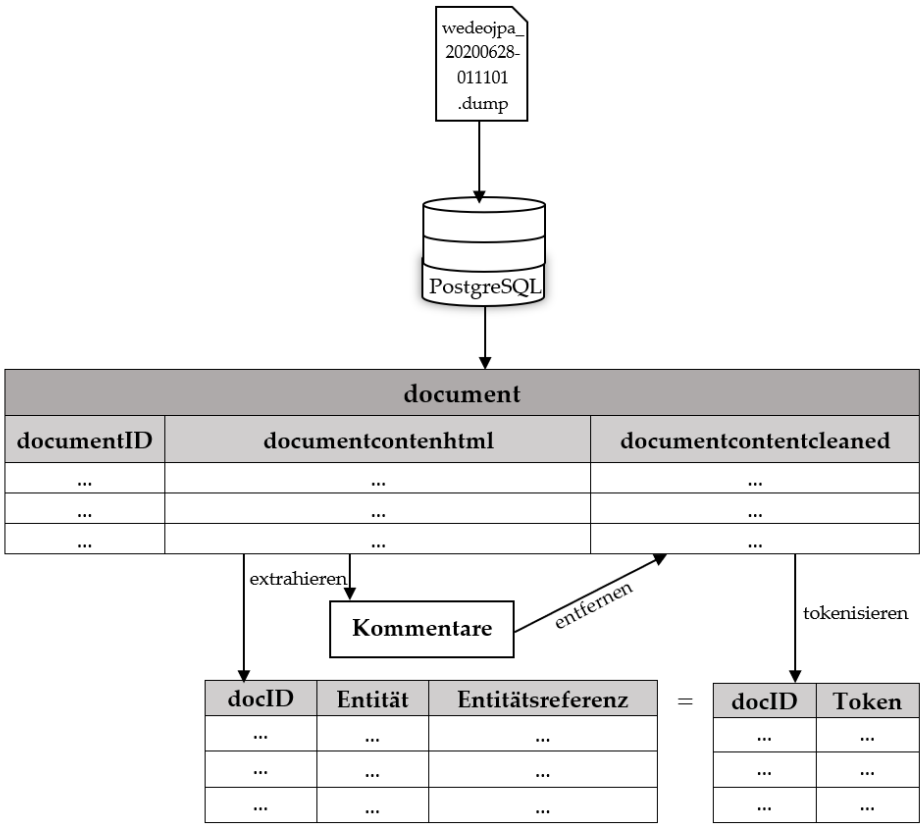


Abbildung 5.1: Workflow

<placename class="tei" data-tei="<placeName type=\'city\' key=\'siteCity_131111137\' synch=\'10550529\'>$1</placeName>">Paris</placename>\r\n15.9.94.
<hi class="tei hi-font-latin" data-tei="<hi rend=\'latintype\'>$1</hi>"><site_briefed class="tei" data-tei="<site_briefed type=\'site\' key=\'site_149\' synch=\'10550530\'>$1</site_briefed>">45\r\nrue Monsieur le Prince</site_briefed>.</hi>
 Lieber Herr <persname class="tei" data-tei="<persName key=\'person_482\' synch=\'10550524\'>$1</persName>">Hartleben</persname>, ich vermuthe daß <comment class="wka709">mein letzter\r\nBrief</comment><note class="tei" data-tei="<note type=\'commentary\' resp=\'Editor\'>$1</note>">nicht überliefert; erschlossenes Korrespondenzstück: Wedekind an Hartleben, 1.9.1894.</note> nicht bis zu Ihnen gelangt <hi class="tei" data-tei="<hi rend=\'strikethrough\'>$1</hi>"><s>se</s></hi> ist. Ich bat Sie dringend\r\ndarum mir doch die beiden Manuscripte „<comment class="wke566"><title_briefed class="tei" data-tei="<title_briefed key=\'work_79\' synch=\'10550531\'>$1</title_briefed>">Schwiegerling</title_briefed></comment><note class="tei" data-tei="<note type=\'commentary\' resp=\'Editor\'>$1</note>">Wedekind hatte Hartleben im Herbst 1893 eine Abschrift seines Stückes „Fritz Schwigerling“ (1892) geschickt [vgl. die erschlossene Korrespondenzstücke Wedekind an Hartleben, 25.10.1893 und Hartleben an Wedekind vom 25.4.1894].</note>“ und „<title_briefed class="tei" data-tei="<title_briefed key=\'work_193\' synch=\'10550532\'>$1</title_briefed>"><comment class="wkae57">Flirt</comment><note class="tei" data-tei="<note type=\'commentary\' resp=\'Editor\'>$1</note>">Hartleben hatte Wedekinds im Frühjahr 1894 entstandene Erzählung „Flirt“ Otto Neumann-Hofer für die von diesem herausgegebenen Zeitschriften „Magazin für Litteratur“ und „Romanwelt“ angeboten, der sie jedoch ablehnte [vgl. KSA 5/1, S. 622].</note></title_briefed>“\r\nhierherschicken zu wollen. Ich adressiere diese Zeilen <comment class="wk6e96">an <persname class="tei" data-tei="<persName key=\'person_997\' synch=\'10550522\'>$1</persName>">S. Fischer</persname></comment><note class="tei" data-tei="<note type=\'commentary\' resp=\'Editor\'>$1</note>">Hartlebens Verleger Samuel Fischer firmierte seit Dezember 1891 unter der Geschäftsadresse Berlin W, Köthener Straße 44.</note>, da ich\r\nmich frage, ob Sie vielleicht Ihre <comment class="wkb384">Wohnung</comment><note class="tei" data-tei="<note type=\'commentary\' resp=\'Editor\'>$1</note>">Berlin, N. W., Karlstraße 32, III. Stock.</note> geändert haben. Für Schwiegerling |\r\nkann ich hier eventuell Verwendung finden. Ich habe die denkbar günstigste\r\n<comment class="wkc425">Gelegenheit</comment><note class="tei" data-tei="<note type=\'commentary\' resp=\'Editor\'>$1</note>">nicht ermittelt.</note>. Darf ich Ihnen also meine Bitte noch einmal dringend wiederholen. <hi class="tei hi-font-latin" data-tei="<hi rend=\'latintype\'>$1</hi>"><persname class="tei" data-tei="<persName key=\'person_1001\' synch=\'10550525\'>$1</persName>">Kainach</persname></hi> wird wie ich durch meinen <persname class="tei" data-tei="<persName key=\'person_352\' synch=\'10550526\'>$1</persName>">Bruder</persname> höre in den nächsten\r\nWochen hier eintreffen. <hi class="tei hi-font-latin" data-tei="<hi rend=\'latintype\'>$1</hi>">Dr</hi> <persname class="tei" data-tei="<persName key=\'person_998\' synch=\'10550523\'>$1</persName>">Goldmann</persname> habe ich noch nicht wieder\r\ngesehen. Seien Sie also bitte so liebenswürdig und | gewähren Sie mir meine\r\nBitte. Sonst geht die Gelegenheit vorüber und ich habe das Nachsehen. \r\n\r\n Ihnen und Ihrer Frau <persname class="tei" data-tei="<persName key=\'person_999\' synch=\'10550527\'>$1</persName>">Gemahlin</persname>\r\nmeine herzlichsten Grüße.
Ihr
<persname class="tei" data-tei="<persName key=\'person_7\' synch=\'10550528\'>$1</persName>">Frank Wedekind</persname>. \r\n \r\n

Abbildung 5.2: HTML Codebeispiel von Dokument *documentid*=691

'Paris 15.9.94. \n45 rue Monsieur le Prince . \n\n\nLieber Herr Hartleben , \nich vermuthe daß mein letzter Brief nicht bis zu Ihnen gelangt se ist. Ich bat Sie dringend darum mir doch die beiden Manuscripte „ Schwiegerling “ und „ FlirtHartleben hatte Wedekinds im Frühjahr 1894 entstandene Erzählung „Flirt“ Otto Neumann-Hofer für die von diesem herausgegebenen Zeitschriften „Magazin für Litteratur“ und „Romanwelt“ angeboten, der sie jedoch ablehnte [vgl. KSA 5/1, S. 622]. “ hierherschicken zu wollen. Ich adressiere diese Zeilen an S. Fischer , da ich mich frage, ob Sie vielleicht Ihre Wohnung geändert haben. Für Schwiegerling | kann ich hier eventuell Verwendung finden. Ich habe die denkbar günstigste Gelegenheit . Darf ich Ihnen also meine Bitte noch einmal dringend wiederholen. Kainach wird wie ich durch meinen Bruder höre in den nächsten Wochen hier eintreffen. Dr Goldmann habe ich noch nicht wieder gesehen. Seien Sie also bitte so liebenswürdig und | gewähren Sie mir meine Bitte. Sonst geht die Gelegenheit vorüber und ich habe das Nachsehen. \nIhnen und Ihrer Frau Gemahlin meine herzlichsten Grüße. \nIhr \nFrank Wedekind . \n'

Abbildung 5.3: Plaintext Codebeispiel von Dokument *documentid*=691

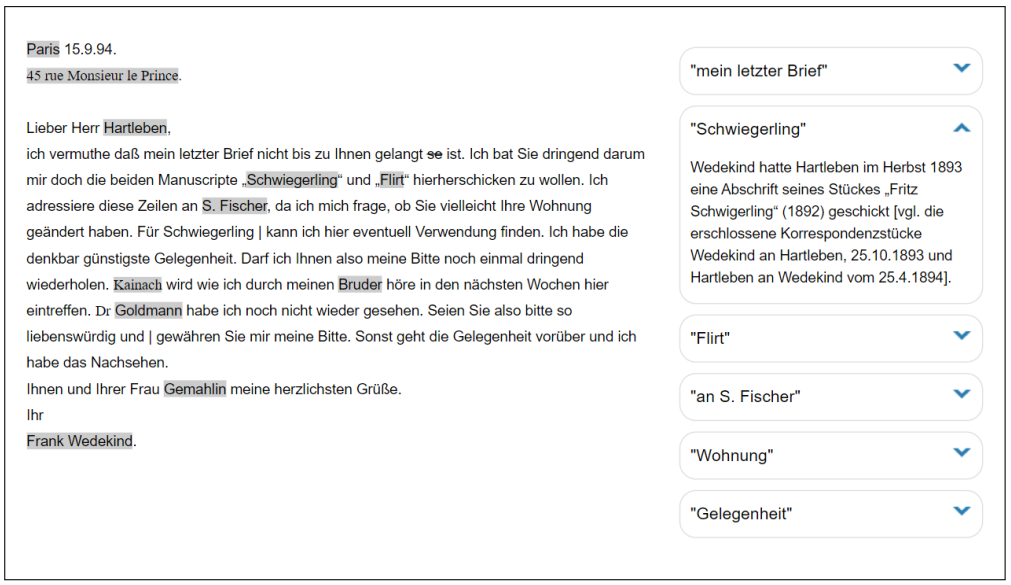


Abbildung 5.4: Darstellung des Dokuments *documentid*=691 in der Edition

5.2.1 *Vorverarbeitung von documentcontentcleaned*

Der erste digitale Schritt des Workflows der Editor*innen ist das Transkribieren der Briefe. Danach folgen Annotationen und Kommentierungen. An dieser Stelle setzt die vorliegende Arbeit an. Für das Training der Named-Entity-Recognition werden die Dokumente in ihrer Rohform (ohne Kommentare) benötigt. Die Briefe sind in *documentcontentcleaned* mit Kommentaren hinterlegt. Nicht alle Kommentare, die in *documentcontenthtml* als solche getaggt sind, wurden in *documentcontentcleaned* entfernt.[5] Kommentare, die noch enthalten sind, werden mittels Regex-Ausdruck aus *documentcontenthtml* gefiltert und die zurückgelieferten Kommentare werden aus *documentcontentcleaned* entfernt. Bei einigen Kommentaren war zusätzliches Entfernen von Zeilenumbrüchen und doppelten Leerzeichen notwendig, um

5 Kommentare, die nicht übernommen werden, besitzen folgende Form: <comment class="*XXX*»<*tagname_start* class="tei"data-tei=«*tagname_start* key='*XXX*' synch='*XXX*'>$1</*tagname_start*>">*annotation_text*</*tagname_end*></comment><note class="tei"data-tei=«note type='commentary' resp='*editor*'>$1</note>»*comment*</note>

oder: <comment class="*XXX*»*annotation_text*</comment><note class="tei"data-tei=«note type='commentary' resp='*editor*'>$1</note>">*comment* </note>

```
"""(?<=<\/comment><note class="tei" data-tei="<note type=\'
    commentary\' resp=\'(?:\w* ?)*\'>\$1<\/note>">)(?:.|\n)
    *?(?=<\/note>)"""
```

eine Übereinstimmung mit den Kommentaren in *documentcontentcleaned* zu erzielen.

5.2.2 *Extrahieren der Annotationen aus documentcontenthtml*

Es werden folgende Annotationen aus den HTML-Dokumenten extrahiert: Person (PER), Ort (ORT), Örtlichkeit (OTK), Werk (WRK) und Ereignis (EVT). Hierfür werden Regex-Ausdrücke verwendet, wobei zwischen den Annotationen mit und ohne zusätzlichen Kommentar durch die Editor*innen differenziert werden muss, da sich diese in ihrem HTML-Aufbau unterscheiden. Mittels OR-Verknüpfung wird sichergestellt, dass nach beiden Ausdrücken gesucht wird.

```
regex_category = '''(?<=<\/tagname_start>">)
                    (?:\w* ?)*
                    (?=<\/tagname>)'''
regex_category_comment = '''(?<=<\/tagname_end>"><comment class="\w*">)
                            (?:\w* ?)*
                            (?=<\/comment>)'''

regex_category_result = '''({})|({})'''.format(regex_category,
                                               regex_category_comment)
```

Die Annotationen Örtlichkeit, Werk und Ereignis besitzen denselben *tagname_start* wie *tagname_end*: site_briefed, title_briefed, event_briefed. Bei den Annotation von Personen und Orten ist *tagname_start* = persName bzw. placeName und für *tagname_end* persname bzw.placename. Die extrahierten Entitäten werden in das IOB2-Format überführt, sodass Entätsreferenzen, die eine Länge von einem Token haben mit der Entität B-XXX, gekennzeichnet werden. Entitätsreferenzen, die sich über mehrere Token zusammensetzen werden so gelabelt, dass der erste Token die Entität B-XXX erhält und alle weiteren mit I-XXX versehen werden. B steht für „Beginning“ und I steht für „Inside“. Token, die keiner Entität zuzuordnen sind, werden mit O für „Other“ markiert.

5.2.3 *Zuordnung der Entitäten aus documentcontenthtml zu den Entitätsreferenzen in documentcontentcleaned*

Nach der Bereinigung der Dokumente, können diese tokenisiert, d.h. in kleinere Einheiten (hier Wörter), zerlegt werden. Hierfür wird das deutsche Modell „de_core_news_sm“ von spaCy verwendet. SpaCy nutzt Wikipedia-

	SpaCy			TreeTagger		
	Dokumente	Sätze	Token	Dokumente	Sätze	Token
Training	736	12874	176973	736	11205	172349
Validierung	184	3176	42574	184	2928	41604
Test	230	4437	59563	230	3910	58003
Gesamt	1150	20487	279110	1150	18043	271956

Tabelle 5.1: Anzahl der Briefe, Sätze und Token je Datensatz (Schema 1)

	Training	Validierung	Test	Gesamt	
B-PER	2028 (2641)	538 (685)	749 (961)	3315 (4287)	4128 (5416)
I-PER	482 (696)	140 (172)	191 (261)	813 (1129)	
B-ORT	713 (995)	167 (227)	252 (359)	1132 (1581)	1156 (1612)
I-ORT	14 (20)	3 (3)	7 (8)	24 (31)	
B-OTK	248 (309)	46 (50)	80 (91)	374 (450)	564 (687)
I-OTK	128 (170)	19 (21)	44 (46)	190 (237)	
B-WRK	389 (508)	82 (94)	142 (204)	613 (806)	899 (1401)
I-WRK	174 (377)	27 (39)	85 (179)	286 (595)	
B-EVT	60 (114)	16 (17)	18 (21)	94 (152)	118 (180)
I-EVT	19 (23)	2 (2)	3 (3)	24 (28)	
	4254 (5853)	1040 (1310)	1571 (2133)	6865 (9296)	

Tabelle 5.2: Anzahl der Entitäten je Datensatz nach Schema 1 (Schema 2), spaCy

Einträge und Zeitungsartikel, um die Verfahren für das deutsche Paket zu trainieren. Da diese in einer standardisierten deutschen Rechtschreibung verfasst sind, wird zur Sicherheit und als Vergleichsmöglichkeit der in der Literatur oft verwendete TreeTagger[6] verwendet, der sowohl für moderne Sprachformen als auch für Mittelhochdeutsch verfügbar ist. Der Nachteil des TreeTaggers ist, dass er kein automatisches Chunking vornimmt, sodass selbst ein Satzsplitting implementiert werden muss. SpaCy auf der anderen Seite führt keine Trennung von Substantivkomposita mit Bindestrich als Verbindung der Wortbestandteile durch. Editor*innen annotieren nur das erste Nomen von substantivischen Zusammensetzungen, deshalb ist im Rahmen der Tokenisierung eine Trennung am Bindestrich wichtig und wird zusätzlich implementiert. Ein Vergleich (Tabelle 5.1) der beiden Pakete zeigt, dass der TreeTagger den Datensatz in 800 Token mehr unterteilt. Das sind im Schnitt 1.5 Token mehr pro Dokument. Wenn die Top 10 PoS-Tags betrachtet werden (siehe Anhang B), lässt sich auch kein bedeutender Unterschied feststellen.

Während des Mergens der Token aus *documentcontentcleaned* mit den extrahierten Entitätsreferenzen und den korrespondierenden Entitäten der HTML-Dateien, gibt es zwei Möglichkeiten, wenn dieselbe Entitätsreferenz mehrmals in einem Dokument vorkommt. Entweder es werden alle Entitätsreferenzen in *documentcontentcleaned* annotiert oder immer nur diejenigen, die als erstes im Text stehen und wiederholende Entitätsreferenzen bekommen das Label „Other“. Letzteres Schema ist konsistent mit dem Vorgehen der Editor*innen und soll dadurch Ziel der Named-Entity-Recognition Anwendung sein. Allerdings kann es je nach Algorithmus im Training zu einer Verringerung der Performanz führen, wenn der Algorithmus das Schema nicht erkennen kann. Deshalb werden beide Annotationsschemata (1. Annotation aller Entitätsreferenzen und 2. Annotation der ersten Entitätsreferenz) getestet und verglichen. Tabelle 5.2 gibt einen Überblick über die Anzahl der Entitäten für Schema 1 und in Klammern die Anzahl der Entitäten für Schema 2.

Insgesamt gibt es ca. 9300 Token, die auf eine Entität verweisen bzw. ca. 7000 Token, die tatsächlich als solche von den Editor*innen annotiert wurden. Die am häufigsten vertretenen Entitäten sind Personen mit einem deutlichen Abstand zu Orten. Fortlaufende „I-XXX“ Entitäten sind konsequenterweise seltener vertreten als die korrespondierenden „B-XXX“ Entitäten. „I-ORT“ und „I-EVT“ sind mit jeweils 24 Referenzen sehr rar. Das wirkt sich auch auf die Ergebnisse des Named-Entity-Recognition Modells aus.

6 https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

Variable	Beschreibung	Typ
word.lower()	Lowercase Version	string
word[-3:]	Prefix der Länge 3	string
word[3:]	Suffix der Länge 3	string
word.length	Anzahl der Zeichen	int
word[0].isupper()	Anfangsbuchstabe groß	boolean
word.isdigit()	Wort besteht aus Zahlen	boolean
postag	Part-of-Speech Tag	string
bos	Satzanfang	boolean

Tabelle 5.3: Feature Auswahl für CRF

5.3 CONDITIONAL RANDOM FIELD IMPLEMENTIERUNG

Regelbasierte Ansätze zu Named-Entity-Recognition Aufgaben liefern zwar oft gute Ergebnisse, sind allerdings im spezifischen Usecase schwer einzusetzen, da es aufgrund der veralteten Sprachform Expert*innen wissen bedarf. Zusätzlich entstehen Probleme durch die Varianz in der Orthographie. Dadurch und aufgrund der Adaptierbarkeit für andere digitale (historisch-kritische) Editionen wird ein Machine Learning Verfahren angewendet. Da bereits annotierte Daten vorliegen können supervised Machine Learning Algorithmen verwendet werden, um ein Named-Entity-Recognition System zu erstellen. Viele Systeme basieren zwar auf Maximum Entropy Markov Modellen, diese besitzen aber das Label Bias (vgl. Abschnitt 2.4.2.2) Problem. Aus diesem Grund wird ein Conditional Random Field, das eine Erweiterung des Maximum Entropy Markov Modells darstellt, gewählt. Die Laufzeit ist zwar höher, da eine globale und keine lokale Normalisierung durchgeführt wird, reduziert aber gleichzeitig den Fehler und erzielt bessere Ergebnisse.

Unter Verwendung der built-in Funktion *crfsuite* von *sklearn* wird ein Conditional Random Field mit Python 3.7.1 implementiert. Gegeben einer Wortsequenz eines Satzes: [... ,w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, ...], wobei w_i das aktuell verarbeitete Wort darstellt, werden die verwendeten Feature in Tabelle 5.3 dargestellt. Für jedes w_i werden Informationen vorangegangener und zukünftiger Wörter innerhalb eines $\pm$ 2 Fensters verwendet. Für diese Kontextwörter werden die Feature word.lower(), word[0].isupper() und die korrespondierenden postags genutzt.

5.4 ERGEBNISSE

Die Ergebnisse werden auf Basis von Precision, Recall und dem F_1-Score evaluiert. Es werden verschiedene Änderungen an dem Baselinemodell (Modell 1) vorgenommen, mit der Intention, Verbesserungen der Vorhersagen zu erzielen. Die Tabelle 5.4 gibt eine Übersicht der Modelle und deren erzielten Ergebnissen, wobei Y für die Testdaten und $\hat{Y}$ für dessen prognostizierte

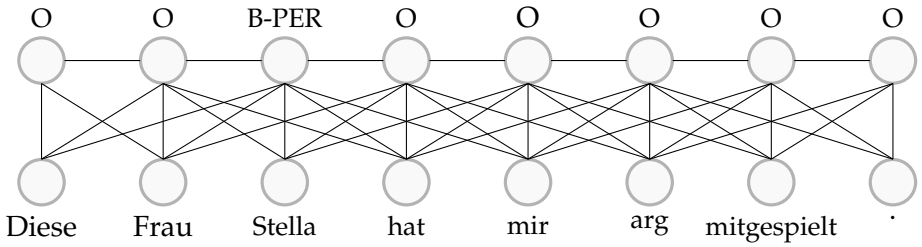


Abbildung 5.5: Graph CRF; Abhängigkeit des Vorhersagewertes y_i durch $x_i \pm 2$

Modell	Y		$\hat{Y}$			Precision	Recall	F_1
	GS	Alle	GS	Alle	_aufbereitet			
1	✓	-	✓	-	-	.629	.453	.521
2	✓	-	✓	-	✓	**.724**	.467	.563
3	-	✓	-	✓	-	.687	.516	.578
4	✓	-	-	✓	-	.535	**.537**	.528
5	✓	-	-	✓	✓	.677	.556	**.605**

Tabelle 5.4: Übersicht Ergebnisse der verschiedenen Modelle

Werte stehen. Hierbei wird jeweils eine Unterscheidung zwischen „GS“ und „Alle“ vorgenommen. GS steht für Goldstandard und gibt die Label so wieder, wie sie von den Editor*innen annotiert wurden. Wie in Abschnitt 5.2 erklärt, annotieren die Editor*innen immer nur die erste Entitätsreferenz bei Mehrfachnennungen in einem Dokument. „Alle“ steht in der Tabelle dafür, dass nicht nur die erste Entitätsreferenz gelabelt wird, sondern, dass jedes Vorkommen mit der dazugehörigen Entität versehen wird. Die Spalte „_aufbereitet“ gibt an, ob die Vorhersagewerte $\hat{Y}$ nach der Prognose noch zusätzlich aufbereitet wurden. Im Rahmen der Aufbereitung wurden die Entitätsreferenzen mit Mehrfachnennungen im Dokument untersucht. Wenn min. eine Entitätsreferenz mit einer Entität gelabelt wurde, wurde dieses Label verwendet, um das erste Vorkommen der Entitätsreferenz damit zu versehen. Alle anderen Duplikate wurden auf „Other“ gesetzt. Dieses Schema folgt dem Annotationsmuster der Editor*innen.

Es wurden 5 verschiedene Modelle untersucht. Modell 1 stellt das Baselinemodell dar, bei dem die Texte nach dem Annotationsschema der Editor*innen gelabelt wurden. Sowohl die Trainings-, als auch die Testdaten folgen diesem Schema. Modell 2 erweitert Modell 1 insofern, dass die Vorhersagen zusätzlich aufbereitet werden, bevor das Modell getestet wird. Modell 3 unterscheidet sich vom Baselinemodell durch das Annotationsschema. Hier wird nicht nur die erste Entitätsreferenz annotiert, sondern alle. Dies gilt sowohl für die Trainings- als auch Testdaten. Modell 4 stellt eine Modifikation von Modell 3 dar. Da das Named-Entity-Recognition Modell dem Annotationsschema der Editor*innen folgen soll, werden in Modell 4 die Vorhersagewerte aus Modell 3 gegen den Goldstandard getestet. Modell 5 erweitert Modell 4 insofern, dass die Vorhersagewerte aufbereitet werden.

5.4.1 *Ergebnisse der Modelle*

Da es sehr viele „O“ („Other“) Label im Datensatz gibt, das Interesse aber der Entitäten gilt, wird dieses Label bei der Berechnung der Evaluationsmaße ausgeschlossen. Danach erzielt das Baseline Conditional Random Field einen F_1-Score von 0.521. Der Precision-Wert ist mit 0.629 höher als der Recall-Wert mit 0.453. Das bedeutet, dass ca. 45% der tatsächlichen Entitäten gefunden wurden (Recall). Wenn eine Entität als solche gelabelt ist, dann ist diese mit 63.8%-iger Wahrscheinlichkeit eine tatsächliche Entität.

Die Vorhersagen sind für wiederholende Entitätsreferenzen in einem Dokument sehr inkonsistent. In manchen Fällen wird die erste Referenz, in anderen Fällen, die zweite und/oder dritte Referenz annotiert. Deshalb werden die Ergebnisse wie in Abschnitt 5.4 beschrieben aufbereitet. Dieser Ansatz (Modell 2) erhöht die Precision um 9.5%-Punkte (0.724), den Recall um 1.4%-Punkte und den F_1-Score um 4%-Punkte.

Wenn nicht nur das erste Vorkommen einer Entität im Dokument von den Editor*innen annotiert würde, dann ergeben sich die Ergebnisse aus Modell 3 wie folgt: Der F_1-Score verbessert sich im Vergleich zum Baselinemodell um ca. 6%-Punkte auf 0.578. Precision und Recall erhöhen sich um 4.9%-Punkte bzw. 6.7%-Punkte (Modell 3).

Da das Ziel eine Übereinstimmung mit der Goldstandard Annotation der Editor*innen ist, werden die Vorhersagen aus Modell 3 gegen die wahren Annotationen der Editor*innen getestet. Wie zu erwarten, sinkt der F_1-Wert ggü. Modell 3. Er ist dennoch (wenn auch nur minimal) größer als das Baselinemodell (Modell 1). Zusätzlich (Modell 5) zu Modell 4 werden die vorhergesagten Ergebnisse wieder aufbereitet. Im Vergleich zu Modell 4 steigt der F_1-Score um 7.7%-Punkte und liegt damit auch deutlich über dem Baselinemodell. Vor allem der Recall liegt ca. 10% Punkte über dem Baselinemodell. Die Precision verbessert sich auch um ca. 5% Punkte.

Wenn die Ergebnisse der Modelle sequentiell betrachtet werden, lassen sich daraus verschiedene Erkenntnisse ziehen. Durch die Modifikation von Modell 1 zu Modell 2 wird deutlich, dass Conditional Random Fields das Annotationsschema, dass immer nur die erste Entität annotiert werden soll, nicht erkennen. Modell 3 zeigt, dass Conditional Random Fields eine konsistente Datengrundlage benötigt. Das bedeutet, dass ein besseres Ergebnis erzielt werden kann, wenn die Editor*innen alle Entitäten annotieren würden. Es ist nicht überraschend, dass die Vorhersagen für ein Modell mit konsistenten Daten keine guten Vorhersagen für die von den Editor*innen annotierten Daten liefern. Dem kann allerdings entgegen gewirkt werden, indem die Vorhersagen im Nachgang so aufbereiten, dass sie dem Schema der Editor*innen entsprechen. Dieser Ansatz liefert sogar über alle anderen Modelle hinweg, die besten Ergebnisse. Das liegt einerseits daran, dass das Modell auf konsistenten Daten trainiert wurde, wodurch die Entitäten der einzelnen Entitätsreferenzen besser erlernt werden können. Zusätzlich ist die Wahrscheinlichkeit einer richtigen Zuordnung für Mehrfachentitätsreferenzen bei diesem Ansatz größer, da alle Label für diese Referenz betrachtet

werden und wenn mindestens eines die Entität erkannt hat, wird diese der ersten Referenz zugewiesen. Dadurch entsteht eine Richtig-Positive (TP) Zuordnung.

Ein weiteres Modell, das nicht in der Tabelle 5.4 dargestellt ist, bereitet die Ergebnisse des Baselinemodells (Modell 1) so auf, dass sie nicht dem IOB2-Format folgen, sondern dem IO-Format. Dadurch erhöht sich der F_1-Score um 2%-Punkte auf auf 54%. Das zeigt, dass das Modell Schwierigkeiten in der Unterscheidung beginnender und fortlaufender Token hat. Diese Erkenntnis ist für „PER"-Entitäten recht eingängig, da in manchen Dokumenten Vor- und Nachnamen vorkommen können, die mit Vorname (z. B. Frank) → „B-PER" und Nachname (z. B. Wedekind) → „I-PER" annotiert werden. In anderen Dokumenten kann dieselbe Person aber nur mit Nachnamen erwähnt werden, wodurch die Zuordnung Nachname (Wedekind) „I-PER" stattfände. Die Wahrscheinlichkeit, dass der Nachname (Wedekind) der Entität „B-PER" oder „I-PER" zugeordnet wird, ist allein auf Basis dieser Informationen nicht eindeutig. Bei der Betrachtung von Tabelle ist die absolute Anzahl der falschen Zuordnungen von tatsächlichen „B-XXX" Entitäten zu „I-XXX" allerdings doch recht gering.

5.4.2 *Ergebnisse der einzelnen Entitäten*

Um noch genauer nachzuvollziehen, wie sich die Ergebnisse zusammensetzen, werden Precision, Recall und der F_1-Score in Tabelle 5.5 in die einzelnen Entitäten aufgeschlüsselt. Die Ergebnisse beziehen sich auf das Baselinemodell (Modell 1) und auf Modell 3, das alle Entitäten, auch die Mehrfachreferenzen, in Betracht zieht. Im Baselinemodell liegen die Precision-Werte von 6 der 10 Entitäten über der Durchschnittsprecision von 0.629. Die Werte für „I-WRK", „I-OTK", „B-EVT" und „I-EVT" sind geringer als der Durchschnittswert. Auffallend ist die kleine Fallzahl dieser Entitäten (85, 44, 18 und 3). Bei den Recall-Werten liegen 4 über dem Durchschnittswert von 0.453. „B-PER", „I-PER", „B-ORT" und „B-OTK". Diese 4 Entitäten sind auch unter den 6 mit einem überdurchschnittlichen Precision-Wert. Somit stellen sie auch die Entitäten mit den Besten F_1-Scores dar. Wenn das Modell nur in Hinblick auf Person-, Ort und Örtlichkeit-Entitäten evaluiert wird, steigt der F_1-Score um 3%-Punkte auf 0.55. Zusammenfassend werden Personen (egal, ob Anfangstoken („B-PER") oder fortlaufenden Token („I-PER")), die Anfangstoken von Orten („B-ORT") und Örtlichkeiten („B-OTK") gut vom Modell erkannt. Die fortlaufenden Token von Orten und Örtlichkeiten haben nur eine geringe Fallzahl und werden vermutlich deshalb nicht gut vom Modell vorhergesagt. Dasselbe Argument trifft auf „B-" und „I-EVT" zu.

Aus dem Rahmen fallen die Entitäten für Werke. Diese liegen in genügender Fallzahl vor, erzielen jedoch geringe Gütemaße. Der Grund wird in der Datenaufbereitung liegen. Wie in Abschnitt 5.2 beschrieben, werden die Annotationen aus den HTML-Dateien in der Form gefiltert, dass die Entitätsreferenz und die korrespondierende Entität extrahiert werden. Danach werden die Annotationen tokenisiert und mit den tokenisierten Dokumen-

Entität	Precision	Recall	F_1-Score	Anzahl
B-PER	.628 (.708)	.483 (.578)	.546 (.636)	749 (961)
I-PER	.657 (.667)	.482 (.460)	.556 (.544)	191 (261)
B-ORT	.651 (.765)	.563 (.763)	.604 (.764)	252 (359)
I-ORT	1.000 (.667)	.143 (.250)	.250 (.364)	7 (8)
B-OTK	.639 (.639)	.450 (.429)	.541 (.513)	80 (91)
I-OTK	.480 (.500)	.273 (.370)	.348 (.425)	44 (46)
B-WRK	.639 (.703)	.324 (.314)	.430 (.434)	142 (204)
I-WRK	.562 (.560)	.212 (.156)	.308 (.245)	85 (179)
B-EVT	.400 (.333)	.111 (.095)	.174 (.148)	18 (21)
I-EVT	.000 (.000)	.000 (.000)	.000 (.000)	3 (3)
Gesamt	.629 (.687)	.453 (.516)	**.521** (**.578**)	1571 (2133)

Tabelle 5.5: Evaluation Conditional Random Field auf Token-Level

ten gemergt. Hierbei wurde sichergestellt, dass bei Entitätsreferenzen, die mehrmals im Dokument vorkommen, gemäß dem Standard der Annotationen immer nur die erste Referenz annotiert wird. Zum besseren Verständnis betrachten wir uns beispielhaft die PER-Entität „Frank Wedekind“. Der durchgeführte Workflow tokenisierte diesen Ausdruck und labelte ihn danach im IOB2-Format, sodass die Token „Frank“ mit „B-WRK“ und „Wedekind“ mit „I-PER“ gelabelt wurden. Danach wurden die Token, mit den Token aus den Dokumenten gemergt,sodass immer das erste Vorkommen von „Frank“ bzw. „Wedekind“ annotiert wird. Allerdings wurde übersehen, dass Entitätsreferenzen aus der HTML-Datei auch mit Token mergen können, die *keine* Entität darstellen und vor der eigentlichen Entität im Dokument stehen. Das führt bei Personen oder Orten weniger zu Problemen, da es selten vorkommt, dass eine Person- oder Ortsreferenz ein Ausdruck ist, der auch keine Entität darstellen kann. Bei Werken kann es aber zu vielen falschen Annotationen kommen. Wird z. B. das Werk „Die Frau vom Meere“ betrachtet, stellt jeder Token ein Problem dar, da jeder eine „Other“ Kategorie darstellen kann oder sogar eine „PER“ Entität (Frau). Wenn einer der Token vor dem eigentlichen Ausdruck des Werkes im Dokument steht, wird dieser mit der „B-“ bzw. „I-WRK“ Entität versehen und der eigentliche Token im Werk erhält die Kategorie „Other“. Das heißt hier entstehen sogar zwei Fehler. Der erste Token wird fälschlicherweise als Entität gelabelt (Falsch-Positiv) und bei dem zweiten Token wird die Entität fälschlicherweise übergangen (Falsch-Negativ). Würde das Modell die Entitäten richtig (d.h. wie von den Editor*innen) gelabelt vorhersagen, würde das Modell dieselben Fehler machen. Es würde nämlich eine Entität für „die Frau vom Meere“ erkennen, obwohl die Datenaufbereitung diese als „Other“ gelabelt hat, wodurch ein Falsch-Positiv entstünde. Für die fälschlicherweise als Entitäten gelabelten Token, würde es diese als „Other“ vorhersagen. Dadurch entstünde ein Falsch-Negativ. Das bedeutet, dass auch wenn das Modell die

tatsächlichen Annotationen der Editor*innen vorhersagen würde, wird im Modell ein geringer F_1-Score erzielt. Ein hoher F_1-Score wäre auch nur gut, wenn sicher wäre, dass die fälschliche Annotation bei der Datenaufbereitung nur sehr selten vorkämen.

Wenn alle Vorkommnisse annotiert werden (Modell 3) verbessern sich die F_1-Scores für „B-PER“ um 9%-Punkte, „B-ORT“ um 16%-Punkte, „I-ORT“ um 11%-Punkte, I-OTK um 8%-Punkte und „I-EVT“ um 14%-Punkte. Wenn nur diese Entitäten bei der Evaluation betrachtet werden, steigt der F_1-Score auf 0.633. Der F_1-Score der Entität „I-WRK“ reduziert sich um 6%-Punkte. Wie bereits erwähnt, entstehen durch die Datenaufbereitung bei dieser Entität Fehler. Der Fehler ist bei diesem Annotationsschema folgenträchtiger, da die Wahrscheinlichkeit von falschen Zuordnungen viel größer ist.

Sowohl für die Ergebnisse des Annotationsschemas 1 als auch 2 erzielen die einzelnen Entitäten durchgängig eine höhere Precision gegenüber dem Recall. Das bedeutet, dass vergleichsweise weniger Entitäten erkannt werden aber wenn sie erkannt werden, sind sie mit hoher Wahrscheinlichkeit richtig.

Zusätzlich zu verschiedenen Annotationsschemata wurde ein weiteres Feature hinzugefügt: Wortembeddings. Für diesen Zweck wird HistWord2Vec[7] verwendet. Da einige Token aus dem Dokumentencorpus trotz der zeitlichen Übereinstimmung mit den Worten aus HistWord2Vec orthographisch nicht übereinstimmten, wurden besagte Token mittels Fuzzymatching den korrespondierenden Vektoren zugeordnet. Aufgrund der langen Berechnungszeit, wurde nach 670 Token gestoppt. Es verbleiben allerdings noch 20% der Token ohne Embedding. Da Wordembeddings in ihrer ursprünglichen Form nicht das passende Datenformat für die Verwendung in Conditional Random Fields besitzen, werden die Embeddings mittels kMeans-Algorithmus in 40 Cluster zusammengefasst. Ähnliche Wörter sollen demnach in dasselbe Cluster fallen und erhalten dieselbe Clusternummer. Token, die kein Wortembedding besitzen, werden auf 99 gesetzt. Durch diesen Ansatz lässt sich der F_1-Score um 1.6%-Punkte erhöhen.

7 https://nlp.stanford.edu/projects/histwords/

6 ZUSAMMENFASSUNG UND AUSBLICK

6.1 ZUSAMMENFASSUNG

Digitale (historisch-kritische) Editionen stellen durch ihre Charakteristik Entwickler*innen vor Herausforderungen. Sie vereinen auf den ersten Blick unvereinbare Konzepte: Geschichte und Moderne, Vergangenheit und Zukunft. Durch diese Eigenschaft begegnen sich im Aufbau von digitalen (historisch-kritischen) Editionen Menschen aus unterschiedlichen Disziplinen, mit unterschiedlichen Zielen und Ansprüchen. Das kann zu Konflikten führen, birgt gleichzeitig aber großes Potential. Das Ziel ist, das breit aufgestellte Know-How zusammenzubringen und effektiv in digitalen (historisch-kritischen) Editionen einzusetzen. Um das zu erreichen, ist ein enger Austausch der diversen Disziplinen miteinander notwendig.

Aus Sicht von Computerwissenschaflter*innen stellen digitale (historisch-kritische) Editionen Information Retrieval Systeme dar, denn sie generieren aus Daten Informationen. Die Gewährleistung des Zugangs zu den Informationen ist in einem erheblichen Ausmaß den Editor*innen zuzuschreiben. Erst durch ihre kritische Auseinandersetzung und strukturelle, verständliche Aufbereitung, unter anderem durch Kommentare und Annotationen in den Dokumente, werden diese für Nutzer*innen zu einer Informationsquelle. Aufgrund der Bedeutung der Editor*innen für die Güte digitaler (historisch-kritischer) Editionen bzw. für die Güte des Information Retrieval Systems besteht die Notwendigkeit Editor*innen in ihrer Arbeit zu unterstützen.

Um die konkreten Anforderungen der Editor*innen zu untersuchen und dadurch sicherzustellen, dass nichts implementiert wird, das nicht benötigt oder gewünscht ist, wurde eine Umfrage generiert. Die Umfrageergebnisse zeigen, dass die Nachfrage nach einem Named-Entity-Recognition System hoch ist. Die Annotation von Dokumenten ist eine Standardaufgabe in digitalen (historisch-kritischen) Editionen. Die meisten befragten Editor*innen annotieren Personen und Orte. Folglich ist es wichtig, dass das Named-Entity-Recognition System diese beiden Entitäten verlässlich erkennt.

Das in der Arbeit generierte Modell erzielt eine Precision von 0.629 und einen Recall von 0.453. Daraus resultiert ein F_1-Score von 0.521. Die besten Einzelwerte erzielen die Entitäten Person, Ort und Örtlichkeit. Zu unterscheiden sind allerdings die Anfangstoken und die fortführenden Token dieser Entitäten. Bei Person-Entitäten sind die F_1-Scores der beiden Tokenarten gleich hoch bei ca. 0.555. Das Modell erkennt auch die Anfangstoken von Orts- und Örtlichkeits-Entitäten mit einem F_1-Score von 0.604 bzw. 0.541. In der Identifikation fortlaufender Token für Orts- und Örtlichkeits-Entitäten performt das Modell allerdings mit F_1-Scores von 0.250 bzw. 0.348. Die Fallzahlen von 7 „I-ORT“ und 44 „I-OTK“ im Testdatensatz sind allerdings sehr

gering, um valide Aussagen treffen zu können. Die Ergebnisse zeigen allerdings auch, dass das Modell teilweise Probleme hat, zwischen „B-“ und „I-“Enitäten zu differenzieren, den Token aber in die richtige Überkategorie einordnet. Wenn die Daten im Nachhinein noch zusätzlich so aufbereitet werden, verbessert sich das Modell sichtlich.

Annotationen können nach unterschiedlichen Schemata durchgeführt werden. Das Annotationsschema, welches dem Modell und der Frank Wedekind Brief Edition zugrunde liegt, verfolgt den Ansatz, immer nur die erste Entitätsreferenz von Mehrfachnennungen im Text zu annotieren. Es ist durchaus in anderen Editionen möglich, dass alle Referenzen annotiert werden. Conditional Random Fields performen besser auf diesem Schema und erzielen im konkreten Anwendungsfall einen F_1-Score von 0.578, welcher sich durch eine Precision von 0.687 und einem Recall von 0.516 ergibt. Die besten Ergebnisse unter den Entitäten, werden auch in diesem Modell bei den Entitäten Person, Ort und Örtlichkeit erzielt. Das gleiche Muster für die fortlaufenden Token wie im Modell zum ersten Schema, lässt sich auch in diesem Modell finden, sodass „I-ORT“ und „I-OTK“ schlechter performen als die korrespondierenden Anfangstoken. Der Unterschied zwischen den „B-“ und „I-OTK“ reduziert sich allerdings von 20%-Punkten auf 9%-Punkte.

Die für die Editor*innen wichtigsten Entitäten (Personen, Orte und Örtlichkeiten) werden von den Modellen am ehesten erkannt. Wenn nur diese Entitäten betrachtet werden erhöht sich der F_1-Score um 3-6%-Punkte.

Das erfolgreichste Modell erzielt einen F_1-Score von 0.605 und wurde mit einem Modell trainiert, das Daten nutzt, die einem konsistenten Annotationsschema folgt, das alle Entitätsreferenzen annotiert. Zusätzlich werden die Vorhersagen im Nachgang, insofern aufbereitet, dass sie dem Annotationsschema der Editor*innen der Frank Wedekind Edition folgen. Wenn nur Person-, Ort- und Örtlichkeit-Entitäten in Betracht gezogen werden erzielt das Modell sogar einen F_1-Score von 0.644.

6.2 AUSBLICK

Die vorliegende Arbeit trägt dazu bei, ein größeres Verständnis für den Einfluss des Annotationsschemas auf die Performanz von Sequenzlabeling Verfahren zu schaffen. Es wurde gezeigt, dass Conditional Random Fields sehr sensibel gegenüber Inkonsistenzen in der Annotation gleicher Referenzen mit verschiedenen Entitäten bzw. Kategorien ist. Einen besseren Ansatz für Dokumente, in denen immer nur eine Referenz von Mehrfachreferenzen annotiert wird, können Neuronale Netze sein. Insbesondere BiLSTM könnten gute Ergebnisse erzielen, da sie durch ihren Aufbau bewusst Informationen aus der Dokumentensequenz abspeichern oder löschen können. Dadurch ist es wahrscheinlich, dass sie das Muster, immer nur die erste Referenz zu annotieren, erkennen und genauso vorhersagen können.

Die Modelle wurden auf zwei wichtigen Maßen evaluiert: Precision und Recall. Bei einer hohen Precision und einem geringen Recall, haben Editor*innen die Sicherheit, dass die Vorschläge des Modells mit großer Wahr-

scheinlichkeit richtig sind, allerdings wird nur ein geringer Teil der Entitäten im Dokument identifiziert. Bei einem hohen Recall und einer geringen Precision, werden viele Entitäten erkannt, allerdings werden auch viele falsche Vorschläge vom System angeboten. Für weitere Verbesserungen ist zu klären, welches Maß für Editor*innen wichtiger ist, sodass gezielter Modifikationen des Modells vorgenommen werden können.

Named-Entity-Recognition dient als Ausgangspunkt für weitere Natural Language Processing Verfahren. Es ist beispielsweise ein Vorverarbeitungsschritt für Relation Extraction. Durch Relation Extraction werden Verbindungen und Beziehungen der einzelnen Entitäten im Dokumentencorpus identifiziert und beschrieben. Diese können anschaulich in Form von Graphen visualisiert werden. Ein weiterer Bereich, der zunehmendes Interesse in digitalen (historisch-kritischen) Editionen erzielt, ist der der Linked (Open) Data.[13] Named Entities werden mit zusätzlichem Hintergrundwissen angereichert, indem sie auf interne Informationen aus anderen Dokumenten verweisen oder auf externe Ressourcen referenzieren. Eine konkrete Möglichkeit ist die Verlinkung der Entitäten von Frank Wedekinds Werken zu ihren Volltexten.[1]

1 16 seiner Werke sind durch Projekt Gutenberg-DE frei zugänglich: https://www.projekt-gutenberg.org/autoren/namen/wedekind.html

Teil II

APPENDIX

A

FRAGEBOGEN

Effizientes Suchen und Editieren von Dokumenten für Nutzer*innen und Editor*innen digitaler (historisch-kritischer) Editionen durch Natural Language Processing

Herzlich willkommen!

Ein wichtiges Merkmal von digitalen (historisch-kritischen) Editionen ist die Suchfunktion, die den Nutzer*innen einen Einblick in den Inhalt ermöglicht und gezielte Informationen schnell zugänglich macht. Sie dient nicht nur Nutzer*innen, sondern auch Editor*innen. Um die Suchfunktion effektiver zu gestalten, erhalten Sie nachfolgend Fragen zur Suche im Allgemeinen, zur Unterstützung des Editionsprozesses (wenn Sie Editor*in sind), zur Präsentation der Suchergebnisse und zu Ihren Erfahrungen/Problemen bei der Suche in digitalen (historisch-kritischen) Editionen.
Vielen Dank, dass Sie an der Umfrage teilnehmen. Die Umfrage wird im Rahmen der Masterarbeit „Optimierung der Suchfunktion und des Editionsprozesses digitaler (historisch-kritischer) Editionen anhand der Frank Wedekind Briefedition" von Tamara Haeder durchgeführt und ist Teil des Projekts "Edition der Korrespondenz Frank Wedekinds als Online-Volltextdatenbank" der Hochschule Darmstadt in Kooperation mit der Frank Wedekind-Gesellschaft e.V.. Die Ergebnisse dieser Umfrage werden im Rahmen einer wissenschaftlichen Arbeit veröffentlicht.

Die Umfrage dauert ca. 15-20 Minuten. Um fortzufahren, klicken Sie auf „Weiter".

Dies ist eine anonyme Umfrage.
In den Umfrageantworten werden keine persönlichen Informationen über Sie gespeichert, es sei denn, in einer Frage wird explizit danach gefragt. Wenn Sie für diese Umfrage einen Zugangscode benutzt haben, so können Sie sicher sein, dass der Zugangsschlüssel nicht zusammen mit den Daten abgespeichert wurde. Er wird in einer getrennten Tabelle aufbewahrt und nur aktualisiert, um zu speichern, ob Sie diese Umfrage abgeschlossen haben oder nicht. Es gibt keinen Weg, die Zugangscodes mit den Umfrageergebnissen zusammenzuführen.

Allgemeines

Nachfolgend werden Ihnen allgemeine Fragen zu Ihrer Nutzung der Suchfunktion in digitalen (historisch-kritischen) Editionen gestellt.

---

1. Waren Sie schon einmal als Editor einer digitalen (historisch-kritischen) Edition tätig?
   (Nur eine Antwortmöglichkeit)
   ☐ Ja
   ☐ Nein

   Filter: Wenn Frage 1 = „Ja“:

   1.1 Wenn Sie eine digitale (historisch-kritische) Edition editieren, welche Merkmale werden von Ihnen annotiert?
   (Mehrere Antworten sind möglich)
   ☐ Personen
   ☐ Orte
   ☐ Ereignisse
   ☐ Spezifische Themen: Politik, Literatur, Kunst, etc.
   ☐ Fremdwörter
   ☐ Wörter in einer Fremdsprache
   ☐ Abkürzungen
   ☐ Formatierungen
   ☐ Sonstiges, und zwar: _____

   Filter: Wenn Frage 1 = „Nein“:

   1.1 Wenn Sie eine digitalen (historisch-kritische) Edition nutzen, für welche Informationen interessieren Sie sich?
   (Mehrere Antworten sind möglich)
   ☐ Einzelne Personen
   ☐ Beziehungen zwischen Personen
   ☐ Orte
   ☐ Ereignisse
   ☐ Sprache
   ☐ Spezifische Themen: Literatur, Kunst, Politik, etc.
   ☐ Sonstiges, und zwar: _____

2. Es gibt verschiedene Möglichkeiten *wie* Nutzer*innen Suchanfragen stellen können. Geben Sie an auf welche Art und Weise Sie suchen.
(Mehrere Antworten sind möglich)

Begriffserklärungen

Ein Wort/ einzelne Wörter: Suche nach einem einzigen Wort oder nach mehreren einzelnen (unabhängigen) Wörtern.

Phrasensuche: Bezeichnung für die Suche nach einer Zeichenfolge, die aus mehreren Wörtern besteht. Hierbei werden die Wörter nicht einzeln, sondern als zusammengesetzter Ausdruck gesucht.

Boolesche Suche: Verwendung von Booleschen Operatoren ("und"/ "oder"/ "nicht") zwischen den Suchbegriffen, um die Suchanfrage genauer zu spezifizieren.

Natürlichsprachliche Suche: Formulierung der Suchanfrage in einer normalen, gesprochenen Sprache.

Wildcard Suche: Verwendung von Wildcards (Platzhaltern) bei der Formulierung der Suchanfrage. Platzhalter sind Symbole (z. B. ?,*), die Zeichen in Suchbegriffen ersetzen, um jede mögliche Kombination des Suchbegriffs zu erhalten.

☐ Ein Wort/ einzelne Wörter
☐ Phrasensuche
☐ Boolesche Suche
☐ Natürlichsprachliche Suche
☐ Wildcard Suche
☐ Sonstiges, und zwar: _____

Filter: Wenn Frage 1.1 min. 1 Antwort hat:

1.1.1 Sie haben angegeben, dass Sie sich für unten genannte Informationen interessieren. Vergeben Sie Zahlen von 1 (wenig) bis 5 (sehr), wie sehr Sie sich dafür interessieren.

	1	2	3	4	5
Wenn Frage 1.1 = „Personen“: Personen	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Orte“ Orte	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Ereignisse“ Ereignisse	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Sprache“ Sprache	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Spezifische Themen: Literatur, Kunst, Politik, etc.“: Spezifische Themen: Literatur, Kunst, Politik, etc.	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Fremdwörter“ Fremdwörter	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Wörter in einer Fremd-sprache“ Wörter in einer Fremdsprache	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Abkürzungen“ Abkürzungen	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Formatierungen“ Formatierungen	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „{Sonstiges}“: {Sonstiges}	☐	☐	☐	☐	☐

Filter: Wenn Frage 1.1 min. 1 Antwort hat:

1.1.1 Sie haben angegeben, dass Sie sich für unten genannte Informationen interessieren. Vergeben Sie Zahlen von 1 (wenig) bis 5 (sehr), wie sehr Sie sich dafür interessieren.

	1	2	3	4	5
Wenn Frage 1.1 = „Einzelne Personen“: Einzelne Personen	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Beziehungen zwischen Personen“: Beziehungen zwischen Personen	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Orte“ Orte	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Ereignisse“ Ereignisse	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Sprache“ Sprache	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „Spezifische Themen: Literatur, Kunst, Politik, etc.“: Spezifische Themen: Literatur, Kunst, Politik, etc.	☐	☐	☐	☐	☐
Wenn Frage 1.1 = „{Sonstiges}“: {Sonstiges}	☐	☐	☐	☐	☐

Filter: Wenn Frage 1.1.1 min 1 Antwort >= 4 hat

1.1.1.1 Sie haben eine hohe Priorität (4 oder 5 Punkte) für die unten genannte Merkmale angegeben. Stellen Sie sich vor, dass Sie zusätzliche Informationen aus anderen Dokumenten benötigen, um dieses Merkmal zu annotieren. Tätigen Sie jeweils im freien Feld eine Suchanfrage, um an die benötigen Informationen zu gelangen. Überlegen Sie sich hierzu eine fiktive Suchanfrage oder stellen Sie eine Suchanfrage, die Sie so schon einmal formuliert haben.

	Suchfeld
Wenn „Personen“ in Frage 1.1.1 >= 4: Personen	
Wenn „Orte“ in Frage 1.1.1 >= 4: Orte	
Wenn „Ereignisse“ in Frage 1.1.1 >= 4: Ereignisse	
Wenn „Sprache“ in Frage 1.1.1 >= 4: Sprache	
Wenn „Orte“ in Frage 1.1.1 >= 4: Spezifische Themen: Literatur, Kunst, Politik, etc.	
Wenn „Fremdwörter“ in Frage 1.1.1 >= 4: Fremdwörter	
Wenn „Wörter in einer Fremdsprache“ in Frage 1.1.1 >= 4: Wörter in einer Fremdsprache	
Wenn „Abkürzungen“ in Frage 1.1.1 >= 4: Abkürzungen	
Wenn „Formatierungen“ in Frage 1.1.1 >= 4: Formatierungen	
Wenn „{Sonstiges}“ in Frage 1.1.1 >= 4: {Sonstiges}	

Filter: Wenn Frage1.1.1 min 1 Antwort >= 4 hat

1.1.1.1 Sie haben angegeben, dass Sie sich für unten genannte Informationen sehr (4 oder 5 Punkte) interessieren. Tätigen Sie jeweils im freien Feld eine Suchanfrage zu diesem Thema. Überlegen Sie sich hierzu eine fiktive Suchanfrage oder stellen Sie eine Suchanfrage, die Sie so schon einmal formuliert haben.

	Suchfeld
Wenn „einzelne Personen“ in Frage 1.1.1 >= 4: Einzelne Personen	
Wenn „Beziehungen zwischen Personen " in Frage 1.1.1 >= 4: Beziehungen zwischen Personen	
Wenn „Orte“ in Frage 1.1.1 >= 4: Orte	
Wenn „Ereignisse“ in Frage 1.1.1 >= 4: Ereignisse	
Wenn „Sprache“ in Frage 1.1.1 >= 4: Sprache	
Wenn „Orte“ in Frage 1.1.1 >= 4: Spezifische Themen: Literatur, Kunst, Politik, etc.	
Wenn „{Sonstiges}“ in Frage 1.1.1 >= 4: {Sonstiges}	

3. Um Ihre Suchanfrage zu vereinfachen und zu verbessern, gibt es verschiedene Möglichkeiten. Welche Unterstützung sollte beim Stellen der Suchanfrage vorhanden sein?
(Mehrere Antworten sind möglich)

Begriffserklärungen

***Autovervollständigung (auch Autocomplete)*:** Eine Funktion, in der die Anwendung den Rest der Suchanfrage erahnt und Ihnen Vorschläge zur Vervollständigung gibt während Sie die Eingabe tätigen.

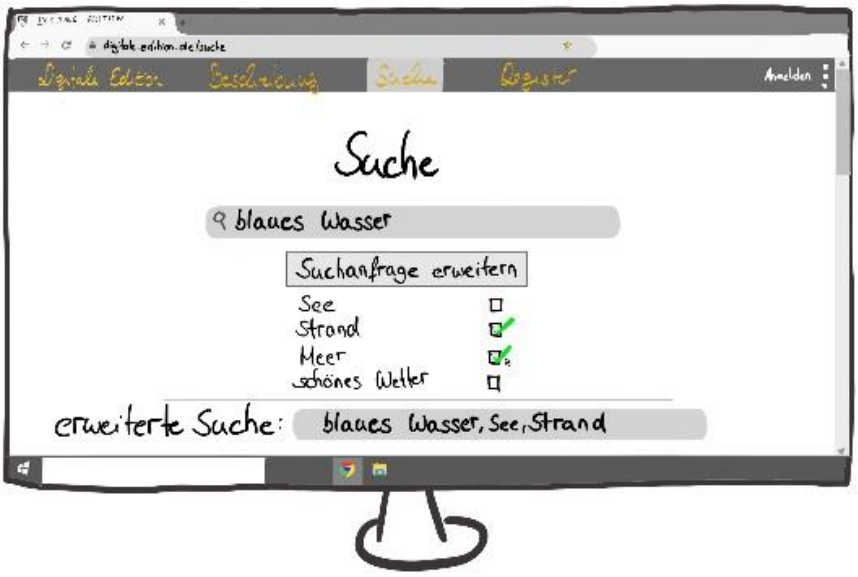
Query Erweiterung: Ein Prozess, bei dem zusätzliche Begriffe (z.B. Synonyme, semantisch verwandte Begriffe) ausgewählt werden und Ihrer Anfrage hinzugefügt werden. Dies dient dazu weitere relevante Dokumente für Sie zu finden.

Query Reformulierung: Ein Prozess, bei dem ihre ursprüngliche Anfrage umformuliert wird, sodass Ihre Worte zu den Worten in den Dokumenten passt und Sie somit mit größerer Wahrscheinlichkeit Ergebnisse zurückbekommen.

☐ Autovervollständigung

☐ Rechtschreibüberprüfung

☐ Query Erweiterung



☐ Query Reformulierung

Filter: Wenn Frage 3 = „Query Erweiterung“ ODER „Query Reformulierung“:

3.1 Sie Suchanfragen Erweiterung/Reformulierung kann entweder automatisiert oder manuell stattfinden.

- *Automatisiert*: Suchanfragen Erweiterung/ Reformulierung findet im Hintergrund statt. Dabei versucht die Applikation durch einen Algorithmus, Ihre Suchanfrage so zu erweitern/umzuformulieren, dass Sie die bestmöglichen Ergebnisse erzielen.
- *Manuell*: Ihnen werden Suchbegriffe/ andere Formulierungen vorgeschlagen, bei denen Sie selbst entscheiden, ob Sie diese Ihrer Suchanfrage hinzufügen bzw. Ihre Suchanfrage umformulieren.

Welche Möglichkeit bevorzugen Sie?
(Nur eine Antwort ist möglich)

☐ Automatisiert
☐ Manuell
☐ Keine, ich finde beide gut.

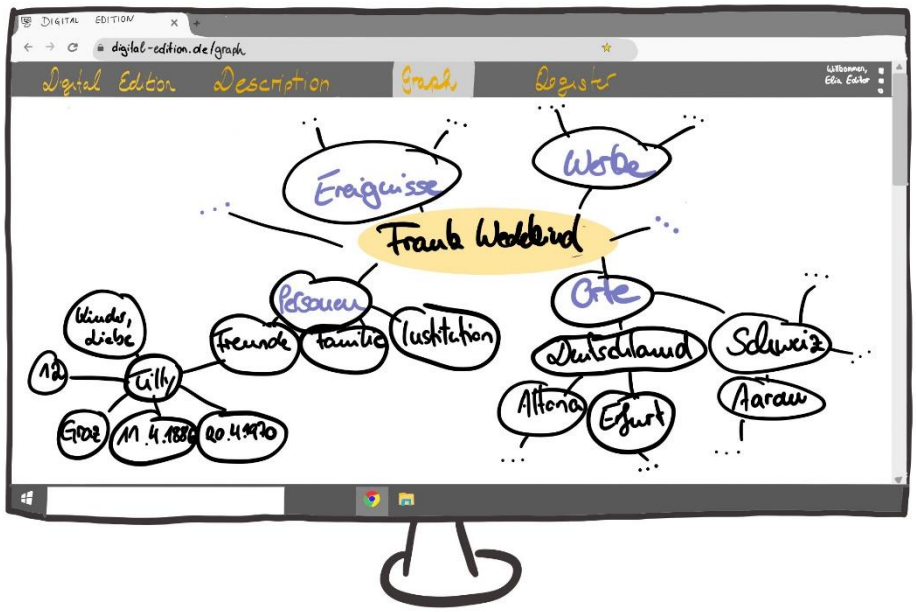
Filter: Wenn Frage 1 = „Ja“

Unterstützung des Editionsprozesses

Nachfolgend werden Ihnen Fragen gestellt, inwieweit Sie durch computergestützte Textverarbeitung bei Ihrem Editionsprozess unterstützt werden können. Geben Sie jeweils an, ob Sie die Methoden als nützlich oder überflüssig erachten. Beispielhaft werden Ihnen Mockups präsentiert, welche sich auf eine Briefedition beziehen.

---

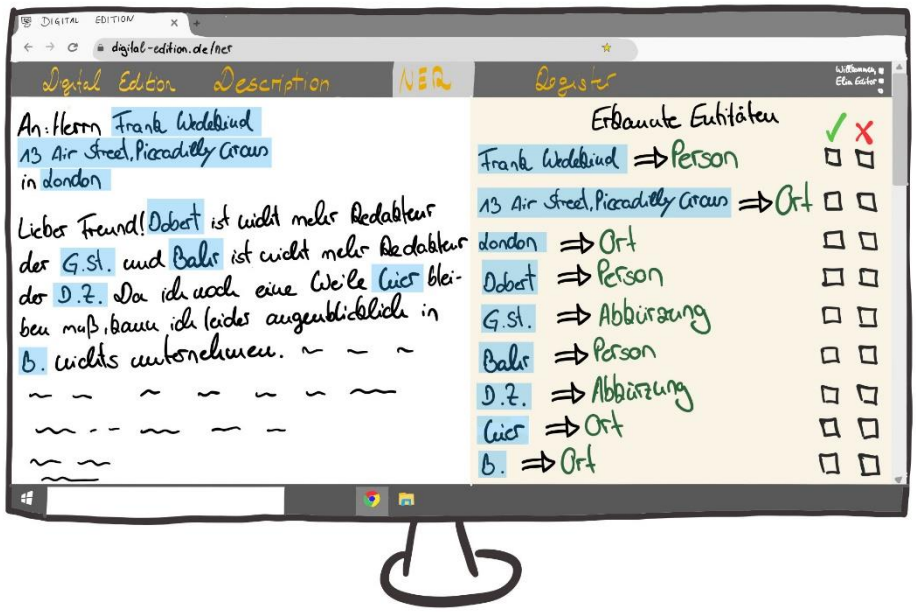
4. Um einen übersichtlichen Einblick in die digitale (historisch-kritische) Edition zu erhalten, gibt es die Möglichkeit Informationen optisch als Graphen mit sogenannten Knoten und Kanten darzustellen. Dabei können einerseits Informationen, die Sie durch Ihre Annotationen bereitgestellt haben, als auch textuelle Informationen abgebildet werden.



☐ Nützlich

☐ Überflüssig

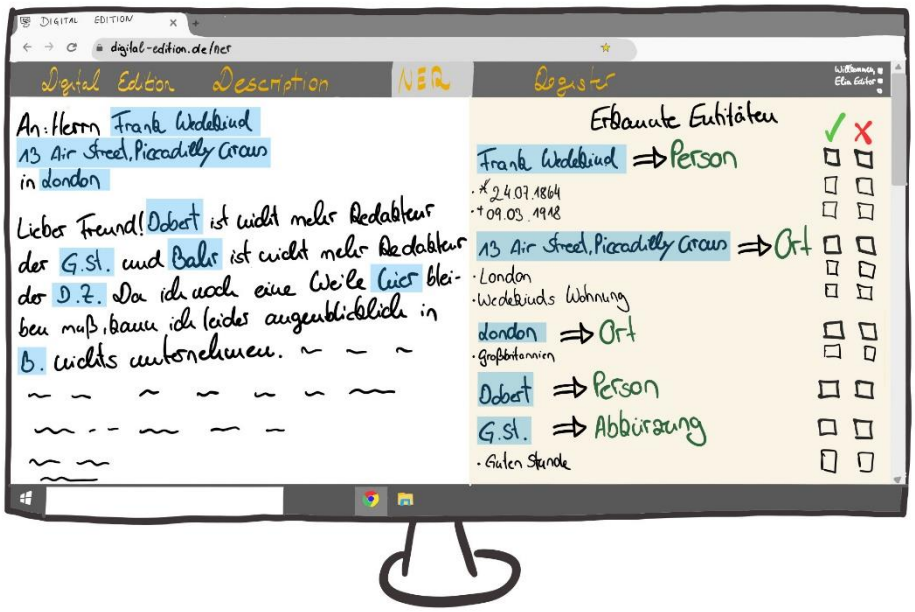
5. Der Annotationsprozess kann mittels Named-Entity Recognition (Eigennamenerkennung) unterstützt werden. Dabei werden automatisch Entitäten (Personen, Orte, Daten, etc.) erkannt. Diese könnten Ihnen vorgeschlagen werden. Sie entscheiden dann, ob der Vorschlag angenommen oder verworfen werden soll.



☐ Nützlich
☐ Überflüssig

6. Zusätzlich zur o.g. automatischen Named-Entity Recognition könnten bei bereits hinterlegen Entitäten, diese erkannt und mit allen zur Verfügung stehenden Informationen an angereichert werden.

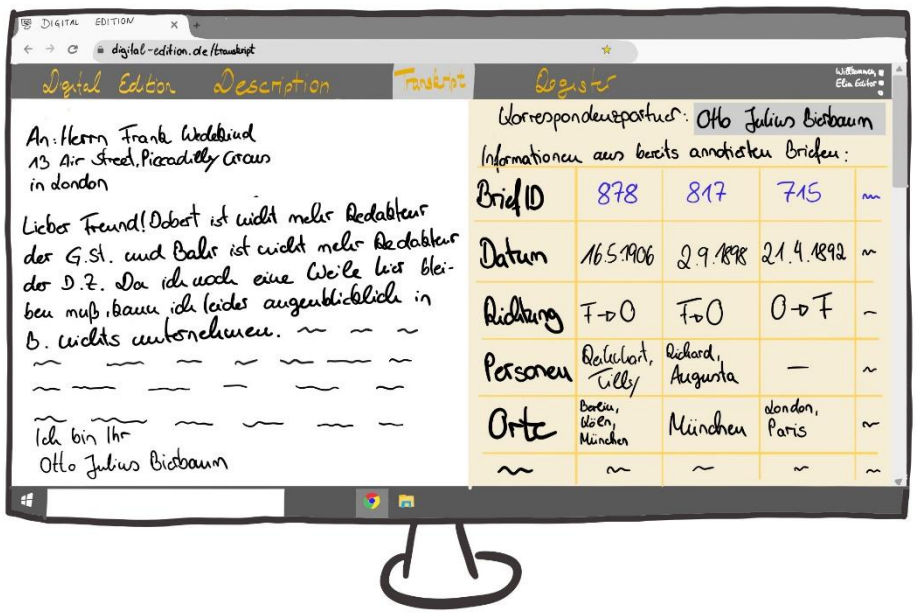
.



☐ Nützlich

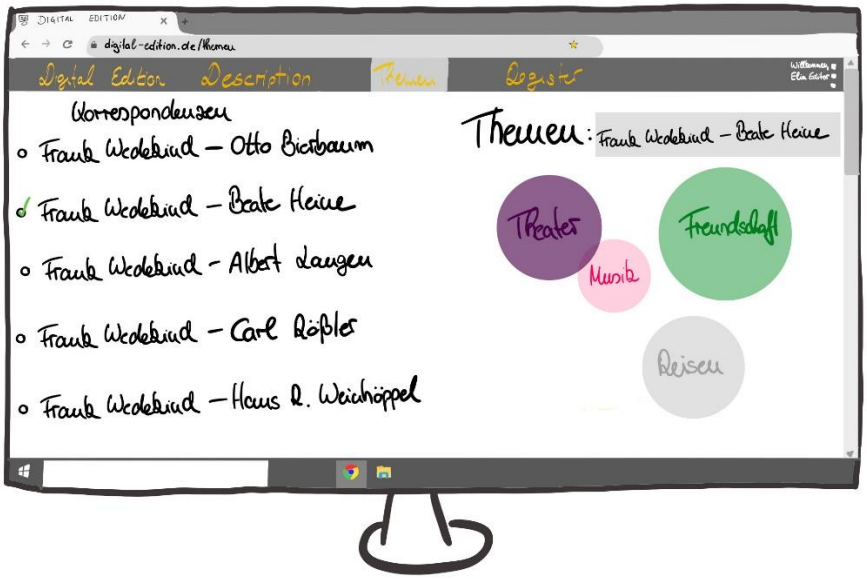
☐ Überflüssig

7. Bei der Eingabe eines neuen Korrespondenzstückes, besteht die Möglichkeit einer Auflistung der Annotationen aller Briefe mit diesem/r Korrespondenzpartner/in. Bei Bedarf können Sie durch Anklicken der BriefID zu diesem Brief gelangen.



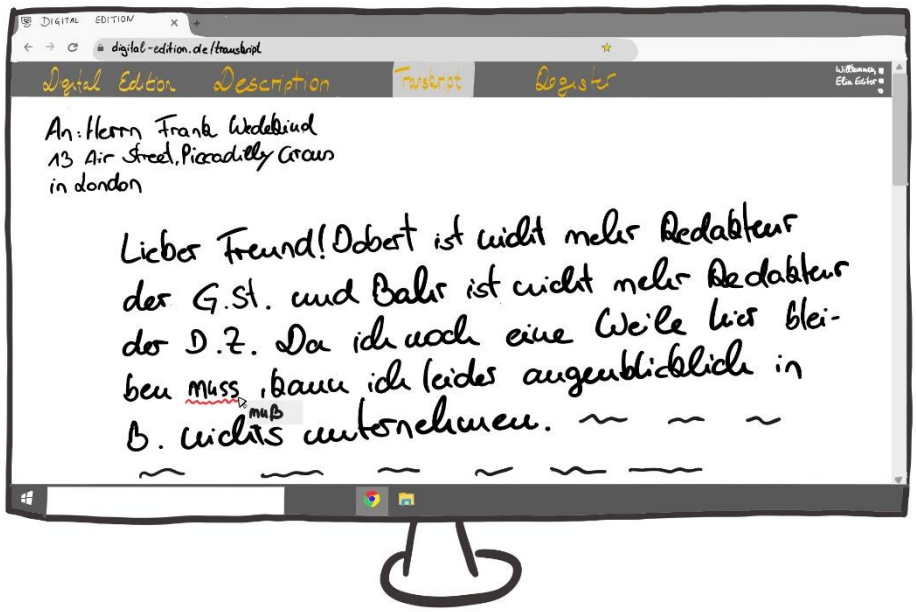
☐ Nützlich
☐ Überflüssig

8. Topic Modellierung bezeichnet ein Verfahren, bei dem anhand eines gegebenen Textes automatisch übergeordnete Themen aus dem Text extrahiert werden. Dies kann bei der Annotation z.B. genutzt werden, indem ein Topic Model zum aktuellen Dokument erstellt wird und Sie durch Anklicken der Themen, zu Dokumenten gelangen, welche dieses Thema auch beinhalten. Neben Topic Modellen auf einzelnen Dokumenten ist es auch möglich Topic Modelle für spezifische Korrespondenzpartner/innen (zweites Bild), Männer vs. Frauen, Privatpersonen vs. Institutionen etc. aufzustellen.



☐ Nützlich
☐ Überflüssig

9. Es ist möglich eine Rechtschreibprüfung basierend auf der Rechtschreibung, die dem Dokumentenkorpus zu Grunde liegt zu implementieren. So können Sie darauf hingewiesen werden, dass ein Wort, dass Sie in der neuen deutschen Rechtschreibung geschrieben haben, im Dokumentenkorpus anders geschrieben hinterlegt ist.



☐ Nützlich

☐ Überflüssig

10. Die Annotation unbekannter Wörtern/Phrasen oder die Suche nach zusätzlichen Informationen zu Wörtern/ Phrasen kann durch verschiedene Methoden unterstützt werden. Geben Sie an, welche Methoden Sie als hilfreich erachten.

Begriffserklärung

***Ontologien*:** Klassifikationssystem, das eine strukturierte Abbildungen des sprachlichen und sachlichen Wissens (graphisch) darstellt.

Kookkurrenzen: Gemeinsames Auftreten zweier Terme in einer übergeordneten Einheit (z.B. einem Satz, Paragraph, etc.). Es besteht die Annahme, dass die Terme voneinander abhängen, wenn sie häufig gemeinsam auftreten.

☐ Thesaurus

☐ Ontologien

☐ Volltextsuche

☐ Kookkurrenzen

☐ Informationen aus Internetquellen

☐ Sonstiges, und zwar: _____

Präsentation der Ergebnisdokumente

Nachdem Sie Ihre Suchanfrage gestellt haben, werden Ihnen relevante Dokumente aufgelistet. Nachfolgend werden Ihnen Fragen zur Präsentation der Ergebnisse gestellt.

11. Wie viele relevante Ergebnisdokumente sollen angezeigt werden?
(Nur eine Antwort ist möglich)

☐ Die ersten 10 relevanten Dokumente.

☐ Die ersten 20 relevanten Dokumente.

☐ Die ersten 30 relevanten Dokumente.

☐ Die ersten 40 relevanten Dokumente.

☐ Die ersten 50 relevanten Dokumente.

☐ Die ersten 51-100 relevanten Dokumente.

☐ Die ersten 101-200 relevanten Dokumente.

☐ Alle relevanten Dokumente

12. Welche Eigenschaften sind Ihnen bei der Auflistung der Ergebnisdokumente wichtig?
(Mehrere Antworten sind möglich)

☐ Sichtbarkeit, nach welchen Kriterien die Reihenfolge der Ergebnisauflistung stattfindet.

☐ Möglichkeit, dass die Reihenfolge der Ergebnisliste nach bestimmten Kriterien erfolgt, die Sie selbst auswählen können (z.B. Datum, Häufigkeit des gesuchten Wortes).

☐ Weitere Filtermöglichkeiten/ Die Möglichkeit, die Ergebnisse präziser einzuschränken.

☐ Sonstiges, und zwar: _____

Filter: Wenn Frage 11 = „Möglichkeit, dass die Reihenfolge der Ergebnisliste nach bestimmten Kriterien erfolgt, die Sie selbst auswählen können (z.B. Datum, Häufigkeit des gesuchten Wortes).“:

12.1 Sie wünschen sich, dass die Reihenfolge der Ergebnisliste nach bestimmten Kriterien (wie z.B. Datum, Häufigkeit des gesuchten Wortes) erfolgt, die Sie selbst auswählen können. Nach welchen Kriterien würden Sie gerne sortieren können?
(Mehrere Antworten sind möglich)

☐ Datum

☐ Autor*in/ Absender*in

☐ Empfänger*in

☐ Schreibort

☐ Häufigkeit des gesuchten Wortes/ der gesuchten Phrase im Ergebnisdokument

☐ Abstand der gesuchten Wörter zueinander im Ergebnisdokument (z.B. direkt nachfolgend, im selben Satz/ Absatz/ etc.)

☐ Position der Wörter/Phrasen (Anfang/ Ende des Dokuments, in der Grußformel, etc.)

☐ Sonstiges, und zwar: _____

Filter: Wenn Frage 11 = „Weitere Filtermöglichkeiten/ Die Möglichkeit, die Ergebnisse präziser einzuschränken.“:

12.2 Sie wünschen sich, dass es weitere Filtermöglichkeiten gibt bzw. die Möglichkeit die Ergebnisse präziser einzuschränken. Welche Filterkriterien sind Ihnen wichtig?
(Mehrere Antworten sind möglich)

☐ Datum

☐ Autor*in/ Absender*in

☐ Empfänger*in

☐ Schreibort

☐ Häufigkeit des gesuchten Wortes/ der gesuchten Phrase im Ergebnisdokument

☐ Abstand der gesuchten Wörter zueinander im Ergebnisdokument (z.B. direkt nachfolgend, im selben Satz/ Absatz/ etc.)

☐ Position der Wörter/Phrasen (Anfang/ Ende des Dokuments, in der Grußformel, etc.)

☐ Annotationen

☐ Sonstiges, und zwar: _____

Filter: Wenn Frage 11.2 = „Annotationen“:

12.2.1 Sie haben angegeben, dass eine Filterung nach Annotationen hilfreich wäre. An welche Annotationen denken Sie hierbei konkret?

Antwort: _____

Erfahrungen mit digitalen (historisch-kritischen) Editionen

Zum Schluss würden wir Ihnen gerne noch ein paar Fragen zu Ihren bisherigen Erfahrungen mit digitalen (historisch-kritischen) Editionen stellen.

13. Wie viele Jahre Erfahrung haben Sie mit digitalen (historisch-kritischen) Editionen?
(Nur eine Antwort ist möglich)

☐ Keine
☐ Weniger als 1 Jahr
☐ 1-5 Jahre
☐ 6-10 Jahre
☐11-15 Jahre
☐16-20 Jahre
☐ 21-25 Jahre
☐ Mehr als 25 Jahre

Filter: wenn Frage 12 ≠ „keine“

13.1 Wie sehr sind Sie mit den inhaltlichen Themen, mit denen sich die digitale (historisch-kritische) Edition befasst, vertraut?
(Nur eine Antwort ist möglich)

☐ sehr
☐ mittel
☐ wenig/ gar nicht

13.2 Wählen Sie alle Dokumententypen, mit denen Sie sich in digital (historisch-kritischen) Editionen beschäftigen, aus.
(Mehrere Antworten sind möglich)

☐ Manuskripte
☐ Briefe
☐ Drucke
☐ Bücher
☐ Kodex
☐ Tafel
☐ Tagebücher
☐ Sonstiges, und zwar: _____

13.3 Und was hat Ihnen bei der Suchmöglichkeit und der Ergebnispräsentation daran am besten gefallen:

Suchmöglichkeit: _____

Ergebnispräsentation: _____

13.4 Und wo sehen Sie Verbesserungspotential:

Suchmöglichkeit: _____

Ergebnispräsentation: _____

13.5 Welche Probleme sind bisher aufgetreten?
(Mehrere Antworten sind möglich)

☐ Eingabe der Suchanfrage war nicht möglich.

☐ Probleme mit Sonderzeichen.

☐ Es wurden keine Dokumente gefunden.

☐ Es wurden Dokumente gefunden aber keine relevanten Dokumente.

☐ Sonstiges, und zwar: _____

Demographische Daten

Vielen Dank, dass Sie an dieser Umfrage teilgenommen haben. Zur besseren Differenzierbarkeit der Anforderungen verschiedener Nutzergruppen, würden wir gerne noch zum Abschluss etwas über Ihre Person erfahren.

14. Wie alt sind Sie?
    - ☐ 0-30 Jahre
    - ☐ 31-40 Jahre
    - ☐ 41-50 Jahre
    - ☐ 51-60 Jahre
    - ☐ 61+ Jahre
    - ☐ Keine Angabe

15. Wie ist ihr biologisches Geschlecht?
    (Nur eine Antwort ist möglich)

    - ☐ Weiblich
    - ☐ Männlich
    - ☐ Keine Angabe
    - ☐ Anderes, und zwar: _____

16. In welchem Bereich sind Sie tätig?

    - ☐ Antwort: _____

17. Was ist Ihr höchster akademischer Grad/Titel?
    (Nur eine Antwort ist möglich)

    - ☐ kein Abschluss
    - ☐ Bachelor
    - ☐ Master
    - ☐ Doktor
    - ☐ Professor
    - ☐ Keine Angabe

18. Welche berufliche Tätigkeit üben Sie derzeit aus?

    Antwort: _____

Vielen Dank für Ihre Teilnahme!

A.1 UMFRAGEERGEBNISSE DER NUTZER*INNEN

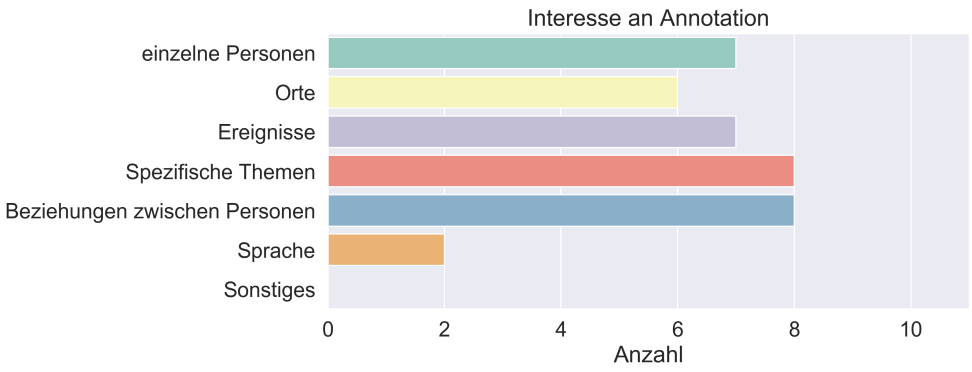


Abbildung A.1: Interesse in Annotationen von Nutzer*innen

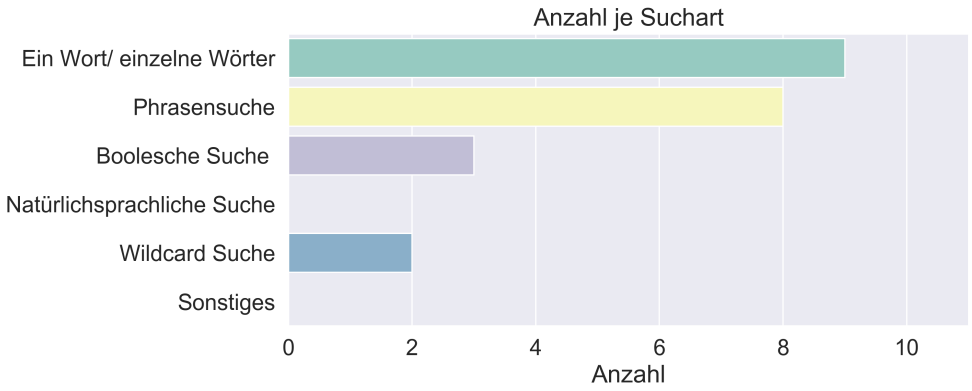


Abbildung A.2: Formulierungsstrategien der Suchanfrage von Nutzer*innen

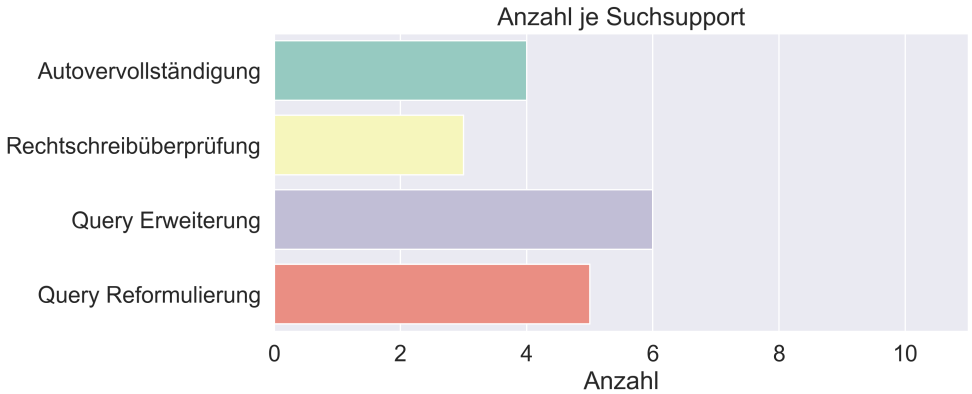


Abbildung A.3: Unterstützungswünsche bei der Suchanfrage von Nutzer*innen

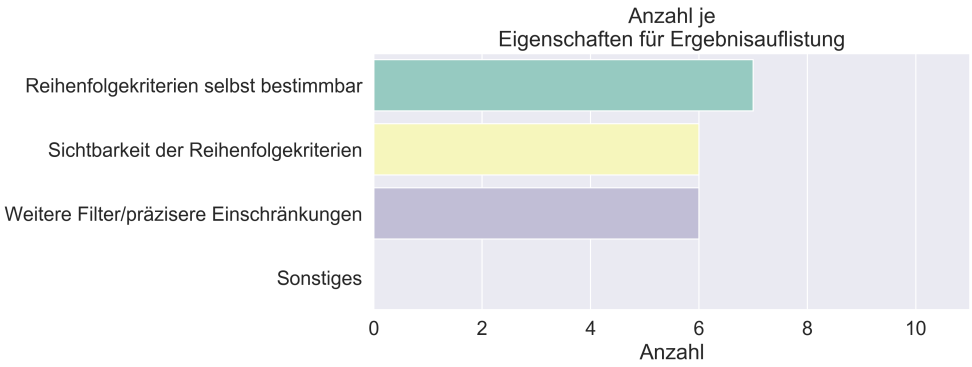


Abbildung A.4: Eigenschaften der Ergebnisauflistung

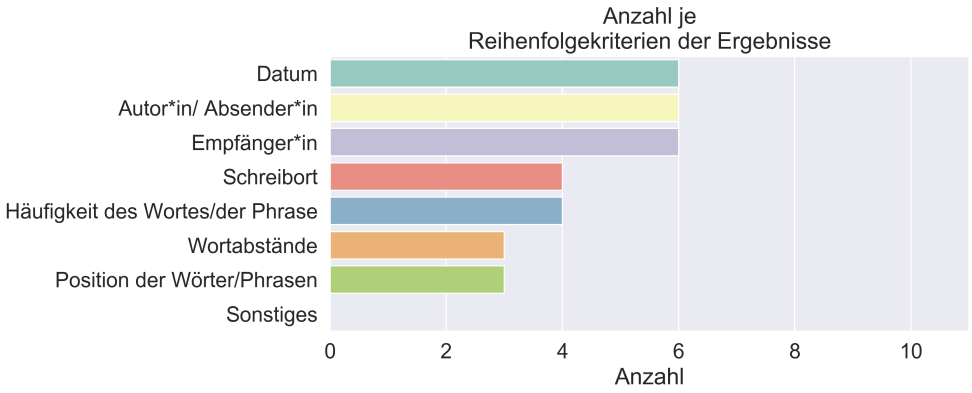


Abbildung A.5: Attribute zur Bestimmung der Reihenfolge

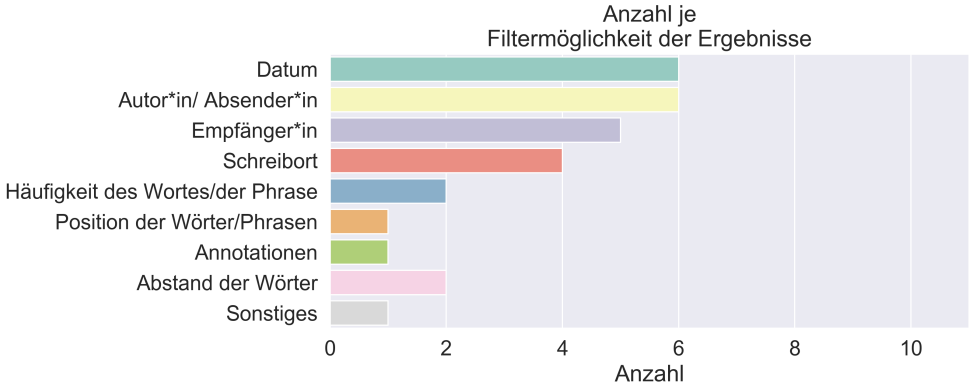


Abbildung A.6: Attribute zur Filterung

B

DATENAUFBEREITUNG UND ANALYSEERGEBNISSE

SpaCy			TreeTagger		
PoS-Tag	Anzahl	Definition	PoS-Tag	Definition	Anzahl
NOUN:	Nomen	40221	NN:	Nomen	41015
PUNCT:	Interpunktion	36578	PPER:	Personal-pronomen	36578
PRON:	Demonstrativ-pronomen, substantivisch	30435	ADV:	Adverb	19407
VERB:	Verb	26059	APPR:	Präposition	17865
DET:	Demonstrativ-pronomen, adjektivisch	24009	$.:	Satzbe-endende Interpunktion	16893
ADV:	Präposition	22654	$,:	Komma	16506
ADP:	Präposition mit Artikel	22424	ART:	(un)bestimm-ter Artikel	14827
ADJ:	attributives Adjektiv	19518	ADJA:	attributives Adjektiv	13746
PROPN:	Eigenname	15968	VVFIN	Vollverb	12624
AUX:	Hilfs-Wort	10051	NE	Eigenname	11209

Tabelle B.1: Vergleich Top 10 PoS-Tags: spaCy vs. TreeTagger

Entität	Falsche Zuordnung
B-PER	I-PER:10, B-WRK:1, O:197
I-PER	B-PER:10 , B-WRK:1, I-WRK:1 , O:58
B-ORT	O:62
I-ORT	O:10
I-OTK	O:6
B-WRK	B-PER:3, O=34
I-WRK	O:30
B-EVT	I-WRK:1, O:7
I-EVT	O:1
O	B-PER:316, I-PER:60, B-ORT:114, I-ORT:7, B-OTK:48, I-OTK:25, B-WRK:79, I-WRK:31, B-EVT:14, I-EVT:3

Tabelle B.2: Falsche Zuordnungen je Entität

LITERATUR

[1] Fabian Barteld, Chris Biemann und Heike Zinsmeister. "Token-based spelling variant detection in Middle Low German texts". In: *Language Resources and Evaluation* 53.4 (2019), S. 677–706.

[2] Matthew J Beal, Zoubin Ghahramani und Carl E Rasmussen. "The infinite hidden Markov model". In: *Advances in neural information processing systems*. 2002, S. 577–584.

[3] Yevgeni Berzak, Michal Richter, Carsten Ehrler und Todd Shore. "Information Retrieval and Visualization for the Historical Domain". In: *Language Technology for Cultural Heritage*. Springer, 2011, S. 197–212.

[4] Peter Boot, Anna Cappellotto, Wout Dillen, Franz Fischer, Aodhán Kelly, Andreas Mertgens, Anna-Maria Sichani, Elena Spadini und Dirk van Hulle. "Advances in Digital Scholarly Editing". In: *DiXiT Conferences in The Hague, Cologne, and Antwerp*. 2017.

[5] Nancy Chinchor. "Appendix E: MUC-7 Named Entity Task Definition (version 3.5)". In: *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*. 1998. URL: https://www.aclweb.org/anthology/M98-1028.

[6] Thierry Declerck, Antonia Scheidel und Piroska Lendvai. "Proppian Content Descriptors in an Augmented Annotation Schema for Fairy Tales". In: *ECAI 2010* (2010), S. 35.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[8] DiXiT. *DiXiT ABOUT*. URL: https://dixit.uni-koeln.de/about/ (besucht am 06. 11. 2020).

[9] Nora Echelmeyer, Nils Reiter und Sarah Schulz. "Ein PoS-Tagger für"das"Mittelhochdeutsche". In: (2017).

[10] Maud Ehrmann, Matteo Romanello, Alex Flückiger und Simon Clematide. "Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers". In: *CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum*. Bd. 2696. CONF. CEUR. 2020.

[11] Gerd Fischer, Matthias Lehner und Angela Puchert. *Einführung in die Stochastik*. Springer, 2015.

[12] Greta Franzini, Melissa Terras und Simon Mahony. "A catalogue of digital editions". In: *Digital scholarly editing: Theories and practices* (2016), S. 161–182.

[13] F Frontini, C Brando, J Ganascia, F Frontini und C Brando. "Domain-adapted named-entity linker using Linked Data To cite this version". In: (2015).

[14] Robert Gaizauskas und Yorick Wilks. "Information extraction: Beyond document retrieval". In: *Journal of documentation* 54.1 (1998), S. 70–105.

[15] Cyril Goutte und Eric Gaussier. "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation". In: *European conference on information retrieval*. Springer. 2005, S. 345–359.

[16] Claire Grover, Sharon Givon, Richard Tobin und Julian Ball. "Named Entity Recognition for Digitised Historical Texts." In: *LREC*. 2008.

[17] Iris Hendrickx, Michel Généreux und Rita Marquilhas. "Automatic pragmatic text segmentation of historical letters". In: *Language Technology for Cultural Heritage*. Springer, 2011, S. 135–153.

[18] Andreas Henrich. *Information Retrieval 1 (Grundlagen, Modelle und Anwendungen)*. Bamberg, 2008. URL: http://www.uni-bamberg.de/minf/IR1-Buch.

[19] Erhard Hinrichs, Marie Hinrichs, Sandra Kübler und Thorsten Trippel. *Language technology for digital humanities: introduction to the special issue*. 2019.

[20] Jing Jiang. "Information extraction from text". In: *Mining text data*. Springer, 2012, S. 11–41.

[21] Epaminondas Kapetanios, Doina Tatar und Christian Sacarea. *Natural language processing: semantic aspects*. CRC Press, 2013.

[22] Frédéric Kaplan und Isabella di Lenardo. "Big data of the past". In: *Frontiers in Digital Humanities* 4 (2017), S. 12.

[23] Roman Klinger und Christoph M Friedrich. "User's choice of precision and recall in named entity recognition". In: *Proceedings of the International Conference RANLP-2009*. 2009, S. 192–196.

[24] Tim William Machan. *Medieval literature: texts and interpretation*. Bd. 79. Mrts, 1991.

[25] Andrew McCallum und Wei Li. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons". In: (2003).

[26] David Nadeau und Satoshi Sekine. "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1 (2007), S. 3–26.

[27] Nils Reiter, Oliver Hellwig, Anette Frank, Irina Gossmann, Borayin Maitreya Larios, Julio Rodrigues und Britta Zeller. "Adapting NLP tools and frame-semantic resources for the semantic analysis of ritual descriptions". In: *Language Technology for Cultural Heritage*. Springer, 2011, S. 171–193.

[28] Patrick Sahle. "What is a scholarly digital edition". In: *Digital scholarly editing: Theories and practices* (2016), S. 19–39.

[29] Helmut Schmid. "Improvements in part-of-speech tagging with an application to German". In: *Natural language processing using very large corpora*. Springer, 1999, S. 13–25.

[30] John B Smith. “Computer criticism”. In: *Style* (1978), S. 326–356.

[31] Elena Spadini. “Producing Scholarly Digital Editions”. In: *Reassembling the Republic of Letters in the Digital Age. Standards, Systems, Scholarship* (2019), S. 251–260.

[32] Caroline Sporleder, Antal van den Bosch und Kalliopi Zervanou. *Language technology for cultural heritage: Selected papers from the LaTeCH workshop series*. Springer Science & Business Media, 2011.

[33] Charles Sutton und Andrew McCallum. “An introduction to conditional random fields for relational learning”. In: *Introduction to statistical relational learning* 2 (2006), S. 93–128.

[34] René Witte, Ralf Krestel, Thomas Kappler und Peter C Lockemann. “Converting a historical architecture encyclopedia into a semantic knowledge base”. In: *IEEE Intelligent Systems* 25.1 (2010), S. 58–67.

[35] Miguel Won, Patricia Murrieta-Flores und Bruno Martins. “ensemble named entity recognition (ner): evaluating ner Tools in the identification of Place names in historical corpora”. In: *Frontiers in Digital Humanities* 5 (2018), S. 2.

[36] Bengong Yu und Zhaodi Fan. “A comprehensive review of conditional random fields: variants, hybrids and applications”. In: *Artificial Intelligence Review* (2019), S. 1–45.