Karlsruher Institut für Technologie
Institut für Technikzukünfte (ITZ)
Teilinstitut Wissenschaftskommunikation

31. Januar 2021

Prüfer:
Dr. Sarah Kohler
Prof. Dr. Markus Lehmkuhl

Masterarbeit

Validierung eines *NER*-Verfahrens
zur automatisierten Identifikation von Akteuren
in journalistischen Texten

-

Validation of a *NER* method
for the automated identification of actors
in journalistic texts

Cecilia Buz

Abstract

Im Vergleich zu manuellen Untersuchungsmethoden ermöglicht der Einsatz von automatisierten Verfahren in der Kommunikationswissenschaft weitaus schnellere Analysen von umfangreichen Textmengen. Eines dieser Verfahren namens *Named Entity Recognition* (*NER*) ist auf die automatisierte Identifikation von Eigennamen in Texten spezialisiert und soll eingehend untersucht und angewandt werden.
Ziel der Arbeit ist die Prüfung der Eignung solch eines Verfahrens für künftige, umfangreiche Akteursanalysen. Diese erlauben umfassende, medienübergreifende Vergleiche in der Berichterstattung, ebenso wie die quantitative Analyse des Vorkommens und der Vielfalt der Akteure über lange Zeiträume.
Da die frei verfügbaren *NER*-Verfahren für ihren Einsatz mit spezifischen Textdaten trainiert und optimiert werden, ist ungewiss, ob ihre Nutzung bei der Analyse von unbekannten journalistischen Nachrichtentexten richtige und präzise Ergebnisse liefert. Dies soll in der vorliegenden Masterarbeit durch eine konkrete Anwendung evaluiert werden. Hierfür werden drei verschiedene *NER*-Verfahren gegenübergestellt und ein Vergleich der Ergebnisse der automatisierten Analyse mit den Ergebnissen aus einer manuellen Inhaltsanalyse desselben Datensatzes vollzogen.
Die Befunde des Vergleichs zeigen eine hohe Übereinstimmung zwischen den händisch erhobenen und den automatisiert identifizierten Akteuren. Doch es wird deutlich, dass die *NER*-Verfahren in der Vorbereitung und Durchführung durch viele Faktoren beeinflussbar sind, wodurch die Ergebnisse sehr variabel sind und das Verfahren im Ganzen schwierig zu validieren ist.

Compared to manual examination methods, the use of automated approaches in communication science enables much faster analyses of extensive text quantities. One of these procedures called *'Named Entity Recognition' (NER)* specializes in the automated identification of named entities in texts and will be examined and applied in detail.
The aim is to test the suitability of such a procedure for future, extensive actor analyses. These allow comprehensive, cross-media comparisons of the general news coverage, as well as the quantitative analysis of the occurrence, frequency and diversity of the named actors or institutions over long periods of time.
Since these *NER* methods are developed and trained using specific annotated text data, it is uncertain whether they will achieve precise and correct identification of entities with unknown journalistic articles. To evaluate that, this work applies three different *NER* methods and compares the outcome of these automated analyses with the results of a manual content analysis. The results show that there is a high concordance between the manually and automatically identified actors. However, it becomes clear that the preparation and implementation of the *NER* methods can be influenced by many factors, which means that the results are very variable and the method as a whole is difficult to validate.

Inhaltsverzeichnis

Hinweis: Aus Gründen der Lesbarkeit wurde im Text die männliche Form gewählt, nichtsdestoweniger beziehen sich die Angaben gleichermaßen auf Angehörige aller Geschlechter.

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

	Englisch	*Deutsch*
AIA	-	automatisierte Inhaltsanalyse
BoW	Bag of Words	[Repräsentation eines Textes als Sammlung von unzusammenhängenden Wörtern]
DGPuK	-	Deutsche Gesellschaft für Publizistik- und Kommunikationswissenschaft
DL	Deep Learning	[Teilgebiet des ML, welches neuronale Netze einsetzt]
HTML	Hypertext Markup Language	Hypertext-Auszeichnungssprache
LDA	Latent Dirichlet Allocation	[Wahrscheinlichkeitsmodell]
ML	Machine Learning	maschinelles Lernen
NE	Named Entity	Eigenname
NER	Named Entity Recognition	Eigennamen-Erkennung
NLP	Natural Language Processing	maschinelle Verarbeitung natürlicher Sprache
NLTK	Natural Language Toolkit	[Python-Toolkit für die Arbeit mit natürlicher Sprache]
NN ➢ CNN ➢ RNN	Neural Networks Convolutional Neural Networks Recurrent Neural Network	Neuronale Netze Faltende Neuronale Netze Rückgekoppelte Neuronale Netze
POS	Part of speech	Wortarten
RKI	-	Robert Koch-Institut
WHO	World Health Organization	Weltgesundheitsorganisation

1. Einleitung

In zahlreichen Bereichen unseres Alltags werden Algorithmen angewandt, die darauf spezialisiert sind, menschliche Sprache zu verarbeiten. Suchmaschinen, Chatbots und Sprachassistenten sind nahezu täglich im Einsatz und verwerten Text- oder Audiodaten maschinell (vgl. Schneider 2014: 40). Einzelhandelsunternehmen werten mit automatisierten Textanalysen ihre Kundenanfragen oder -rezensionen aus, aber auch in der Wissenschaft werden vermehrt algorithmische Lösungen angewandt, um Erkenntnisse aus Bild- und Textdateien zu erlangen (vgl. Niekler 2016: 2). Rechtswissenschaftler können damit beispielsweise allumfassend vergangene Rechtsprechungen auswerten und künftige Beschlüsse auf diesen Daten stützen (vgl. Evans et al. 2007: 1018). In den Wirtschaftswissenschaften werden hingegen digitale Daten aus den Finanzmärkten für die Untersuchung verhaltensökonomischer Prozesse genutzt (vgl. Boumans/Trilling 2016: 8).

Innerhalb der Sozialwissenschaften hat sich in den vergangenen Jahren das interdisziplinäre Arbeitsfeld der *Computational Communication Science* als Schnittstelle zwischen der angewandten Informatik und der Kommunikationswissenschaft gebildet (vgl. Domahidi et al. 2019: 3877). Dort steht die Nutzung computergestützter Methoden im Mittelpunkt, um die Inhalte großer Textsammlungen mittels Algorithmen zu analysieren, darin neue Zusammenhänge und Muster zu identifizieren und diese Datenstrukturen zu visualisieren (vgl. Grimmer/Stewart 2013: 267). Der Fokus dieser Masterarbeit liegt auf der Anwendung solcher digitalen Methoden, um Medien- und Kommunikationsdaten zu untersuchen. Dies ermöglicht die systematische Auswertung von Nachrichtenbeiträgen, Reden oder nutzergenerierten Online-Inhalten und somit die Erlangung von Erkenntnissen über die gegenwärtigen Medieninhalte sowie Meinungsbildungsprozesse in der Gesellschaft (vgl. Strippel et al. 2018: 8).

Aufgrund der kontinuierlichen Entstehung solcher digitalen Inhalte, deren Masse kaum zu überblicken ist, steigt die Relevanz ihrer automatisierten Analyse (vgl. Sommer et al. 2014: 14). Insbesondere für quantitative Untersuchungen, die aufgrund ihres Umfangs kaum händisch zu bewältigen sind, eignet sich der Einsatz computergestützter Maßnahmen sehr, setzt jedoch selektive Informatikkenntnisse bei den Kommunikationswissenschaftlern voraus (vgl. Jannidis et al. 2017: 95).

Die Masterarbeit beleuchtet daher Inhalte aus der Computerlinguistik, Informatik und Kommunikationswissenschaft, um eine automatisierte quantitative Analyse durchzuführen. Schwerpunkt ist hierbei die sogenannte *Named Entity Recognition (NER),* um Personen- und Organisationsnamen aus journalistischen Texten automatisiert zu extrahieren.

Dafür wird zunächst in dem ersten Teil der Arbeit eine Bestandsaufnahme der vorhandenen computergestützten Verfahren, die in der Kommunikationswissenschaft genutzt werden, durchgeführt. Es wird untersucht welche Verfahrensarten existieren und welches ihre jeweiligen Vor- und Nachteile sind. Im Mittelpunkt steht hierbei die automatisierte Inhaltsanalyse, bei der es explizit um die Erfassung und Untersuchung von Inhalten aus Textdaten geht.
Im Anschluss werden die Grundlagen der maschinellen Verarbeitung von natürlicher Sprache erläutert, um die Funktionsweise hinter dem Prozess der *NER* nachvollziehen zu können. Nachdem veranschaulicht wird, welche Verfahren dafür zu Verfügung stehen und wie sie implementiert werden, erfolgt die letztendliche Auswahl des zu nutzenden Codes für den empirischen Teil der Arbeit. In diesem Zuge wird ein Datensatz bestehend aus deutschsprachigen Nachrichtenartikeln der *dpa*, des *SPIEGELs* und der *WELT* für die maschinelle Verarbeitung aufbereitet. Die journalistischen Texte wurden für verschiedene manuelle Inhaltsanalysen des Lehrstuhls genutzt und die dort identifizierten Akteure dienen als Vergleichsgrundlage für die automatisiert erkannten Eigennamen. Es wird untersucht inwieweit die identifizierten Akteure und die Häufigkeit ihres Vorkommens übereinstimmen. Ebenso ist von Interesse, wie hoch der Anteil an irrelevanten Begriffen ist, bei denen es sich nicht um Eigennamen handelt.
Ziel dieser Masterarbeit ist nicht die klassische Beantwortung einer empirischen Fragestellung aus dem Bereich der Wissenschaftskommunikation, sondern viel mehr die Anwendung und Validierung einer digitalen Forschungsmethode für ihren zukünftigen Einsatz bei umfangreichen Akteursanalysen in der quantitativen Forschungsarbeit. Dafür werden die getätigten Arbeitsschritte bei der Anwendung des *NER*-Verfahrens sowie die auftretenden Herausforderungen strukturiert dargestellt.

Bisher gehören digitale Methoden in der Kommunikationswissenschaft nicht zu dem Ausbildungsstandard und in der entsprechenden Grundlagenliteratur existieren noch keine etablierten Gebrauchsrichtlinien (vgl. Strippel et al. 2018: 8). Zwar liegen zahlreiche Ausarbeitungen vor, die untersuchen, wie automatisierte Verfahren die inhaltliche Auswertung von Texten unterstützen können, dennoch herrscht keine Einigkeit darüber, welchen Qualitätsanforderungen diese genügen müssen (vgl. Niekler 2016: 179).
Mit zunehmender Verbreitung und Bedeutung solcher Methoden steigt allerdings die Notwendigkeit der methodologischen Diskussion über ihren Einsatz im kommunikationswissenschaftlichen Forschungsbereich (vgl. Strippel et al. 2018: 9). An dieser Stelle kann die Masterarbeit einen Beitrag leisten und aufzeigen, wo noch Aufklärungs- und Standardisierungsbedarf besteht und welche Hindernisse für ihren alltäglichen Einsatz überwunden werden müssen.

2. Einsatz automatisierter Verfahren in der Kommunikationswissenschaft

Computergestützte Verfahren umfassen ein breites Spektrum an Prozessen in der Kommunikationswissenschaft. Sie wurden zur Unterstützung von Arbeitsschritten entwickelt, sei es, um diese zu vereinfachen oder zu beschleunigen (vgl. Wettstein 2016: 124). Dazu zählen beispielsweise Vorgänge, wie die Filterung und Selektion von Beiträgen aus umfangreichen Textdatenbanken. Dafür werden häufig sogenannte *Web-Crawling* und *Web-Scraping* Technologien eingesetzt, mit denen zahlreiche webbasierte Datenquellen automatisch nach bestimmten Suchbegriffen durchsucht, die relevanten Beiträge identifiziert und schließlich extrahiert werden (vgl. Günther/Scharkow 2014: 112). In diesem Bereich ist unter anderem die automatisierte Datenakquise über eine *Rich Site Summary (RSS)* gängig. Dabei handelt es sich um die Bereitstellung von Daten über eine *RSS*-Liste, die eine stets aktualisierte Erfassung und Speicherung von textbasierten Nachrichtenangeboten zulässt (vgl. Trilling 2014: 73). Andere computergestützte Prozesse umfassen beispielsweise die automatische Verschlagwortung von Texten oder die Bereinigung und Aufbereitung der erhobenen Daten (vgl. Sommer et al. 2014: 13).
Komplizierter als die Verfahren zur Suche, Extraktion und Archivierung von digitalen Inhalten ist die analytische Arbeit mit ihnen (vgl. Niekler 2018: 15). Dies wird als *Text Mining* bezeichnet und beschreibt zunächst jegliche algorithmenbasierte Analyseprozesse, bei denen unstrukturierte Textdaten von Software erschlossen werden (vgl. Schneider/Zimmermann 2010: 36). Fokus der Masterarbeit liegt auf ebendieser Art von Informationsextraktion, bei der ein Computeralgorithmus zur Erfassung der Inhalte verwendet wird und dadurch eine automatisierte Form der klassischen Inhaltsanalyse ermöglicht (vgl. Scharkow 2012: 45).

2.1 Die automatisierte Inhaltsanalyse (AIA)

Die Inhaltsanalyse zählt zu den zentralen empirischen Erhebungsmethode der Kommunikationswissenschaft, um einheitlich und intersubjektiv nachvollziehbar Medieninhalte zu erfassen (vgl. Rössler/Geise 2013: 269). Diese Methode der Datenerhebung wird hauptsächlich genutzt, um die Themen der Medienberichterstattung sowie das Vorkommen von Akteuren oder Aussagen zu untersuchen, wobei in quantitativen Analysen die Inhalte durch ihre Zählung und Einteilung in Kategorien systematisch erhoben werden (vgl. Wettstein 2016: 5). Solche Analysen sind elementar, damit aus Textdaten letztlich Wissen entstehen kann (vgl. Graff/Theobald 2010: 195). Sie lassen erkennen, welche Inhalte in den Medien vermittelt werden und zeigen damit gesellschaftliche Prozesse auf, da die Annahme besteht, dass die Öffentlichkeit sich maßgeblich mit den veröffentlichten Inhalten auseinandersetzt oder anders betrachtet, die Medien gesellschaftlich relevante Inhalte thematisieren (vgl. Niekler 2016: 2).

Durch die Analyse bestimmter Variablen der Berichterstattung können Rückschlüsse über die Kommunikatoren als auch über die Rezipienten gezogen werden. So erlaubt zum Beispiel die Untersuchung der Medieninhalte in Kombination mit dem Online-Informationsverhalten der Rezipienten Erkenntnisse über mögliche Arten der Einflussnahme der Medien im Meinungsbildungsprozess (vgl. Rössler 2017: 255).

In der Vergangenheit wurden Inhaltsanalysen überwiegend manuell durchgeführt und meist nur computergestützt aufbereitet (vgl. Scharkow 2012: 46). Doch bereits vor 20 Jahren wurde bei einer Methodentagung der Deutschen Gesellschaft für Publizistik und Kommunikationswissenschaft (DGPuK) zusammengetragen und diskutiert, wie Teile oder gar der gesamte Prozess dieser zentralen Datenerhebungsmethode automatisiert werden können (vgl. Sommer et al. 2014: 9). Trotz der zahlreichen bestehenden Herausforderungen wurde großes Potenzial in der Automatisierung gesehen und eine enorme Weiterentwicklung in den kommenden Jahren vorausgesagt (vgl. ebd.: 10). Die Forschung dazu blieb in den darauffolgenden Jahren allerdings in einem überschaubaren Rahmen (vgl. ebd.). In einer Untersuchung von Früh und Früh wurden Studien aus führenden sozial- und kommunikationswissenschaftlichen Fachzeitschriften aus den Jahren 2000 bis 2009 analysiert und festgestellt, dass die Inhaltsanalyse zwar dominierend als Methode eingesetzt, jedoch nur in seltenen Fällen ein automatisiertes Verfahren dabei angewandt wurde (vgl. Früh/Früh 2015: 38).

Erst im Jahr 2012 gelang dem Kommunikationswissenschaftler Michael Scharkow mit der Entwicklung eines computergestützten Programms ein entscheidender Fortschritt (vgl. Sommer et al. 2014: 10). Er setzte dafür trainierte Algorithmen (s. Kap. 2.2.2) in den Prozess der Inhaltsanalyse ein und zeigte auf, inwieweit sich solch ein Verfahren für den kommunikationswissenschaftlichen Forschungsalltag eignet (vgl. Scharkow 2012: 16).

Es folgten verschiedene anwendungsbezogene Ausarbeitungen und zusätzliche Tagungen in der Kommunikationswissenschaftsgemeinschaft, um unter anderem die Schwierigkeiten bei der Analyse der neuen Medienformate zu diskutieren (vgl. Sommer et al. 2014: 10). Es wurde mehrheitlich die Erweiterung der Theorien und Methoden angestrebt, um auf den Wandel der Medienlandschaft und die *Datafizierung* reagieren zu können (vgl. Hepp 2016: 229). Letzteres steht für die Repräsentation des sozialen Lebens in computerisierten Daten und die Tatsache, dass immer mehr Teile unserer Kommunikation in Datenform vorliegen und wertvolle Erkenntnisse aus diesen digitalen Informationen erhalten werden können (vgl. ebd.).

Durch die computergestützte Analyse von Online-Nachrichtenartikeln und nutzergenerierten Beiträgen können neuartige Untersuchungseinheiten analysiert werden, wie beispielsweise *Hyperlinks* und *Hashtags*, wodurch komplexe Vernetzungsstrukturen zwischen den verschiedenen

Kommunikationsteilnehmern ermittelt werden können (vgl. Günther/Scharkow 2014: 112). Als Herausforderung bei der Arbeit mit Web-Inhalten gilt grundsätzlich die Beschaffung einer repräsentativen Stichprobe aufgrund des großen Umfangs der Inhalte im Internet (vgl. Lewis et al. 2013: 39). Bei Beiträgen aus den sozialen Medien kommt hinzu, dass die Archivierungsmechanismen unbekannt sind und unklar ist, ob die Kommunikationsforscher prinzipiell Zugang zu der Gesamtheit der Beiträge oder nur zu einem Ausschnitt davon erhalten (vgl. ebd.: 40). Kritisch ist nicht nur, dass die Inhalte sehr dynamisch und flüchtig sind, sondern auch, dass die Aussagekraft von nutzergenerierten Erhebungsdaten stark eingeschränkt ist, da die Äußerungen aktiver Online-Nutzer nicht repräsentativ für die Gesellschaft sind (vgl. Naab/Sehl 2014:129).

Es gilt festzuhalten, dass unabhängig von dem Anwendungsbereich mittlerweile zahlreiche datengetriebene Prozesse in der Kommunikationswissenschaft eingesetzt werden, die erlauben bestimmte Teilschritte oder ganze Abläufe der Inhaltsanalyse automatisch durchzuführen (vgl. Wettstein 2014: 17). Derzeitig werden in Studien, die digitale Methoden einsetzen, hauptsächlich einzelne Analyseaufgaben automatisiert (vgl. Eisenegger et al. 2020: 4, Stoll et al. 2020: 113, Boberg et al. 2020: 4, Burggraaff/Trilling 2020: 124). Abbildung 1 bietet in diesem Zusammenhang einen detaillierten Überblick über die verschiedenen inhaltsanalytischen Ansätze, die genutzt werden können.

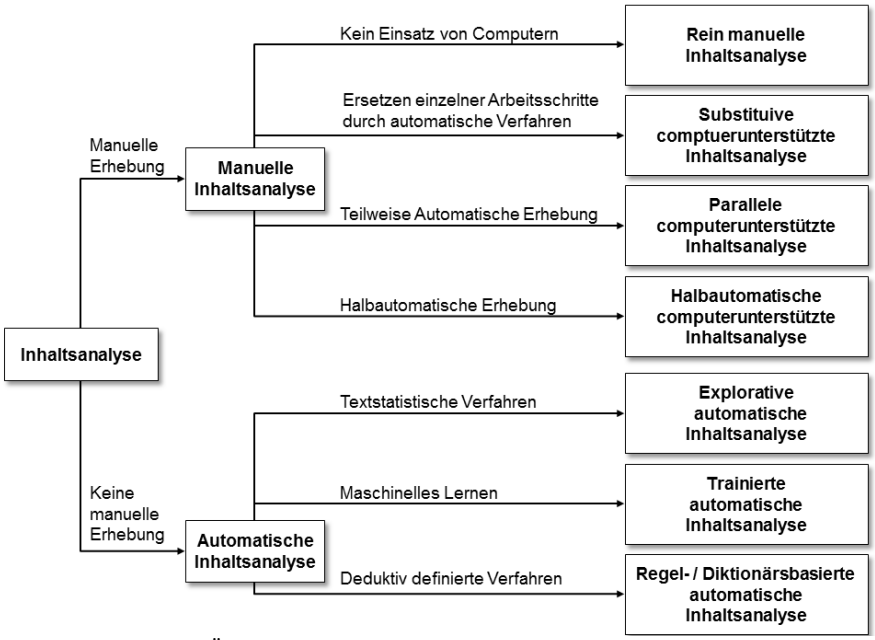


Abb. 1: Übersicht unterschiedlicher inhaltsanalytischer Ansätze (Quelle: Wettstein 2016: 25)

Die obige Abbildung verdeutlicht die Fülle und Vielfalt an Methoden bei der inhaltlichen Auswertung von Texten mit und ohne computergestützte Maßnahmen. Die schlussendliche Auswahl des Verfahrens hängt von der Forschungsfrage, den zu untersuchenden Variablen und der Datengrundlage ab (vgl. van der Meer 2016: 954).

Neben der Möglichkeit einzelne Teilaufgaben durch die Automatisierung zu ersetzen, können auch parallel manuelle sowie maschinelle Arbeitsschritte durchgeführt werden und ihre Ergebnisse kombiniert werden (vgl. Wettstein 2014: 18). Im nachstehenden Unterkapitel wird zunächst allgemein erläutert, welche Stärken und Schwächen die automatisierte Erhebungsform verglichen zu der klassischen, manuellen Methode aufweist. Anschließend werden auch die verschiedenen, oben abgebildeten, Verfahrensarten der AIA gegenübergestellt und das im Fokus dieser Arbeit stehende Verfahren eingeordnet.

2.2 Vor- und Nachteile der AIA

Bei der quantitativen Untersuchung von massenmedial verbreiteten Nachrichtentexten sehen sich Kommunikationswissenschaftler heutzutage mit einer gewaltigen Datenmenge konfrontiert, die nur sehr zeitintensiv und arbeitsaufwendig manuell bearbeitet werden kann (vgl. Maier et al. 2018: 63). Vor der eigentlichen Erhebung ist zunächst die Erstellung eines Codebuches notwendig. Darin wird mittels Kategorien definiert, welche Textbestandteile als Analyseeinheiten von Interesse sind, um die Forschungsfrage zu beantworten (vgl. Brosius et al. 2016: 157). Um diese Inhalte standardisiert und nachvollziehbar zu erfassen, müssen die verschiedenen Ausprägungsmöglichkeiten der Kategorien festgelegt werden. Durch dieses Kategoriensystem können dann systematisch die Textinhalte den entsprechenden Ausprägungen in Zahlenform zugeordnet und somit codiert werden (vgl. ebd.: 153). Mit diesen Codierungen wiederum lassen sich die Inhalte in einer einheitlichen Form sammeln und anschließend übergreifend auswerten. Bevor es zu der eigentlichen Codier-Arbeit kommen kann, müssen allerdings die Reliabilität und Validität des Erhebungsinstruments sichergestellt werden. Das heißt, dass überprüft werden muss, ob anhand des erstellten Codebuchs, unabhängig von Zeitpunkt und Codierer, stets die gleiche Codier-Entscheidung getroffen wird (vgl. ebd.: 51). Bei mehreren Codierern muss zusätzlich mittels zeitintensiver Schulungseinheiten und Pretests kontrolliert werden, ob die Texte auf dieselbe Art und Weise verstanden und codiert werden. Nur so kann die Übereinstimmung der Codierungen verschiedener Codierer, die Intercoderreliabilität, gewährleistet werden (vgl. Rössler/Geise 2013: 281). Mithilfe einer AIA können diese Codier-Entscheidungen jedoch von einem Computeralgorithmus übernommen und somit zeitliche und personelle Ressourcen erheblich eingespart werden (vgl. Rössler 2017: 195).

> „Das zentrale Anliegen seit den Ursprüngen der computergestützten Inhaltsanalyse war […] die eigentliche Codierung zu automatisieren, sodass man letztlich auf die Unterstützung von teuren, immer wieder neu zu schulenden, Fehler machenden und insgesamt schwer kontrollierbaren menschlichen Codiererinnen und Codierern verzichten kann.“ (Scharkow 2013: 290)

Ob die Algorithmen in einer AIA weniger Fehler machen als die menschlichen Codierer gilt es zu prüfen. Sicher ist jedoch, dass sie weitaus weniger Zeit für die Analyse von Textdaten benötigen. Ein Beispiel dafür ist eine im Jahr 2015 durchgeführte automatisierte Inhaltsanalyse aller englischen *Wikipedia*-Artikel, die zu dem Zeitpunkt 8,2 GB umfassten. Es wurden 16.000 Dokumente pro Minute verarbeitet, wodurch der Datensatz in knapp vier Stunden ausgewertet wurde (vgl. Rehurek 2015 zitiert nach Nunez-Mir et al. 2016: 1270). Eine Leistung, die manuell nicht ansatzweise in so kurzer Zeit durchführbar wäre.
Eine computergestützte Methode ist verglichen zu einer manuellen Erhebung leichter skalierbar, da sie problemlos auf größere digitale Textbestände angewendet werden kann, sobald sie einmal aufgesetzt ist (vgl. Lewis et al. 2013: 38). Die Nutzung von größeren Datensätzen kann wiederum dazu verhelfen, Untersuchungen mit geringem Umfang zu erweitern und so deren statistische Aussagekraft zu optimieren (vgl. van Atteveldt et al. 2019: 2).
Das Kategorienschema, nach dem die Texte analysiert werden, ist überdies bei einer AIA jederzeit erweiterbar und ohne viel Aufwand erneut nachträglich anwendbar (vgl. Brosius et al. 2016: 180). Neben der erhöhten Effizienz und Kapazität, weist sie auch eine erhöhte Reliabilität auf, denn „ein Computer codiert im besten Fall 24 Stunden am Tag und wird eine heute getätigte Zuordnung in einem Monat übereinstimmend wiederholen können“ (Rössler 2017: 200). Die Auswertung von größeren Textmengen, die über lange Zeiträume erstellt wurden, lässt daher Forschung auf ganz neuen Skalen sowie die Beantwortung andersartiger Fragestellungen zu (vgl. Lewis et al. 2013: 36).

Dennoch ist die Konzeption solcher automatisierten Methoden derzeitig durchaus arbeitsintensiv und komplex, da noch keine standardisierten Verfahren existieren und ihre Güte stark von den individuell programmierten Vorgaben abhängt (vgl. Rössler 2017: 200). Auch herausfordernd ist der generelle Zugang und die Aufbereitung der zu analysierenden Textdaten. Meist sind die Daten zu unstrukturiert für die sofortige maschinelle Verarbeitung und bei digitalen Nachrichtenbeiträgen erkennen Analysealgorithmen, verglichen zu menschlichen Codierern, nicht so leicht überflüssige Inhalte, wie Navigationselemente, Werbeanzeigen oder Leserkommentarspalten (vgl. Günther/Scharkow 2014: 112).
Wenn aufgrund von bestimmten Texteigenschaften die automatisierte Selektion der relevanten Inhalte nicht gelingt, beruhen die Kontrolle, die Optimierung der Verarbeitungsschritte oder gegebenenfalls die aufwendige, nachträgliche Datenbereinigung wieder auf den menschlichen Codierern und stellen keine Arbeitserleichterung dar (vgl. Wirth 2001 zitiert nach Günther/Scharkow 2014: 113). Dies wird in Kapitel 6 bei der Datenaufbereitung und Anwendung des automatisierten Verfahrens anhand einiger Beispiele aufgezeigt und im Detail ausgeführt.

Eine AIA eignet sich gut, wenn es bei der Erhebung hauptsächlich auf eine hohe Anzahl an Beiträgen, Vergleiche über längere Zeiträume oder die Häufigkeitsauszählung bestimmter Untersuchungseinheiten ankommt (vgl. Brosius et al. 2016: 175). Nützlich und zeiteffizient ist sie vor allem im Falle der Erkennung von Schlagwörtern oder der Operationalisierung von Forschungsfragen mittels Einzelworten und Wortkombinationen (vgl. ebd.).

Ein großer Vorbehalt gegenüber AIA liegt allerdings darin begründet, dass für zahlreiche Analyseaufgaben komplexe Inhalte erhoben werden müssen und dafür eine gewisse Sprachkompetenz sowie Interpretationsfähigkeit notwendig ist (vgl. Wettstein 2014: 16). Computeralgorithmen verfügen bisher noch nicht über menschliches Textverständnis, Weltwissen oder Abstraktionsfähigkeit (vgl. ebd.:17). Die inhaltliche Bedeutung der Wörter oder Sinnzusammenhänge werden nicht erkannt, wodurch die Bewertung von Sachverhalten erschwert ist. Ambiguitäten, doppelte Verneinungen oder rhetorische Stilmittel, wie Ironie, können nicht einfach erfasst werden (vgl. Rössler/Geise 2013: 271). Dies kann daher je nach Gattung der Texte, die es zu untersuchen gilt, zu mehr oder weniger großen Schwierigkeiten führen. „Der Spiegel ist beispielsweise schwerer mit einer AIA zu untersuchen, weil der Schreibstil oft mehrdeutig und relativ komplex ist“ (Brosius et al. 2016: 175).
Zusammenfassend lässt sich sagen, dass die automatisierte Inhaltsanalyse weitaus besser darauf ausgelegt ist, große Datensätze zu bearbeiten, aber nur begrenzt latente Bedeutungen oder die Feinheiten der menschlichen Sprache erkennen kann (vgl. Lewis et al. 2013: 37). Sie ist gegenwärtig noch nicht in der Lage solche Analyseaufgaben vollumfänglich zu übernehmen und dabei die Validität in dem Maße zu erfüllen, wie ein menschlicher Codierer (vgl. Schwotzer 2014: 63). Nichtsdestotrotz ist, je nach Forschungsfrage, bei einer Inhaltsanalyse nicht immer die Erfassung der Komplexität eines Textes gefordert, sondern vielmehr die bewusste Reduktion und gezielte Selektion von Informationen (vgl. Brosius et al. 2016: 191). In solchen Fällen stellt die AIA eine vielversprechende Möglichkeit dar.

Überdies wird in der Fachliteratur mehrfach betont, dass die automatisierten Methoden die etablierten manuellen Erhebungsmethoden der Kommunikationswissenschaft nicht ersetzen, sondern ergänzen sollen (vgl. Grimmer/Stewart 2013: 270). In diesem Zusammenhang äußern Boumans und Trilling: „automated methods are not equivalent to manual methods” (Boumans/Trilling 2016: 9). Dies verdeutlicht, dass mit automatisierten Analysen gegenwärtig keine gleichwertigen Erhebungen möglich sind. Sie sollen sinnvoll dort eingesetzt werden, wo manuelle Methoden aus Kapazität- oder Kostengründen an ihre Grenzen stoßen, um methodische Lücken zu schließen und damit zur Weiterentwicklung der Forschung beitragen (vgl. Nunez-

Mir et al. 2016: 1271). Durch einen kombinierten Ansatz können die Stärken beider Methoden genutzt werden, die Kontextsensitivität der traditionellen, manuellen Inhaltsanalyse sowie gleichzeitig die Kapazitätsvorteile, algorithmische Genauigkeit und Reproduzierbarkeit von computergestützten Methoden (vgl. Lewis et al. 2013: 38).

Ebenso vorteilhaft scheint die Möglichkeit die Erhebung geeigneter Kategorien aus manuellen Inhaltsanalysen durch automatisierte Analysen mit vergleichsweise geringem Codier-Aufwand fortzusetzen und somit durchgeführte Studien durch umfangreiche Langzeituntersuchungen zu erweitern (vgl. Günther/Scharkow 2014: 112). Andere Publikationen heben hervor, dass automatisierte Verfahren auch zur Methodentriangulation genutzt werden können (vgl. Graaf/van der Vossen 2013: 440). Damit gemeint ist ihr Einsatz bei der Beobachtung des gleichen Untersuchungsgegenstandes, um die Ergebnisse, die manuell erhoben wurden, zu bestätigen oder zu widerlegen (vgl. ebd.).

Wettstein betont darüber hinaus in der Publikation ‚*Best of both worlds*' die Vorzüge einer halbautomatisierten Inhaltsanalyse, bei der bestimmte manuelle Schritte des Analyseprozesses durch computergestützte Maßnahmen ergänzt und mittels eines geeigneten PC-Programms durchgeführt werden. Das Programm lernt aus den Eingaben des Codierers und wird dadurch kontinuierlich trainiert und verfeinert. Schließlich kann es den menschlichen Codierern automatisch Codier-Entscheidungen zur Überprüfung vorschlagen und sie dadurch entlasten. Notwendig sind eine Verbindung und ein ständiger Austausch zwischen der Eingabemaske für die Codierungen, den zu analysierenden Textdaten und dem Analyseprogramm (vgl. Wettstein 2014: 18). Ein standardisiertes System, welches die Erfassung von Textdaten, die automatische und manuelle Codierung und die Datenanalyse vereint, wäre sicherlich hilfreich. Die übergreifende Nutzung eines führenden Programms dafür hat sich allerdings noch nicht durchgesetzt, da es sich bei den meisten Untersuchungsmethoden um individuelle Lösungen für spezifische Forschungsfragen handelt, die nicht universell einsetzbar sind (vgl. Rössler 2017: 200).

Grundsätzlich sind nach Scharkow stets nur einzelne Analyseschritte automatisierbar, niemals die Forschungsmethode an sich (vgl. Scharkow 2012: 50). Damit gemeint ist, dass auch bei halb- oder vollautomatisierten Inhaltsanalysen weiterhin die Forscher mit dem nötigen Fachwissen die automatisiert erstellten Codier-Ergebnisse bewerten und letztlich den fundamentalen Forschungsbeitrag leisten, indem sie selber die logischen Schlussfolgerungen und das Wissen aus den Daten für den jeweiligen Anwendungskontext ziehen (vgl. ebd.).

2.3 Verschiedene Verfahrensarten der AIA

Neben den eingangs vorgestellten Automatisierungsstufen der AIA lassen sich innerhalb der existierenden Verfahren generell verschiedene Dimensionen zu deren Unterscheidung feststellen. So gibt es für die Analyse von Texten überwachte und unüberwachte Verfahren mit induktiven oder deduktiven Methoden und statistische sowie semantische Herangehensweisen.
In der Forschungsliteratur lässt sich dabei keine allgemeingültige, einheitliche Klassifizierung ausmachen. Häufig wird aber zwischen regelbasierten, trainierten und explorativen Verfahren unterschieden (vgl. Scharkow 2012: 58; vgl. Boumans/Trilling 2016: 8; vgl. Rössler 2017: 196). Abbildung 2 soll die verschiedenen Verfahrensarten zum Einstieg zunächst grob einordnen, um im Anschluss zu den verschiedenen Bereichen konkrete Anwendungsmöglichkeiten erläutern zu können.

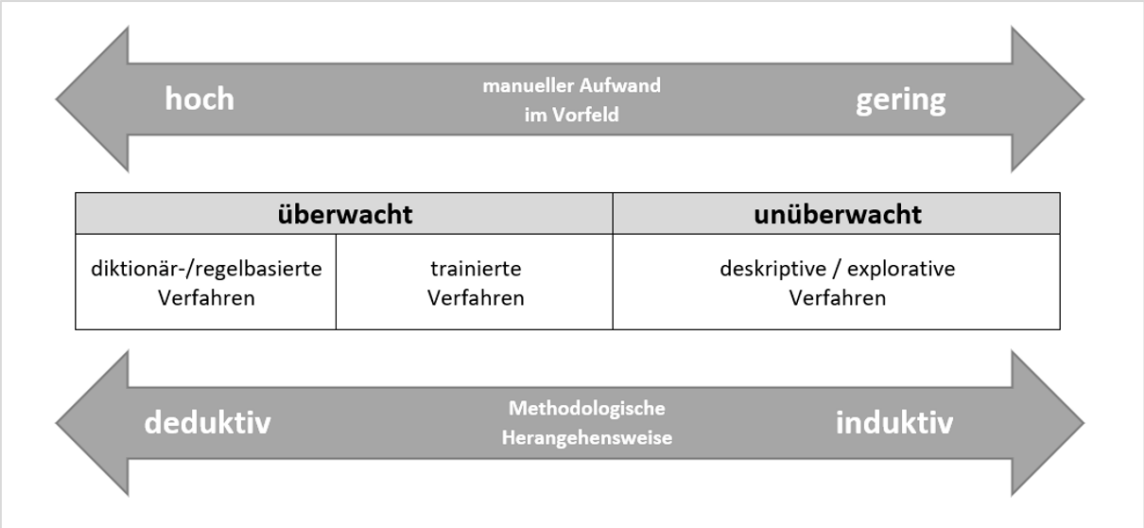


Abb. 2: Verfahrensarten automatisierter Inhaltsanalysen
(Quelle: Eigene Darstellung in Anlehnung an Rössler 2017:196 und Boumans/Trilling 2016:10)

Deduktive Ansätze werden hauptsächlich verwendet, um Inhalte basierend auf vorab definierten Kategorien zu analysieren, während induktive Ansätze angewandt werden, um unbekannte Muster zu erkennen (vgl. van der Meer 2016: 953). Bei Letzterem trifft der Computeralgorithmus die Entscheidung darüber, was in einem Datensatz bedeutsame Inhalte sind, während bei deduktiven Verfahren der Forscher dies definiert (vgl. Boumans/Trilling 2016:10). Solche überwachten Analysen versprechen daher eine höhere Validität und leichtere Interpretation der Ergebnisse, verglichen zu den autonom ablaufenden, unüberwachten Analysemodellen (vgl. Scharkow 2013: 291). Der Vorbereitungsaufwand vor ihrer Nutzung ist jedoch weitaus höher und zeitintensiver.
Bei dem anschließend zuerst vorgestellten, deduktiven Verfahren ist die Einflussmöglichkeit auf die automatisiert vorgenommene Codierung durch den Forscher somit noch vergleichsweise groß, während dies bei den später genannten Prozessen weiter abnimmt.

2.3.1 Diktionär- und regelbasierte Verfahren

Unter dem weit verbreiteten diktionärbasierten Verfahren wird die automatisierte Informationsextraktion mittels Schlagworten oder Wortlisten verstanden (vgl. Scharkow 2012: 60). Bei diesem deduktiven Vorgehen handelt es sich um einen einfachen Vergleich von bestimmten Zeichen oder Suchbegriffen, die im Vorfeld als maschinenlesbares Wörterbuch definiert werden (vgl. Wettstein 2014: 20). Um mit solch einem Verfahren die relevanten Inhalte in den zu analysierenden Textdaten automatisch zu identifizieren, können eigene Begriffslisten erstellt oder bereits verfügbare Wörterbücher genutzt und individuell adaptiert werden (vgl. Züll/Mohler 2001: 4). Darin müssen eindeutige Bezeichnungen sowie deren Synonyme hinterlegt sein und in bestimmten Fällen die Wortstämme der Begriffe aufgelistet werden, um die verschiedenen Deklinationen der Wörter miteinzuschließen (vgl. Rössler 2017: 198).

Ein klassischer Anwendungsbereich hierfür ist die computergestützte Erfassung von Themen journalistischer Texte. Diese basiert auf einer Vorauswahl von Schlüsselwörtern wie beispielsweise ‚Einbruch' oder ‚Mord' mit denen das Thema ‚Kriminalität' identifiziert werden soll (vgl. Schwotzer 2014: 59). Die Wortkombination ‚Selbstmord' müsste jedoch explizit ausgeschlossen werden, da sie den Begriff ‚Mord' beinhaltet, aber nicht in die Kategorie ‚Kriminalität' fällt (vgl. ebd.). Die Schwierigkeit liegt daher darin, mit den Schlüsselwörtern so viel wie möglich von einem Thema abzudecken, ohne aber zu viele Begriffe auszuwählen, die mehrdeutig sind oder in andere Themengebiete fallen (vgl. Lind et al. 2018: 4002). Deutlich wird hier, dass bei diesen Verfahren die Beschreibung der vorgegebenen Suchtermini stark vokabularabhängig und nicht leicht in andere Sprachen übertragbar ist (vgl. Boumans/Trilling 2016: 12).

Auch die Bestimmung der Tonalität oder Haltung von Texten kann mit diktionärbasierten Verfahren durch eine Stimmungsanalyse erfolgen. Notwendig dafür ist, dass bestimmte Begriffe vorab trennscharf in ‚positiv und negativ' oder ‚höflich und beleidigend' kategorisiert werden, sodass der Algorithmus bei der Identifikation der Schlagwörter die Textinhalte in diese Kategorien einordnen und mittels ihres anteiligen Vorkommens schließlich eine Klassifikation des Textes vornehmen kann (vgl. Graff/Theobald 2010: 207).

Um mit dem diktionärbasierten Ansatz wiederum die Nennung von Akteuren in einem Text automatisiert zu erkennen, müsste ein Wörterbuch mit allen Variationen der Namen dieser Organisationen und Personen sowie gegebenenfalls deren Berufsbezeichnung oder Positionen innerhalb einer Institution erstellt werden (vgl. Lind et al. 2019: 4002). Jegliche Informationen müssen bei diesem Verfahren im Vorfeld vorliegen, wodurch keine unbekannten Akteure in Texten identifiziert werden können.

Die Erkennung von unbekannten Akteuren wäre maximal durch den Einsatz regelbasierter Verfahren möglich, die sogenannte reguläre Ausdrücke beinhalten (vgl. Scharkow 2013: 300). Diese codierten Ausdrücke beschreiben die Suchkriterien nach logischen Regeln in formaler Sprache (vgl. Lane et al. 2019: 344). Ein Beispiel für solch einen Suchalgorithmus ist unten abgebildet und gelb markiert. Darin ist die Suche nach allen großgeschriebenen Wörtern in einem exemplarischen Text und deren Ausgabe definiert.

```
import re
text = 'Am 11. August wurde der weltweit erste Corona-Impfstoff in Russland zugelassen.'
regex = re.findall(r'(\b[A-Z][a-z]*\b)',text)
print(regex)
```

```
['Am', 'August', 'Corona', 'Impfstoff', 'Russland']
```

Abb. 3: Beispiel für einen regulären Ausdruck
(Quelle: Screenshot aus eigenem Python-Code in JupyterLab Interface)

Reguläre Ausdrücke sind sehr effizient, weisen jedoch auch Limitationen auf, da sie sich nur auf die Eigenschaften der Zeichen beziehen, aus denen die Wörter eines Textes bestehen. Bei der Suche nach zwei nacheinander stehenden großgeschriebenen Wörtern, könnten damit in zahlreichen Sprachen sehr schnell alle vorkommenden Vor- und Nachnamen oder Titel und Zunamen ermittelt werden. Nichtsdestotrotz würden auch viele irrelevante Ergebnisse erhalten werden, bei denen es sich nicht um Akteure handelt. Außerdem würden keine Personen, Parteien oder Institutionen identifiziert werden, deren Name nur aus einem Wort besteht.

Die diktionär- und regelbasierten Ansätze gelten als überwachte Verfahren, weil jegliche Handlungsvorschriften des Algorithmus durch den Forscher vorgegeben sind und alle möglichen Ausprägungsmöglichkeiten des Untersuchungsgegenstands festgelegt werden (vgl. Rössler 2017: 198). Da bei dieser Art der AIA allerdings nur vorher determinierte Inhalte erkannt werden können, sind die Einsatzmöglichkeiten und die Anwendungstiefe recht begrenzt (vgl. Stoll et al. 2020: 113).
Zusätzlich besteht das Risiko, dass nicht alle benötigten Ausprägungen einer Untersuchungsvariable erfasst werden, weil sie unzureichend definiert worden sind (vgl. Boumans/Trilling 2016: 12). Ebenfalls kritisiert wird, dass der Aufwand hinter der Entwicklung der Wörterbücher oder der Definition der regulären Ausdrücke mitunter größer als bei manuellen oder anderen automatisierten Verfahrensarten ist (vgl. Scharkow 2013: 300). Vor allem, weil die computerlesbaren Entscheidungsregeln und Wortlisten nicht auf den Codebuch-Definitionen vorheriger manueller Analysen aufbauen, sondern eine andersartige Definitionsarbeit und zusätzlichen Erstellungsaufwand bedeuten (vgl. ebd.). Eine andere Form des überwachten Verfahrens, welche keine vordefinierten Codier-Regeln benötigt, um Akteure in Texten zu identifizieren, wird nachfolgend vorgestellt.

2.3.2 Trainierte Verfahren

Bei dieser Art von Verfahren werden *Machine Learning*-Algorithmen eingesetzt, die anhand von speziell angefertigten Trainingsdokumenten mit richtigen Klassifikationen, eigenständig die Codier-Zuordnungen und -Regeln erlernen (vgl. Maier et al. 2018: 63).

Dieser Prozess ist nicht rein induktiv, da auch hier im Vorfeld manueller Aufwand nötig ist, um einen Trainingsdatensatz mit richtigen Zuordnungen und maschinenlesbaren Kennzeichnungen zu erstellen. Diese Kennzeichnungen werden ‚Annotationen' genannt und können verschiedene Informationen beinhalten, mit denen der Algorithmus trainiert wird.

Während die zuvor vorgestellten diktionär- und regelbasierten Verfahren mehr konzeptionelle Vorarbeit von dem Forscher erfordern, sind solche überwacht lernenden Verfahren vor allem auf viele und zuverlässig annotierte Trainingsbeispiele angewiesen (vgl. Scharkow 2012: 60). In diesen annotierten Beispieltexten werden von dem Algorithmus statistische Zusammenhänge und Strukturen erkannt, woraus ein Vorhersagemodell erstellt wird, welches letztendlich auf andere Testdaten angewandt werden kann (vgl. Kelm et al. 2020: 3). Mit jedem zusätzlichen Beispiel in dem Trainingsdatensatz kann der Algorithmus dazulernen und seine Leistung optimieren (vgl. Augenstein et al. 2017: 69).

Abbildung 4 stellt diesen Vorgang des überwachten maschinellen Lernens (ML) plakativ dar. Nachdem im ersten Schritt der Algorithmus annotierte Daten zum Training als Input erhält, werden im zweiten Schritt Daten ohne Annotationen genutzt, um zu überprüfen, ob diese von dem Algorithmus korrekt klassifiziert werden. Dieses Training sollte bis zu dem Erhalt zuverlässiger Kennzeichnungen durch den Algorithmus durchgeführt werden.

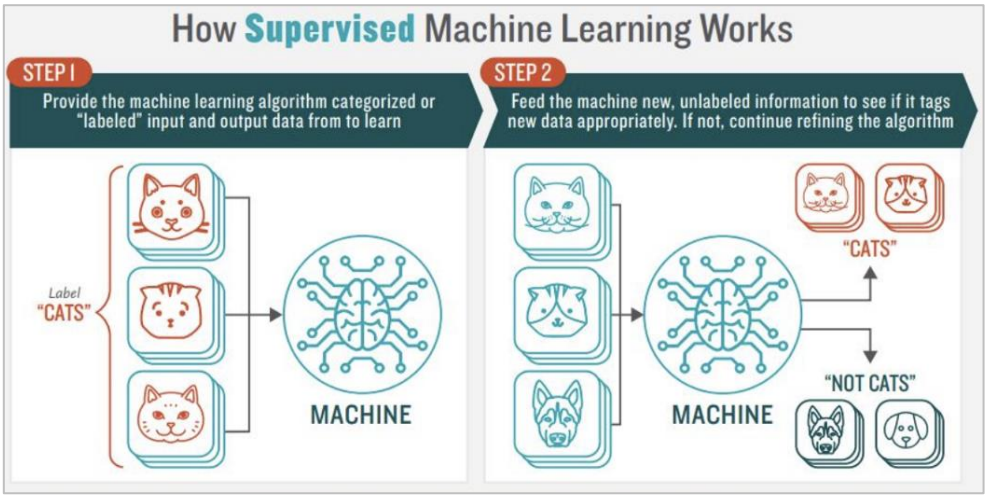


Abb. 4: Vereinfachte Darstellung maschinellen Lernens
(Quelle: Leonel 2018 - Supervised Learning - https://bit.ly/3lqrOMX)

Ein Vorteil dieser überwachten Methoden gegenüber dem diktionärbasiertem Ansatz besteht darin, dass sie einfacher in andere Themengebiete überführbar sowie leichter zu validieren sind (vgl. Scharkow 2013: 300).

Denn wenn eine bereits manuell vorgenommene Codierung von dem Algorithmus repliziert werden kann, liefert der Vergleich der Ausgabe der Maschinencodierung und der Handcodierung eine klare Bewertung (vgl. van der Meer 2016: 956). Nichtsdestotrotz besteht die Gefahr, dass die Vorhersagen des Algorithmus inkorrekt sind oder das erlernte Modell zu systematischen Fehlern führt, wenn die Annotationen des zugrundeliegenden Trainingsdatensatzes fehlerhaft oder nicht vollständig sind (vgl. Niekler 2016: 44).
Es kann auch zu einem sogenannten *overfitting* kommen, wenn die Ergebnisvorhersagen des Algorithmus mit dem Trainingsdatensatz sehr gut sind, doch bei den Testdaten sehr schlecht ausfallen (vgl. Kossen/Müller 2019: 123). Solch eine Ergebnisverzerrung kommt dadurch zustande, dass der Algorithmus sich auf die Inhalte der Trainingstexte spezialisiert und somit ‚überanpasst' (vgl. ebd.).
Im Trainingsprozess können dem *ML*-Algorithmus aber auch Gewichtungen für bestimmte erlernte Zusammenhänge mitgegeben werden, um die Relevanz bestimmter Daten zu verstärken oder zu ignorieren und somit das spätere Verhalten des *ML*-Modells zu beeinflussen (vgl. Stoll et al. 2020: 129). Dies zeigt, wie komplex der Vorgang hinter dem überwachten maschinellen Lernen von Computeralgorithmen ist und dass das umfassende sowie ausgewogene Training solch eines Algorithmus maßgeblich entscheidend für die Qualität der Ergebnisse ist (vgl. Wettstein 2016: 125).

Wie umfangreich solch ein Trainingsdatensatz sein muss, um beispielsweise Texte automatisiert in bestimmte Kategorien einzuordnen, wird in der Literatur unterschiedlich angegeben, da dies von der Komplexität der Untersuchungseinheit abhängt.
Wüest et al. schätzen etwa 100 annotierte Texte als ausreichend ein (vgl. Wüest et al. 2011: 8). Scharkow empfiehlt stattdessen eine Anzahl von etwa 300 Trainingsdokumenten zur automatisierten Erkennung bestimmter Themen wie ‚Politik' oder ‚Sport', um eine zuverlässige Klassifikation durch den Algorithmus zu gewährleisten. Um akzeptable Klassifikationsergebnisse für ein Thema wie ‚Kriminalität' zu erhalten, geht er von weit über 400 Trainingstexten aus, da es sich um eine komplexere Kategorie handele (vgl. Schwarkow 2011: 202). Boumans und Trilling geben eine Größenordnung zwischen 100-500 Trainingstexten an, verweisen jedoch gleichzeitig darauf, dass es bereits umfangreiche, annotierte Textkorpora für zahlreiche Sprachen gibt (vgl. Boumans/Trilling 2016:14).
Ebenso wie es frei zugängliche Code-Packages mit Algorithmen gibt, welche mittels dieser annotierten Textkorpora trainiert wurden, sodass Kommunikationswissenschaftler sie für ihre Analysen nicht selbst codieren und trainieren müssen (vgl. ebd.). Diese Code-Packages werden mehrheitlich auf *Open-Source*-Plattformen publiziert, wo die Verfasser auch die genutzten

Trainingsdatensätze angeben und die Anwendungsspezifika notieren. Wichtig ist in dem Zusammenhang die Überprüfung, ob die zu untersuchenden Datensätze der eigenen Analyse sich zu stark von den Trainingsdaten des zu nutzenden Algorithmus unterscheiden. Denn bei der Arbeit mit spezifischen Untersuchungsgegenständen, wie multilinguale Texte oder nutzergenerierte Inhalte, können die gängigen, bereits trainierten *Open-Source*-Algorithmen schlechte Klassifikationsleistungen aufweisen, da sie hauptsächlich mit Beispielen aus generischen, monolingualen Nachrichtenbeiträgen trainiert wurden (vgl. Eftimov et al. 2017: 5).
Für die im Fokus dieser Arbeit stehende automatisierte Erfassung von Akteuren wird ebendiese vortrainierte Verfahrensart genutzt und evaluiert. Es kann davon ausgegangen werden, dass die Nutzung von *ML*-Algorithmen sich gut eignet, da diese mit journalistischen Datensätzen trainiert wurden und hier auch an journalistischen Texten angewandt werden.

Andere kommunikationswissenschaftliche Anwendungsbereiche für überwachte, trainierte Verfahren sind die Identifikation von Nachrichtenfaktoren und *Frames* in der Berichterstattung. Das Phänomen des *Framing* beschreibt, wenn in einem Medienbeitrag eine spezifische Sichtweise eingenommen wird oder ein Thema kontextualisiert wird (vgl. Matthes 2008: 158). Diese Art und Weise der Darstellung kann bestimmte Aspekte eines Themas stärker betonen und somit die Einordnung und Meinungsbildung des Rezipienten beeinflussen (vgl. Maier et al. 2018: 138).
Während es als recht leicht eingestuft wird, einen Akteur als Variable zu codieren, ist die Erkennung von *Frames* komplexer, da es sich um eine abstrakte Kategorie handelt, welche ebenfalls für menschliche Codierer schwerer zu entdecken und definieren ist (vgl. Matthes 2008: 157). Dennoch existieren bereits verschiedene Ausarbeitungen in diesem Bereich, bei denen *Frames* wie ‚Wirtschaftliche Konsequenzen' oder ‚Konflikte' dank zahlreicher Beispiele dieser abstrakten Konzepte von Algorithmen automatisiert erkannt werden konnten (vgl. Burscher et al. 2014: 193; vgl. Matthes/Kohring 2008: 275). Für solche Untersuchungen, ebenso wie für Tonalitätsanalysen, wird eine weitaus größere Anzahl an Trainingstexten im vierstelligen Bereich empfohlen (vgl. Rudkowsky et al. 2018: 143).
Auch für die Identifikation von Nachrichtenfaktoren kann ein trainierter Algorithmus eingesetzt werden. Dieser kann beispielsweise die Ereignisorte in den Medieninhalten eigenständig ermitteln sowie anschließend automatisch die geografische Distanz zu dem Publikationsort des Artikels berechnen und somit den Nachrichtenfaktor ‚räumliche Nähe' bestimmen (vgl. Maier et al. 2018: 63).
Um allerdings Nachrichtenfaktoren wie ‚Überraschung' und ‚Prominenz' zu identifizieren, ist mehr Kontextwissen erforderlich (vgl. Scharkow 2011: 554). Solche Konstrukte sind für einen

Algorithmus schwierig aus Trainingstexten zu erlernen, weswegen sich dafür konkrete Vorgaben oder Wortlisten, wie bei dem diktionärbasierten Verfahren, besser eignen (vgl. ebd.).
In der Praxis sind diese überwachten Methoden am besten anwendbar, wenn umfangreiche annotierte Textkorpora zu ihrem Training genutzt wurden. Wenn jedoch kein Kategorisierungsschema oder Trainingsdatensatz verfügbar ist, kann eine unbeaufsichtigte Methode hilfreich sein, bei der relevante Textelemente induktiv gefunden werden (vgl. van der Meer 2016: 959). Solche unüberwachten Verfahren werden hierauf als letzte Verfahrensart vorgestellt.

2.3.3 Unüberwachte Verfahren

Unüberwachte Verfahren erfordern den geringsten Aufwand im Vorfeld der Analyse, da keine manuellen Regelspezifikationen für ihren Einsatz notwendig sind (vgl. Rössler 2017: 196). Im Gegensatz zu den wörterbuchbasierten und überwacht trainierten Ansätzen werden hierbei Muster und Wortcluster in einem Textdatensatz mittels unbeaufsichtigten maschinellen Lernens identifiziert. Statt nach vordefinierten Kategorien zu suchen, werden durch den Algorithmus eigene Zuordnungen vorgenommen. So liefert diese Methode dem Forscher zum Beispiel Informationen darüber, welche übergreifenden Themen in den analysierten Texten gefunden werden können (vgl. van der Meer 2016: 957). Dieses induktive Verfahren ermittelt eigenständig Zusammenhänge und grobe Strukturen und ermöglicht dadurch den schnellen Erhalt eines Überblicks über eine große Textsammlung. Es wird dabei in deskriptive und explorative Verfahren unterschieden.

Bei den deskriptiven Verfahren geht es um die einfache Ermittlung von Wortstatistiken, wie die Auszählung bestimmter Zeichen sowie die Bestimmung der Textlänge oder die Berechnung von Worthäufigkeiten. Dies kann von einem Computer, verglichen zu einem Menschen, weitaus schneller und reliabler erledigt und im gleichen Zuge grafisch dargestellt werden (vgl. Scharkow 2012: 61).

> „Obwohl die Beschreibung von Texten durch Häufigkeiten und Mittelwerte auf den ersten Blick trivial erscheint, können doch verschiedene interessante und wissenschaftlich relevante Konzepte mit textstatistischen Maßen operationalisiert werden.“ (Scharkow 2012: 61)

Durchführbar sind damit beispielsweise Wortschatzanalysen, bei denen die Wortfrequenzen aufgelistet und über verschiedene Medien hinweg verglichen werden, um Rückschlüsse auf die jeweiligen Kommunikatoren ziehen zu können (vgl. Brosius et al. 2016: 177).
Ebenso wird in Themenfrequenzanalysen dieses Verfahren angewandt, um zu analysieren, wie häufig im Zeitverlauf über ein bestimmtes Thema berichtet wird und somit auf dessen Relevanz in der Medienöffentlichkeit geschlossen (vgl. Niekler 2016: 7).

In einer Untersuchung von Fu et al. wurden beispielsweise die Anzahl der medial verbreiteten Nachrichten zu dem Thema ‚Zika-Virus' ermittelt und daraufhin die Anzahl der *twitter*-Beiträge zu dem gleichen Thema als Verlaufskurve darübergelegt. So konnte unter anderem erkannt werden, dass der Anstieg der auf den Zika-Virus bezogenen *Tweets* mit einer in den englischsprachigen Medien geäußerten Ankündigung der *World Health Organization (WHO)* zusammenhing (vgl. Fu et al. 2016: 1701).

Als komplexere unüberwachte Vorgänge gelten die explorativen Verfahren, wie die *Co-Occurence*-Analyse, bei der untersucht wird, welche Wörter gemeinsam auftreten, um daraus Wortnetzwerke und -cluster erstellt werden können (vgl. Waldherr et al. 2019: 6). Die zugrundeliegende Annahme bei der Betrachtung gemeinsam auftretender Wörter ist, dass diese auch semantisch zusammenhängen (vgl. Krippendorff 2004 nach Scharkow 2012: 66). Auf diesem Vorgang basiert ebenfalls das *Document-Clustering,* welches in unüberschaubaren Datenmengen ähnliche Dokumente oder Textklassen identifiziert und gruppiert (vgl. Rössler 2017: 197).

Als letztes Beispiel für explorative Verfahren ist das *Topic Modelling* anzuführen, welches ermöglicht einen Text auch mehreren Themen zuzuordnen. Diese Auswertung der inhaltlichen Themenzusammensetzung funktioniert auf der Grundlage einer statistischen Analyse der Kohärenz und Ähnlichkeit von Wortmustern (vgl. Boberg et al. 2020: 5). Häufig angewandt wird dabei das Wahrscheinlichkeitsmodell *Latent Dirichlet Allocation (LDA),* welches allen Signalwörtern eines Dokuments ein Thema mit Gewichtung zuordnet und somit die thematischen Anteile ermitteln kann (vgl. Blei 2012: 78). Die Abbildung 5 veranschaulicht exemplarisch, wie das statistische Modell die Verteilung errechnet und hebt farblich hervor aus welchen Themen sich diese zusammensetzt.

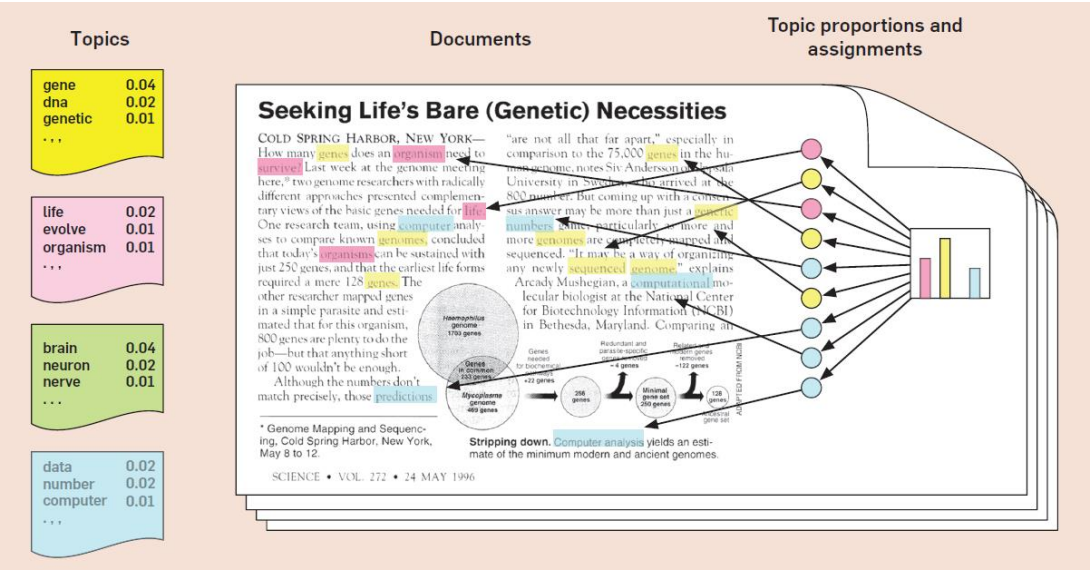


Abb. 5: Darstellung der Funktionsweise von LDA
(Quelle: Blei 2012: 78)

Grundlegend ist dabei die Hypothese, dass ein Thema stets die Wörter bestimmt, die für das Verfassen der Berichterstattung darüber genutzt werden. Demnach ist ein Text zu einer bestimmten Angelegenheit immer eine Mischung aus Wörtern, die für deren Erklärung und Darstellung notwendig sind (vgl. Niekler 2016: 70).

Auch die Identifikation von Akteuren in Texten ist unüberwacht möglich. Dabei werden nur wenige Beispielnamen benötigt, nach denen der explorative Algorithmus eigenständig in Datensätzen sucht. Die Grammatik und Satzstruktur der Sätze, in denen die Namen vorkommen, werden untersucht und kontextbezogene Hinweise gespeichert (vgl. Nadeu/Sekine 2007: 5). Damit versucht der Algorithmus andere Arten von Namen mit ähnlichen Eigenschaften zu finden, die in ähnlichen Kontexten auftreten. Dieser Lernprozess wird dann erneut auf die neu gefundenen Beispiele angewendet, um neue relevante Zusammenhänge zu entdecken (vgl. ebd.). Die gefundenen textstatistischen Korrelationen müssen jedoch von den Kommunikationsforschern im Detail geprüft und gedeutet werden (vgl. Scharkow 2012: 70).

Während bei unüberwachten Verfahren wenig manueller Aufwand vor deren Anwendung anfällt, muss nach ihrem Einsatz viel Arbeit in die Interpretation und Validierung der extrahierten Informationen gesteckt werden (vgl. Waldherr et al. 2019: 6). Es muss beurteilt werden, ob inhaltlich sinnvolle Zusammenhänge identifiziert sowie stimmige und ausreichende Cluster und Kategorien gefunden wurden. Genauso wie evaluiert werden muss, welche Aussagen die Ergebnisse über den Datensatz zulassen und ob sie neue Erkenntnisse für die Forschung liefern.
Bei den zuvor beschriebenen deduktiven Verfahren wird all dies vor ihrem Einsatz definiert, um gezielt Hypothesen zu prüfen, wodurch sie weitaus leichter zu interpretieren und validieren sind (vgl. Scharkow 2013: 291). Durch die Anwendung von unüberwachten Verfahren wird hingegen versucht Forschungsfragen und Hypothesen aus den zugrundeliegenden Daten abzuleiten. Aufgrund dieser Offenheit für unterschiedliche Interpretationen durch den Forscher, weisen diese unbeaufsichtigten Methoden größere Schwierigkeiten bei der Validierung auf (vgl. Boumans/Trilling 2016: 16).

Die verschiedenen bisher vorgestellten Verfahrensarten im Bereich des *Text Minings* sollen den Umfang und die Vielfalt an automatisierten Analysemöglichkeiten aufzeigen. Es wurde sichtbar, dass Akteure auf verschiedene Arten und Weisen in Texten identifiziert werden können. Bevor jedoch vertieft wird, welche maschinellen Verarbeitungsschritte dafür notwendig sind, soll hervorgehoben werden, warum die Ermittlung von Akteuren grundsätzlich für die Kommunikationswissenschaft von Bedeutung ist.

2.4 Relevanz der Akteursidentifikation in der Kommunikationswissenschaft

Um die konkrete Relevanz von *Named Entity Recognition* für kommunikationswissenschaftliche Forschungsfragen darzustellen, werden an dieser Stelle explizit der Nutzen hinter der Identifikation von Akteuren in journalistischen Texten herausgearbeitet und einige konkrete Anwendungsbeispiele aus gegenwärtigen Untersuchungen zusammengefasst.

Für eine grundsätzliche Analyse der Medieninhalte ist seit jeher bedeutsam, welche Personen, Unternehmen oder Organisationen in der Berichterstattung involviert sind (vgl. Schneider 2014: 41). Die vorkommenden Eigennamen in einem Text bieten dabei eine spezifische Informationsquelle, da sie meist den inhaltlichen Kern des Nachrichtenartikels darstellen (vgl. Hirschmann 2019: 50). In den Medien werden häufig Personalisierungen zur Komplexitätsreduktion verwendet und die automatisierte Erkennung von ihren Namen lässt bedeutsame Erkenntnisse über den Inhalt zu (vgl. Boberg et al. 2020: 12). Die Akteure sind häufig Handlungsträger in der journalistischen Berichterstattung und bestimmen das gesellschaftliche oder politische Geschehen (vgl. Rössler 2017: 140). Bestimmte Akteure, wie Greta Thunberg oder die AfD, stellen darüber hinaus nicht nur eine Person oder Partei dar, sondern werden als Repräsentant dezidierter Inhalte oder Meinungen angesehen (vgl. Boberg et al. 2020: 13).

Außerdem liegt bei der Identifikation von Personen mithilfe ihrer Namen meist keine große Sprachabhängigkeit vor, wodurch auch sprachübergreifende Studien möglich sind und die Berichterstattung in verschiedenen Regionen und Ländern verglichen sowie möglicherweise lokale Unterschiede beobachtet werden können (vgl. Niekler 2016: 2).

Viele Untersuchungen im Bereich der Journalismus- und Kommunikationsforschung befassen sich mit der Frage wie häufig spezifische Akteure erwähnt werden (vgl. Boumans/Trilling 2016: 11). Dabei werden nicht nur Analysen durchgeführt, die aufzeigen, wie oft ein Akteur in den Nachrichten vertreten ist, sondern auch wie sich dies zwischen den verschiedenen Medien unterscheidet oder im Zeitverlauf entwickelt. Dadurch können Aussagen über die Sichtbarkeit und Relevanz bestimmter Personen, Unternehmen oder Institutionen getroffen sowie Veränderungen in der Akteurskonstellation innerhalb der Berichterstattung erkannt werden (vgl. Strippel et al. 2018:7). Die Nennung von bestimmten Akteuren kann ebenfalls spezielle Phasen der Berichterstattung aufzeigen. Ein Beispiel hierfür ist die Untersuchung von Kolb zu der medialen Darstellung von Umweltproblemen durch Autoabgase. Seine Analyse verdeutlicht, dass in der Anfangsphase der Berichterstattung zunächst vermehrt Wissenschaftler im Zusammenhang mit der Thematik genannt werden. Ab einem gewissen Zeitpunkt wird das Thema in den Medien jedoch politisiert und ein auffälliger Rückgang wissenschaftlicher Akteure ist feststellbar, während weitaus mehr politische Akteure in den Beiträgen vorkommen (vgl. Kolb 2005: 207).

Es können überdies auch Interaktionsprozesse oder Verbindungen zwischen Akteuren in öffentlichen Diskussionen nachvollzogen werden, sodass Vernetzungsmuster sowie häufig genannte Personen oder Organisationen gar als Meinungsführer ermittelbar werden (Maier et al. 2014: 105). Meistens ist von Interesse, welche Akteure in den Medien genannt werden oder sich eigenständig Gehör verschaffen und in öffentlichen Debatten einbringen. Doch es kann auch aufschlussreich sein zu ermitteln, wer über keine öffentliche Stimme verfügt oder wessen Ansichten in den Medien kaum öffentliche Relevanz zugestanden werden (vgl. Brosius/Schwer 2008: 154).

Durch vergleichende Akteursanalysen in der Berichterstattung zu gleichen Themen von verschiedenen Medientiteln können Rückschlüsse auf ihre Qualität, Seriosität oder politische Ausrichtung gezogen werden. Weiterführend kann untersucht werden, ob die Auswahl der Akteure gemäß der redaktionellen Linie stattfindet. Ebenso interessant ist, ob mehr Akteure mit übereinstimmender als konträrer Meinung zitiert werden oder ob sich Akteure bereits stark genug in der Öffentlichkeit etabliert haben, sodass ihre Äußerungen zitiert werden, selbst wenn diese nicht mehrheitskonform sind oder mit der Auffassung der Redaktion übereinstimmen (vgl. Kepplinger 1989: 12).

Bei einer aktuellen AIA zu dem Thema Covid-19 wurden zum Beispiel die Facebook-Beiträge alternativer Nachrichtenmedien untersucht. Dabei handelt es sich um Medien, deren Urheber keine neutralen journalistischen Vermittler sind, „sondern politisch motivierte Bürger, die publizistisch ihre eigene Meinung vertreten“ (Schweiger 2017: 43). Bei der Untersuchung wurden aus über 115.000 Meldungen die 20 meistgenannten Akteure ermittelt. Dabei konnten die Kommunikationsforscher rechtspopulistische Darstellungen im Umgang mit der Thematik erkennen, da beispielsweise die AfD verhältnismäßig oft genannt wurde, obwohl sie keinen maßgeblichen Einfluss auf die getätigten politischen Entscheidungen und beschlossenen Verordnungen in diesem Zeitraum hatte (vgl. Boberg et al. 2020: 13). Auch die vermehrte Identifikation des türkischen Präsidenten Erdogan in den Beiträgen war auffällig, da dieser Akteur weitaus häufiger genannt wurde als andere Staatsoberhäupter angrenzender oder von der Pandemie stark betroffener Länder. Der sichtbare Fokus der Facebook-Beiträge auf Erdogan, im Zusammenhang mit seinem als bedrohlich dargestellten Beschluss, Flüchtlinge nach Europa fliehen zu lassen, ließ eine deutliche Anti-Migrationshaltung der Beiträge im Kontext der Coronakrise erkennen (vgl. ebd.: 12).

Durch die automatisierte Identifikation von Akteuren ist es somit möglich, interessante Erkenntnisse über die Medientitel und Kommunikatoren selbst zu erhalten, wenn in ihren Veröffentlichungen beispielsweise eine sehr eingeschränkte Akteursauswahl erfolgt.

Für einen gelungenen Meinungsbildungsprozess der Gesellschaft wird eine vielfältige Berichterstattung mit einer Vielzahl an Standpunkten und entgegengesetzten Sichtweisen als wertvoll angesehen. Daher kann das Spektrum der genannten Akteure mitsamt ihren Äußerungen als Indikator für eine gehaltvolle Berichterstattung gewertet werden (vgl. Schweiger 2017: 32). Bei einer Studie zu der Corona-Medienberichterstattung in der Schweiz wurde beispielsweise erhoben, welche Wissenschaftler in den Nachrichtentexten erwähnt werden, um das Maß an Diversität in der Berichterstattung zu überprüfen (vgl. Eisenegger et al. 2020: 10). Dabei wurde eine stark unausgeglichene Geschlechterverteilung aufgezeigt, sowie die Tatsache, dass Wissenschaftler von ausländischen Institutionen kaum Resonanz erhielten (vgl. ebd.: 15). Die Untersuchung des Vorkommens weiterer Akteure aus anderen gesellschaftlichen Sphären zeigte, dass in über 80% der Beiträge Akteure, wie Wirtschafts- und Regierungsvertreter oder Behördenrepräsentanten zu Wort kamen und ihre Ansichten und Forderungen kundtaten (vgl. ebd.: 11). Mittels eines Vielfaltsindex wurde die Repräsentation von solchen Experten zwischen verschiedenen Schweizer Medientiteln verglichen und Unterschiede in der Expertenauswahl je nach Beitragsstil und Medientyp festgestellt (vgl. ebd. 15). Deutlich wurde unter anderem, dass die untersuchten Online-Ausgaben von Abonnementzeitungen dabei überdurchschnittlich vielfältig sind und nicht so eine starke Konzentration auf bestimmte Akteure aufweisen wie die analysierten Printtitel (vgl. ebd.:14).

Auch in einer aktuellen Studie von Burggraaff und Trilling wurden, basierend auf der Untersuchung der genannten Akteure, Unterschiede in der Berichterstattung zwischen Online- und Printmedien sowie Populär- und Qualitätsmedien erkannt. Ihre Analyse holländischer Medientitel zeigt zum Beispiel auf, dass die Populärmedien weitaus mehr Personen referenzierten als die Qualitätsmedien oder dass Politiker häufiger in Online-Medien als in Printmedien vorkommen (vgl. Burggraaff/Trilling 2020: 121).

Beide zuletzt genannten Untersuchungen haben gemeinsam, dass die automatisierte Identifikation der Akteure in den Texten mit weiteren Arbeitsschritten ergänzt wurde, um bestimmte Eigenschaften der Akteure, wie ihre Prominenz, ihren Beruf oder ihre Reputation zu erfassen. Dies wurde nicht automatisch als Zusatzinformation erhalten, sondern musste nach der Erkennung der Eigennamen separat durch einen Abgleich mit einem Verzeichnis wie *Pubmed*, *DBpedia* oder *Wikipedia*, durchgeführt werden, um die dort hinterlegten Informationen über die entsprechenden Personen zu erhalten (vgl. ebd.: 120).

Der automatisierte Erhalt von Zusatzinformation bei der Identifikation von Akteuren weist daher großes Potenzial für künftige *NER*-Analysen auf. Derzeitig existierten dafür nur speziali-

sierte *Named Entity Linking* Tools, die mit angebundenen externen Datenquellen die Eigennamen in den Texten mit entsprechenden Wissens- und Informationsseiten verknüpfen (vgl. Maynard et al. 2016: 53).

Ein anderweitiger Anwendungsbereich, bei der die Extraktion von Eigennamen hilfreich ist, sind Medienresonanzanalysen, die untersuchen wie oft bestimmte Institutionen in der Berichterstattung vorkommen und damit beispielsweise die Medienreputation von Organisationen oder Institutionen, wie Universitäten, abbilden (vgl. Vogler/Schäfer 2020: 3148). Diese Analysen müssen sich dabei nicht nur auf das simple Vorkommen der Organisation beschränken, sondern können auch den Zusammenhang zu den getätigten PR-Aktivitäten abbilden oder den generellen Kontext untersuchen, in dem sie genannt werden (vgl. Boumans/Trilling 2016: 16).

Wenn bei der automatisierten Erhebung allerdings nur die Kookkurrenz anderer Wörter als Indiz für eine positive oder negative Berichterstattung über eine Person oder Institution ausgewertet wird, ist Vorsicht geboten. Es muss sichergestellt werden, dass die Äußerungen, die im Kontext genannt werden, sich tatsächlich auf die Akteure beziehen und diese in den Texten nicht bloß als Sprecher agieren, der sich zu einem Thema positiv oder negativ äußert (vgl. Eisenegger et al. 2020: 14).

Um bei der maschinellen Verarbeitung von natürlicher Sprache die Beziehungen zwischen den vorkommenden Wörtern und einzelnen Akteuren und somit automatisiert den Kontext zu erkennen, sind viel weitreichendere syntaktisch-semantische Analysen erforderlich. Van Atteveldt nutzte beispielsweise eine semantische Netzwerkanalyse, um die Darstellung und Rolle von politischen Akteuren in Zeitungsberichten zu ermitteln (vgl. van Atteveldt 2008: 50). In seiner Ausarbeitung wird deutlich, wie komplex die notwendigen Arbeitsschritte sind, damit maschinell identifiziert werden kann, wer das Subjekt oder Objekt in der Berichterstattung ist (s. Anhang [1], S. 100).

Solch eine umfangreiche computerlinguistische Analyse kann im Rahmen dieser Arbeit nicht durchgeführt werden, daher ist wichtig an dieser Stelle festzuhalten, dass bei der durchzuführenden Extraktion von Akteuren aus den journalistischen Texten nicht ersichtlich sein wird, ob sie mit eigenen Äußerungen zitiert werden, oder selbst Thema der Berichterstattung sind.

Um die grundsätzliche Komplexität der Prozesse hinter der automatisierten Identifikation und Extraktion von Akteuren aus Textdaten nachvollziehen zu können, werden die notwendigen computerlinguistischen Grundlagen im nächsten Kapitel ausgeführt.

3. Natural Language Processing

Das Ziel maschineller Verarbeitung natürlicher Sprache, besser bekannt als *Natural Language Processing (NLP),* ist es, die menschliche Sprache für Maschinen lesbar und verwertbar zu machen, damit diese den Inhalt und die Informationen daraus nutzen können (vgl. Beysolow 2018: 1). Durch *NLP* kann gesprochene und geschriebene Sprache computerbasiert erfasst werden, sodass Prozesse wie Übersetzungen, Beantwortung von Fragen oder Rechtschreibüberprüfungen automatisiert ausgeführt werden können (vgl. Shelar 2020: 324).

3.1 *NLP*-Grundlagen

Wie im zweiten Kapitel ersichtlich wurde, sind für die maschinelle Verarbeitung von Texten unterschiedliche Techniken möglich. Die Mehrheit der Verfahrensarten, die vorgestellt wurden, basieren auf dem sogenannten *Bag-of-words*-Ansatz, der einen Text als eine unsortierte Ansammlung von Wörtern versteht (vgl. Wettstein 2014: 22). Nur aus dem gemeinsamen Auftreten von Begriffen und bestimmten Worthäufigkeiten werden Informationen über den Text erhalten, ihrem Inhalt und ihrer Textumgebung wird keine Beachtung geschenkt (vgl. ebd.). Es wird angenommen, dass die isolierte Betrachtung von Wörtern oder Wortpaaren ausreichen kann, um die konnotative Bedeutung des Analysetextes zu erfassen (vgl. van der Meer 2016: 954). Es kann außerdem eine *tf-idf*-Berechnung durchgeführt werden, die abgekürzt für *term frequency - inverse document frequency* steht (vgl. Lane et al. 2019: 71). Sie berechnet die Häufigkeit eines Wortes in einem Dokument, anteilig zu dessen allgemeinen Vorkommen in allen vorhandenen Dokumenten und gewichtet es entsprechend. Dahinter steht die Annahme, dass ein Wort, welches oft in einem Text vorkommt, für diesen wichtig ist. Doch Wörter, die in vielen Dokumenten häufig vorkommen, wie Artikel oder Konjunktionen, sind nicht informativ. Wörter mit hohen *tf-idf*-Werten bilden somit meist die zentralen Themen eines Dokuments ab (vgl. Schneider 2014: 43). Bei dem *BoW*-Ansatz werden Texte unter anderem in große Matrizen überführt und die Worthäufigkeit vereinfacht in Zahlenform dargestellt, um zu erfassen, ob und wie oft sie in den Sätzen vorkommen. Abbildung 6 stellt exemplarisch solch eine simplifizierte Repräsentation von zwei Sätzen als Matrix dar.

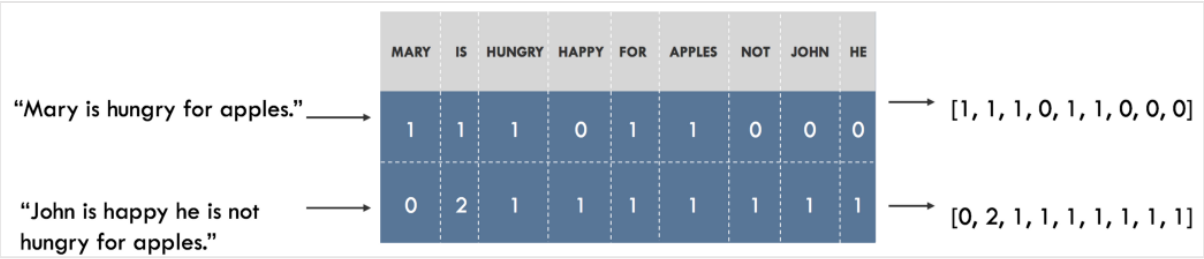


Abb. 6: Repräsentation von Text in Form einer Matrix (Quelle: Ameisen 2018 - https://bit.ly/2S3nEgq)

Die Abildung zeigt, dass in dem zweiten Satz die Zusammengehörigkeit der Wörter ‚*not*' und ‚*hungry*' verloren geht, da jedes einzelne Wort für sich steht. Da bei dieser lexikalischen Herangehensweise weder die Reihenfolge der Wörter im Satz noch der Kontext ausgewertet werden, müssen sich die zu bearbeitenden Fragestellungen durch Wörter und Wortkombinationen operationalisieren lassen (vgl. Rössler 2017: 196).

Bei der Informationsextraktion einzelner Begriffe wird in der Literatur von der Analyse von *unigrams* gesprochen, während bei der Identifikation längerer Wortkombinationen von sogenannten *n-grams* die Rede ist (vgl. Stoll et al. 2020: 115). Die Untersuchung von *unigrams* ist sehr gängig, kann jedoch nicht immer forschungsrelevante Ergebnisse liefern, da einzelne Begriffe je nach Kontext sehr unterschiedlich genutzt werden und andere Bedeutungen aufweisen (vgl. Scharkow 2013: 292). In vielen Fällen hilft bereits die Extraktion von *n-grams* gegenüber der traditionellen *unigram*-Analyse, um beispielsweise die Syntax von Negationen berücksichtigen zu können (vgl. Scharkow 2013: 292). Doch auch wenn viele Untersuchungen auf diesem rein wortbasierten *BoW*-Ansatz beruhen und sinnvolle Verarbeitungsergebnisse aufweisen, werden in *NLP* immer mehr computerlinguistische Ansätze eingesetzt, die weitaus mehr Sprachelemente berücksichtigen können (vgl. van der Meer 2016: 960).

Solche komplexeren *ML*-Algorithmen nutzen mehr Informationen als grammatikalische und erlernte statistische Zusammenhänge. Sie können die Wörter mittels neuronaler Netze in formale Sprache überführen und dadurch Informationen über die Semantik erhalten (vgl. Wettstein 2014: 24). Das bedeutet, dass Wörter in einzelne Vektoren umgewandelt werden, um ihre inhaltliche Bedeutung und Nähe zu anderen Wörtern mit ähnlichem Sinngehalt mathematisch abzubilden (vgl. Song et al. 2018: 24).

Abbildung 7 visualisiert anhand von drei Beispielen eine mehrdimensionale Darstellung von Wörtern als Werte im Raum. Je enger Wörter semantisch zusammenhängen, desto näher stehen sie als Vektoren zusammen. Semantisch unzusammenhängende Begriffe werden folglich auch räumlich weiter voneinander entfernt dargestellt.

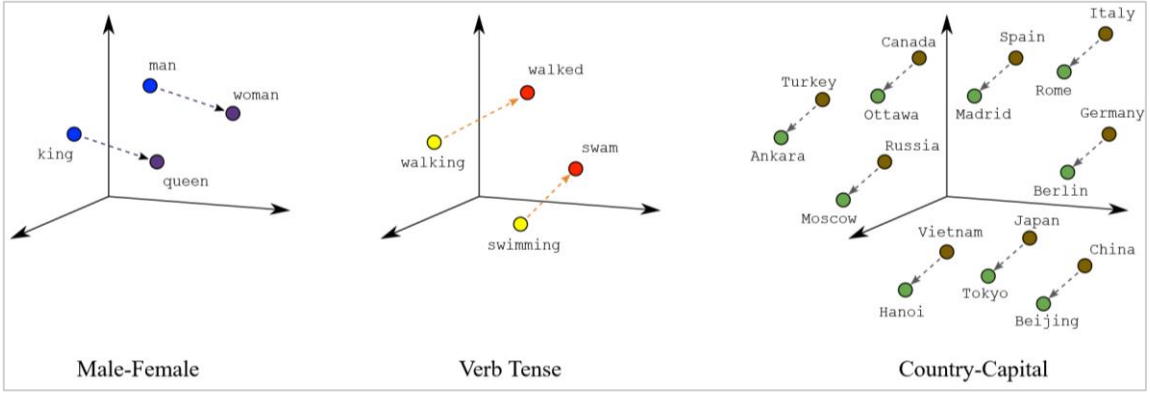


Abb. 7: Mehrdimensionale Darstellung von Wörtern als Vektoren in einem Raum (Quelle: Google Developers 2020 - https://bit.ly/3h3wRiP)

Diese Repräsentation von Wörtern als dreidimensionale Gleitkommazahlen ist ein relativ neuartiger Ansatz und wird als *word embeddings* bezeichnet (vgl. Rudkowsky et al. 2018: 140). Er ist stark sprachabhängig und benötigt eine erhöhte Rechenleistung, doch er ermöglicht eine präzisere Verarbeitung menschlicher Sprache (vgl. Stoll et al. 2020: 120). Im Vergleich zu der wortbasierten, statistischen Texterschließung des *BoW*-Ansatzes verarbeiten diese linguistischen Konzepte die Texte auf Satzebene. Durch die Nutzung neuronaler Netze können tiefergehende Muster und Zusammenhänge aus den Datenmengen erlernt werden. In diesem Kontext wird von *Deep Learning (DL)* gesprochen, welches ein Teilgebiet des *ML* darstellt.
Vor allem, wenn nur wenig Trainingsdaten für den spezifischen Anwendungsfall vorhanden sind, haben sich solche Ansätze in Vergleichsstudien als überlegenere Methodik im Umgang mit natürlicher Sprache erwiesen (vgl. Augenstein 2017: 70; vgl. Yadav/Berhard 2019: 1; vgl. Li et al. 2020: 5). In vielen Anwendungskontexten, unter anderem auch bei *NER*, werden daher gegenwärtig *word embeddings* genutzt, um die Semantik der Wörter zu ermitteln und die relevanten Bestandteile der Sätze exakter zu erkennen (vgl. Shelar et al. 2020: 325).

Unabhängig davon, ob in einem *NLP*-Prozess der *BoW*-Ansatz oder *word embeddings* verwendet werden, sind für die maschinelle Verarbeitung natürlicher Sprache grundsätzlich verschiedene aufeinanderfolgende Arbeitsschritte notwendig. Diese werden nachfolgend gebündelt dargestellt.

3.2 Verarbeitungsschritte in einer *NLP-Pipeline*

Voraussetzung für die Durchführung jeglicher *NLP*-Aufgaben ist die Errichtung einer sogenannten *Processing Pipeline*, welche am Anfang die unstrukturierten Textdaten einliest und alle anschließenden Verarbeitungsschritte umfasst (vgl. Lane et al. 2019: 4). Bei der am Ende erhaltenen strukturierten Ausgabe kann es sich um einzelne Elemente, Listen oder ganze Dateien handeln, je nachdem, was innerhalb der jeweiligen Prozesskette spezifiziert wird.
Die für diese Masterarbeit in Programmiersprache aufgesetzten *Pipelines* und erhaltenen Ausgabedateien stehen als separate *HTML*-Dateien zur Verfügung. An dieser Stelle soll anhand von Abbildung 8 eine exemplarische *Processing Pipeline* aufgeführt werden, welche die unterschiedlichen Komponenten abstrakt und übersichtlich illustriert.

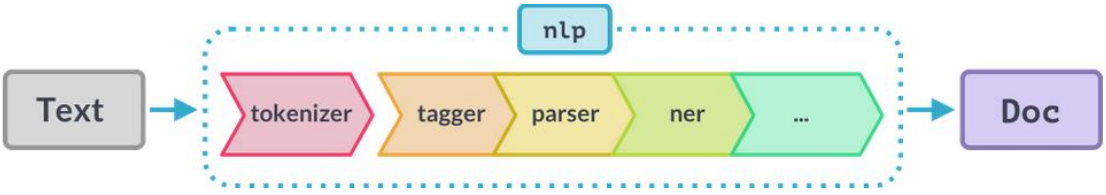


Abb. 8: Komponenten einer Processing-Pipeline
(Quelle: https://www.datacamp.com/community/blog/spacy-cheatsheet)

Als erste *Pipeline*-Komponente ist in der obigen Abbildung ein *tokenizer* aufgeführt, welcher den eingelesenen Datensatz, bestehend aus zahlreichen Texten, in einzelne Sätze segmentiert und die Sätze wiederum in einzelne Token zerteilt. Als Token gilt jeglicher Bestandteil eines Satzes, dazu zählen alle Wörter sowie Interpunktionszeichen und Symbole (vgl. Jockers/Thalken 2020: 238). Das Zerteilen der Textdaten stellt die Grundlage für alle weiterführenden Prozesse dar, die Satzstruktur und Reihenfolge der Wörter bleiben dabei erhalten.

Abhängig davon, was für Informationen aus einem Text gezogen werden sollen, müssen zunächst innerhalb der *Pipeline* Textbereinigungsschritte (*Preprocessing*) erfolgen. Für die allgemeine Auszählung von Worthäufigkeiten ist es beispielsweise hilfreich, die häufig vorkommenden ‚Stoppwörter' zu entfernen. Dabei handelt es sich um die zuvor thematisierten häufigen Wortarten, wie Präpositionen oder Auxiliarverben, die vermehrt in Texten vorkommen, aber keine wesentliche Bedeutung haben und keine Aussage über den Inhalt ermöglichen (vgl. Scharkow 2013: 295).
Für einige Auswertungen kann auch von Vorteil sein, alle Wörter auf ihre Wortstämme zu kürzen. Dieser Vorgang wird als *lemmatization* bezeichnet und steht für die Umwandlung von deklinierten Wörtern in ihre Grundform und von konjugierten Verben in ihren Infinitiv.
So werden beispielsweise aus Ausdrücken wie ‚In Trumps Rede nannte er…' oder ‚die Aufgabe des Deutschen Roten Kreuzes …' die Grundformen ‚Trump' und ‚nennen' sowie ‚deutsch rot Kreuz'. Durch diese Veränderung kann schneller und leichter zusammengefasst werden, welche Wörter sich wiederholen, ohne dass die verschiedenen gebeugten Formen einzeln gezählt werden. Wie an dem Begriff des ‚Deutschen Roten Kreuzes' sichtbar wird, kann *lemmatization* jedoch auch die Identifikation von Eigennamen beeinträchtigen.
Genauso kann das Beseitigen von Stoppwörtern in einigen Fällen zielführend sein und zu einer schnelleren Verarbeitung der Daten verhelfen. Es kann aber auch zu inhaltlichen Trugschlüssen führen, da Wörter wie ‚nicht' oder ‚kein' den Inhalt von Aussagen essentiell beeinflussen und nicht zwingend entfernt werden sollten (vgl. Lane et al. 2019: 291).
Dieses *Preprocessing* stellt Vorverarbeitungsschritte dar, die optional einsetzbar sind und maßgeblich die final erhaltene Ergebnisqualität bestimmen können (vgl. Kovalchuk et al. 2019: 22). Welche Bereinigungsabläufe nötig sind, muss daher vor jeder Analyse abgewogen oder im Prozess getestet werden.

In der obigen *Pipline* (Abb. 8) wird nach dem *tokenizing* das *tagging* als gängiger computerlinguistischer Schritt zur Datenverarbeitung aufgeführt. Mit einem *tagger* kann die Kennzeichnung der sogenannten *Parts-of-Speech (POS)* erfolgen.

Damit gemeint ist die Bestimmung der Wortarten der einzelnen Token als beispielsweise Nomen, Verben oder Adjektive (vgl. Jockers/Thalken 2020: 238). Der Beispielsatz in Abbildung 9 zeigt dessen Einteilung in elf Token und die durch *POS-tagging* zugeordneten Wortarten.

	0	1	2	3	4	5	6	7	8	9	10	11
Token	Das	Robert-Koch-Institut	in	Berlin	warnt	vor	der	unkontrollierten	Einnahme	von	Antibiotika	.
Wortart	DET	PROPN	ADP	PROPN	VERB	ADP	DET	ADJ	NOUN	ADP	NOUN	PUNCT

Abb. 9: Token und Wortarten eines Beispielsatzes
(Quelle: Screenshot des Outputs aus eigenem SpaCy Code)

In der Abbildung wird außerdem sichtbar, dass neben den Nomen ‚Einnahme' und ‚Antibiotika' die Wörter ‚Robert-Koch-Institut' und ‚Berlin' als *proper nouns* klassifiziert werden. Dabei handelt es sich um die Kennzeichnung sogenannter Eigennamen, die für *Named Entity Recognition* essentiell ist.

Als weiterer Verarbeitungsschritt kann ein *parser* eingesetzt werden. Als *parsing* wird der Vorgang bezeichnet, bei dem jeder Satz in seine syntaktischen Strukturen heruntergebrochen wird, um die Beziehung der einzelnen Satzbestandteile und ihre Abhängigkeiten untereinander abzubilden (vgl. ebd.). Jedem Token wird dabei ein individueller Status bezüglich seiner syntaktischen Eigenschaften in der Struktur zugewiesen (vgl. Hirschmann 2019: 52). Auch diese grammatikalischen Abhängigkeiten können für ein besseres Verständnis visualisiert werden. Abbildung 10 zeigt anhand des vorherigen Beispielsatzes auf, wie die einzelnen Token miteinander in Verbindung stehen. Es wird beispielsweise erkannt, zu welchen Wörtern die Präpositionen gehören oder mit welchen Nomen die Artikel zusammenhängen.

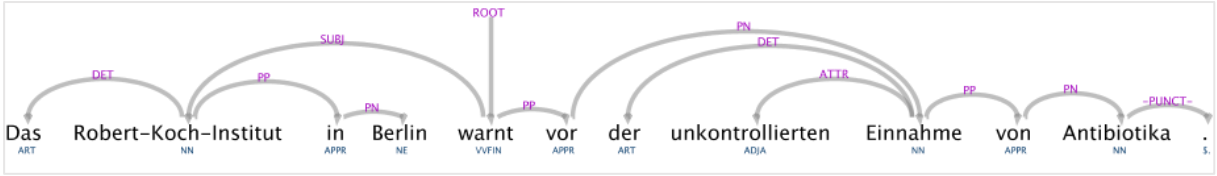


Abb. 10: Vereinfachte visuelle Darstellung des Syntaxbaums des Beispielsatzes
(Quelle: Screenshot des Outputs aus dem WebLicht Tool)

Eine wichtige Erkenntnis an dieser Stelle ist, dass es unterschiedlich schnelle und exakte *tagger* und *parser* gibt (vgl. Choi et al. 2015: 395). Dies ist abhängig davon, ob sie regelbasiert oder mittels komplexeren Verfahren des *ML* oder durch den Einsatz neuronaler Netze erstellt worden sind. Komplexere Algorithmen erzielen häufig präzisere Ergebnisse, doch gegenwärtig benötigen sie auch weitaus mehr Verarbeitungszeit (vgl. Vychegzhanin/Kotelnikov 2019: 76).

Viele derzeitig genutzte *tagger* arbeiten gewöhnlich datengesteuert und haben die *tagging*-Regeln automatisch anhand eines manuell annotierten Korpus aus Beispieltexten gelernt (vgl. van Atteveldt 2008: 45). Auch die geläufigen *parser* basieren auf riesigen annotierten Korpora, bestehend aus Text in natürlicher Sprache mitsamt maschinenlesbaren Kennzeichnungen, mit denen sie trainiert werden können (vgl. van Atteveldt 2008: 47).

Um die Token und ihre Beziehungen untereinander richtig zu klassifizieren, sind außerdem morphologische Herangehensweisen möglich, welche die äußere Gestalt der Wörter untersuchen. Dabei werden zum Beispiel Merkmale wie die Wortendung eines Tokens betrachtet oder ob es mit einem Großbuchstaben beginnt. Basierend auf diesen Beobachtungen wird die statistisch wahrscheinlichste Wortart vorhergesagt. Doch auch durch die Ermittlung der Position des Wortes in einem Satz sowie die Betrachtung der vorangehenden und nachfolgenden Begriffe kann kalkuliert werden, ob es sich beispielsweise um ein Subjekt oder Prädikat handelt (vgl. Rössler 2017: 198).
Die verschiedenen Mechanismen hinter *tagging* und *parsing* stellen die Basis für *Named Entity Recognition* dar und die dafür genutzten Algorithmen sind ebenfalls bei *NER* im Einsatz. (vgl. Maynard et al. 2016: 24). Auch für die Erkennung von Eigennamen existieren daher verschiedene Ansätze und Unterschiede in der Qualität der Identifikationsleistung. Um diese Unterschiede nachvollziehen zu können, werden in den nächsten Unterkapiteln die Funktionsweise von *NER* sowie die dafür nutzbaren Algorithmen, Trainingskorpora und Verfahrensmodelle aufgeführt.

3.3 *Named Entity Recognition* als Bestandteil von *NLP*

Wie eingangs erwähnt, wird *Named Entity Recognition* zur Extraktion von Informationen aus unstrukturierten Texten angewandt, mit dem Ziel benannte ‚reale Objekte' zu identifizieren, die aus Eigennamen bestehen (vgl. Marrero et al. 2012: 482). Dabei kann es sich beispielsweise um die Erkennung der Namen von Personen, Orten, Unternehmen oder staatlichen und nichtstaatlichen Institutionen sowie Parteien handeln (vgl. Schneider 2014: 41).

Das Konzept der automatisierten Erkennung von Eigennamen wurde bereits im Jahr 1996 auf der *Message Understanding Conference (MUC)* thematisiert, auf der die erste Definition von *Named Entities (NEs)* und die Konkretisierung ihrer maschinellen Identifikation erfolgte (vgl. Marrero et al. 2012: 482). Auf weiteren Konferenzen, wie der *CoNLL* (*Computational Natural Language Learning)* im Jahr 2003 und der *ACE (Automatic Content Extraction)* in 2008, wurden die Schwierigkeiten bei der allumfassenden Beschreibung und Erfassung von Eigennamen in verschiedenen Sprachen detailliert diskutiert, optimiert und evaluiert.

> "Since its introduction some twenty years ago, named entity (NE) processing has become an essential component of virtually any text mining application and has undergone major changes." (Ehrmann et al. 2020: 1)

Der Großteil der *NER*-Forschung wurde zunächst der englischen Sprache gewidmet, doch auch Modelle für die deutsche Sprache sind seit der *CoNLL-2003* Konferenz umfassend entwickelt

und optimiert worden (vgl. Nadeu/Sekine 2007: 2). Mittlerweile befasst sich ein Großteil der Ausarbeitungen sogar mit der Ermöglichung der Mehrsprachigkeit und der Sprachunabhängigkeit der Verfahren (vgl. ebd.).

Bei dem Einsatz eines *NER*-Verfahrens werden nicht nur die Eigennamen in einem Text identifiziert, zusätzlich werden diese auch einer bestimmten *NE*-Klasse zugeordnet (vgl. Eftimov et al. 2017: 3). In der Forschungsliteratur bestehen verschiedene Auffassungen dazu, ob nur die Erkennung (*detection*) oder auch die Klassifikation (*classification*) der Eigennamen unter den Begriff *NER* fällt (vgl. Maynard 2016: 25; vgl. Li et al. 2020: 2; vgl. Pinto et al. 2016. 4; vgl. Benikova et al. 2014: 2524).

Das zugrundeliegende Verständnis des *NER*-Begriffs in dieser Arbeit basiert auf der konkreten Anwendung des Verfahrens, welches die Identifikation und Klassenzuordnung der Eigennamen miteinschließt. Die vier gängigen *NE*-Kennzeichnungen umfassen die Klassen *PER, ORG, LOC* und *MISC* für die Kategorien Person, Organisation, Ort und Sonstiges (vgl. Faruqui/Padó 2010: 130). Bei dem zuvor gewählten Beispielsatz werden die zwei identifizierten Eigennamen mit folgenden Klassen gekennzeichnet:

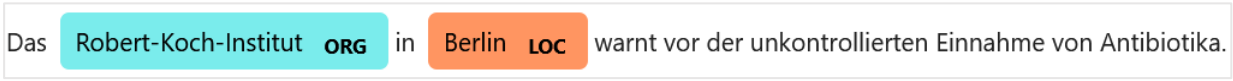


Abb. 11: Identifizierte Eigennamen im Beispielsatz visualisiert mit displaCy (Quelle: Screenshot des Outputs aus eigenem SpaCy Code – large model)

Der Eigenname ‚Berlin' wird erkannt und korrekt als Ort klassifiziert, ebenso wie die Wort-Kombination ‚Robert-Koch-Institut' ohne den Artikel markiert und der Klasse ‚Organisation' zugeteilt wird.

Je nach Verständnis und Definition davon, was als Eigenname gilt und in welche Klasse dieser einzuteilen ist, unterscheiden sich die verfügbaren *NER*-Verfahren (vgl. Marrero et al.: 484). So existieren *NER*-Algorithmen, die auf die Erkennung von über hundert verschiedenen *NE*-Klassen trainiert wurden (vgl. Li et al. 2020: 3). Diese anderen Kategorien können Wörter wie Mengenangaben, Nationalitäten, Jahreszahlen sowie Uhrzeiten umfassen (vgl. Wettstein 2014: 20). In südostasiatischen *NER*-Verfahren werden sprachbedingt auch Abkürzungen, Marken und bestimmte zeitliche Ausdrücke individuell gekennzeichnet (vgl. Yadav/Bethard 2019: 2).

Außerdem werden abhängig von dem wissenschaftlichem Analyseumfeld, in dem *NER* eingesetzt wird, auch domänenspezifische *Named Entities* definiert und die Algorithmen auf deren Erkennung trainiert. Unter anderem können in dem Forschungsbereich der Medizin, der Biochemie oder den Ernährungswissenschaften mit angepassten *NER*-Verfahren fachspezifische Begriffe wie Bakterien, Proteine, Gene oder Krankheitssymptome in Textdaten identifiziert werden (vgl. Vychegzhanin/Kotelnikov 2019: 72).

Darüber hinaus existieren mittlerweile Verfahren, die mit historischen Texten arbeiten können. Dafür werden vorab optische Zeichenerkennungsmethoden genutzt, um die altertümliche Schrift in Textdaten umzuwandeln und daraufhin *NER*-Verfahren einzusetzen, welche auf die veraltete Schreibweise von Personen- oder Ortsnamen, wie ‚Carolsruhe' statt ‚Karlsruhe' sowie die früheren Rechtschreibkonventionen spezialisiert sind (vgl. Ehrmann et al. 2020: 2).

Durch die Identifikation von *Named Entities* lassen sich mitunter auch einige der typischen W-Fragen (Wer, was, wo, wann?) beantworten, die die relevanten Ereignisse einer journalistischen Berichterstattung zusammenfassen (vgl. Marrero et al. 2012: 484). Deshalb eignet sich *NER* auch ideal als Grundlage für die Entwicklung von Chatbots und Sprachassistenten, da schnell alle relevanten Informationen maschinell erfasst werden können und so zum Beispiel aus Textbestandteilen in Emails direkt Kalendereinträge generiert werden können (vgl. Shreyas 2018: 1242).

Ebenso enthalten Nachrichten- und Verlagsdatenbanken große Mengen an Online-Inhalten, deren korrekte Verschlagwortung wichtig ist, um sie optimal nutzen zu können (vgl. Gasser et al. 2018: 181). Die Identifikation der *Named Entities* ermöglicht hier eine erste automatische Inhaltserkennung und ist auch umgekehrt nutzbar. Das bedeutet, dass bei der Eingabe von bestimmten Eigennamen relevante Beiträge identifiziert und angezeigt werden können. *NER* hilft somit bei der Klassifizierung von Inhalten und stellt damit eine wichtige Basis für jegliche Text- und Produktsuchmaschinen dar (vgl. Shelar et al. 2020: 325).

Unabhängig von dem letztlichen Einsatzgebiet bestehen die zu identifizierenden Eigennamen häufig nicht nur aus einem einzelnen Text-Token, sondern auch aus mehreren zusammengesetzten Wörtern oder Zahlen. Daher ist für ihre Verarbeitung die Betrachtung auf Token-Ebene nicht ausreichend, stattdessen müssen sogenannte *chunks* erfasst werden. Dabei handelt es sich um die gesamte Textspanne, die den Anfang und das Ende einer bestimmten Phrase umfasst (vgl. Li et al. 2020: 5). Abbildung 12 stellt solche *chunks* dar und verdeutlicht welche Token als Nominalphrase identifiziert und mittels *Named Entity Recognition* schlussendlich als Person klassifiziert werden.

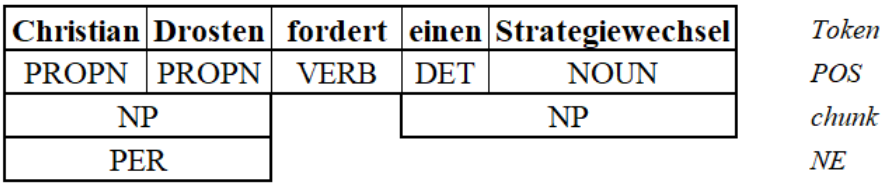


Abb. 12: Unterscheidung von Token, PO-Tags, Chunks und NEs
(Quelle: Eigene Darstellung in Anlehnung an Versley/Björkelund 2016: 245)

Das *Chunking* stellt einen grundlegenden Schritt für die korrekte Informationsextraktion aus Texten dar. Die Erkennung von Mehrwortsequenzen und Grenzen der *Named Entities* kann bei der Identifikation von Eigennamen einen maßgeblichen Einfluss auf die Ergebnisse erzielen (vgl. Kang et al. 2012: 17). Dies kann sich bei deutschsprachigen Texten schwieriger gestalten als in Texten, die auf Englisch oder in romanischen Sprachen verfasst sind. Dort stellt die Großschreibung von Eigennamen ein wertvolles Indiz für die *ML*-Algorithmen dar, während im Deutschen nicht nur Eigennamen, sondern auch Nomen großgeschrieben werden (vgl. Didakowski et al. 2007: 158).
Ein Beispiel, welches die damit verbundene Problematik aufzeigt, ist der kurze Nebensatz ‚*…weil Karl Software entwickelt*'. Durch die Großschreibung des Substantivs ‚*Software*' kann es bereits zu einer Fehlinterpretation kommen, sodass der Begriff als Nachname eingestuft und der Eigenname ‚*Karl Software*' ausgegeben wird (vgl. ebd.: 160).
Neben der Schwierigkeit der Erkennung der korrekten Grenzen von Eigennamen, ist die grundsätzliche Auffassung und Definition von *Named Entities* nicht trivial. Bei *NEs* handelt es sich im Wesentlichen um Namen realer Personen, Organisationen und Orte, die ein eindeutiges Bezugsobjekt haben (vgl. Maynard 2016: 27). Dies bedeutet, dass zum Beispiel der generische Begriff ‚Premierminister' streng genommen keine *NE* darstellt, da er sich auf eine Gruppe von möglichen Personen bezieht. Nichtsdestotrotz kann es hier zu einer *NE*-Klassifikation kommen, genauso wie einige *NER*-Verfahren abhängig von dem Kontext Wörter wie ‚Gott' oder ‚Jesus' als Eigennamen einordnen (vgl. ebd.). Je nach Anwendungsbereich des Verfahrens kann dies als relevantes, irrelevantes oder falsches Ergebnis gewertet werden.

Eine andere Herausforderung bei Textanalysen in Verbindung mit *NER*-Verfahren ist die Tatsache, dass nach einer ersten Erwähnung des Namens einer Person oder Organisation im weiteren Verlauf des Textes häufig mit Personal- und Possessivpronomen oder bestimmten Nominalphrasen Bezug auf diese genommen wird (vgl. van Atteveldt 2008: 98). So wird in der Berichterstattung für ‚Angela Merkel' auch die Bezeichnung ‚Merkel' oder ‚die Kanzlerin' gewählt. Diese Anaphern werden von *NER*-Algorithmen nicht automatisch als bereits bekannte Eigennamen zugeordnet.
Wenn jedoch für die Analyse relevant ist, an wie vielen und welchen Stellen Bezug auf die Akteure genommen wird, können mit einer sogenannten Anapher-Auflösung diese Ausdrücke in die Eigennamen umgewandelt werden (vgl. Wüest et al. 2011: 13). Dieser Zusatzschritt kann regelbasiert umgesetzt werden, sodass bei der Erkennung eines Pronomens geprüft wird, ob ein Eigenname mit dem entsprechenden Geschlecht in den vorhergehenden Sätzen erwähnt und daraufhin dadurch ersetzt wird (vgl. ebd.).

Auch hier sind mittlerweile datengetriebene *ML*-Algorithmen verfügbar und fähig, bei Bedarf diese Problematik zu lösen (vgl. Poesio et al. 2016: 98). Für die Analyse innerhalb dieser Arbeit wird keine ergänzende Anapher-Auflösung angewandt, da von Interesse ist, welche manuell codierten Akteure grundsätzlich erkannt werden. Dennoch kann es vorkommen, dass in einem Text bekannte Personen nach einmaliger vollständiger Namensnennung im Anschluss nur noch mit ihrem Nachnamen oder gar Spitznamen erwähnt werden (vgl. Wüest et al. 2011: 12). Damit auch in diesem Fall der Algorithmus diesen Eigennamen im gesamten Text als denselben Akteur identifiziert und zusammenfasst, müssen diese Pseudonyme entweder im Vorfeld einheitlich ersetzt oder im Nachhinein händisch zusammengefasst werden.

Als Letztes ist die Schwierigkeit der Zuteilung eines Eigennamens in die korrekte Klasse anzuführen. In der Forschungsliteratur wird zwischen interner und externer Evidenz unterschieden, wenn es um das Wissen geht, auf welches zurückgegriffen wird, um die Eigennamen korrekt zu klassifizieren (vgl. Rössler 2007: 36). Interne Evidenz beschreibt jegliche Hinweise zu dem Wort, welche aus lexikalischen Ressourcen gewonnen werden können. Wie die Tatsache, dass im Allgemeinen Wörter mit den Endungen -burg, -dorf oder -heim auf einen Ort hinweisen und der *NE*-Klasse ‚LOC' zugeordnet werden können. Die Mehrdeutigkeit von Wörtern erschwert ihre korrekte Klassifikation basierend auf bloßer Wortebene jedoch stark. Tabelle 1 zeigt exemplarisch einige Begriffe, die abhängig von ihrem Zusammenhang keinen Eigennamen darstellen oder unterschiedlichen *NE*-Klassen zuzuordnen werden können.

Wort	**Mögliche NE-Klasse oder Leseart**
Essen	Ortsangabe (LOC) oder Substantiv (keine NE)
Philipp Morris	Personenname (PER) oder Organisation (ORG)
Das Weiße Haus	Organisation (ORG) oder Ortsangabe (LOC)
Bauer	Personennane (PER), Beruf (keine NE) oder Schachfigur
Zeppelin	Personenname (PER) oder Objekt (keine NE)

Tab. 1: Mehrdeutigkeit von Wörtern
(Quelle: Eigene Darstellung in Anlehnung an Rössler 2007: 47)

Die externe Evidenz beschreibt hingegen die Hinweise, die aus dem Kontext auf Satzebene erhalten werden. Dabei helfen Namenszusätze oder Funktionsbezeichner, wie *die chinesische Stadt Xiaogan* oder *Landtagsfraktionschef Stoch*, unbekannte Wörter als Eigennamen zu identifizieren und korrekt zu klassifizieren (vgl. Didakowski et al. 2007: 160).
Insgesamt erfordert die Erkennung und Klassifikation von Eigennamen einen robusten Umgang mit unbekannten Wörtern, da es unmöglich ist, alle Bedeutungsvarianten jeglicher Wörter zu erfassen. Der menschliche Wortschatz erweitert sich ständig und in der journalistischen Berichterstattung treten regelmäßig Neologismen und neue Akteure auf (vgl. Maynard et al. 2016: 28).

Diese Komplexität und Mehrdeutigkeit von menschlicher Sprache bleibt weiterhin eine der Hauptherausforderungen aller *NER*-Verfahren (vgl. ebd.: 27). Die Untersuchung der vorliegenden Arbeit soll daher helfen zu beurteilen, wie stark die genannten Schwierigkeiten die Qualität der Ergebnisse einer *NER*-Analyse beeinträchtigen.

Nachdem in diesem Kapitel die grundsätzliche Funktionsweise und die linguistischen Herausforderungen bei *Named Entity Recognition* vorgestellt wurden, wird hierauf beleuchtet inwieweit sich die verschiedenen verfügbaren *NER*-Verfahren unterscheiden.

3.4 Unterscheidung verschiedener *NER*-Verfahren

Die Computerforschung, die darauf abzielt, *Named Entities* in Texten automatisch zu identifizieren, besteht aus einer immensen Auswahl an Verfahrensarten, Sprachmodellen und Algorithmen (vgl. Nadeau/Sekine 2007: 1). Wie in Kapitel 2 erläutert wurde, kann *NER* nicht explizit einer Verfahrensart zugeordnet werden, da die Identifikation von Akteuren in Texten auf allen vorgestellten Weisen möglich ist. Diese reichen von handgefertigten Definitionen bis hin zu Ansätzen des maschinellen Lernens.

Diktionär- und regelbasierte *NER*-Methoden identifizieren Eigennamen anhand menschlich definierter Wörterbücher oder logischer Regeln, die auf bestimmte Merkmale der Wörter abzielen (vgl. Eftimov et al. 2017: 5). Alle anderen Verfahren zur Informationsextraktion setzen auf den Gebrauch von *machine learning* und *deep learning* Algorithmen (vgl. Li et al. 2020: 4).

Unüberwacht lernende Algorithmen eignen sich besonders um neue, domänenspezifische *NE*-Klassen zu erkennen. Die überwacht lernenden Algorithmen hingegen basieren auf Trainingsbeispielen aus annotierten Textkorpora, die von menschlichen Experten erstellt werden. Ein Beispiel hierfür liefert die Abbildung 13, in der dargestellt ist, wie die vorhandenen Eigennamen anhand ihrer Position im Satz gekennzeichnet und die Informationen zu ihrer jeweiligen Klasse mitgeliefert werden.

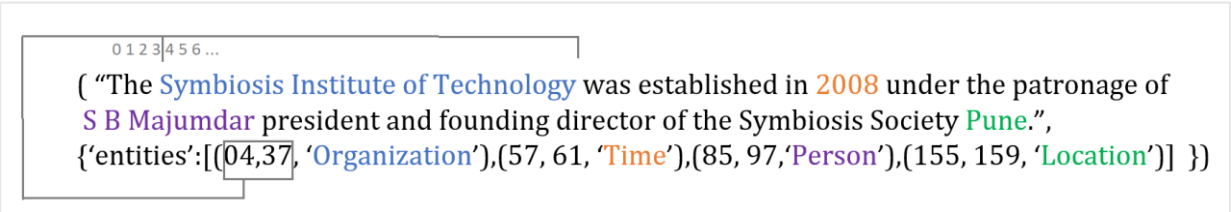


Abb. 13: Beispielhafte NE-Annotation eines Trainingstexts
(Quelle: Eigene Darstellung in Anlehnung an Shelar et al. 2020:8)

Nach dem Training mit solchen annotierten Daten werden von dem Algorithmus Vorhersagemodelle erzeugt, mit denen sie in anderen Texten die Eigennamen erkennen und deren entsprechenden Klassen bestimmen können (vgl. Eftimov et al. 2017: 5).

Für deutsche Analysen wurden diese überwachten *ML*-Algorithmen zu Beginn mit annotierten Nachrichtenartikeln der *Frankfurter Rundschau* trainiert (vgl. Faruqui/Padó 2010: 130). Nach und nach wurden zusätzliche Medientitel in den Trainingssatz hinzugefügt, ebenso wie deutsche *Wikipedia*-Einträge, *Tweets* und *YouTube*-Kommentare (vgl. Li et al. 2020: 2). In englischen Datensätzen wurden darüber hinaus *IBM*-Computerhandbücher, Pflegehinweise und transkribierte Telefongespräche annotiert (vgl. Taylor et al. 2003: 5).

Bei einer Ausarbeitung von Augenstein et al. wurden 19 verfügbare Trainingskorpora verglichen und zahlreiche Unterschiede herausgearbeitet. Die verschiedenen annotierten Textdaten unterscheiden sich nicht nur in ihrem generellen Umfang, sondern auch in der Zusammensetzung der dort auftretenden Eigennamen. Diese Zusammensetzung der Texte stellt einen wesentlichen Einflussfaktor beim Training der Algorithmen dar und der gewählte Textkorpus wirkt sich erheblich auf die Leistung der damit trainierten Algorithmen aus (vgl. Augenstein et al. 2017: 76).
Tabelle 2 soll beispielhaft darstellen, wie verschieden die verfügbaren Korpora sind und wie unterschiedlich die Ausgewogenheit der darin vorkommenden Eigennamen ausfallen kann:

Korpusname	Genre	Entitäten	PER	LOC	ORG
CONLL	Nachrichtenartikel	20.061	6.600	7.140	6.321
OntoNotes MZ	Magazintexte	8.150	2.895	3.569	1.686
MSM 2013	Twitter Beiträge	2.815	1.660	575	580
ACE CTS	Telefonkonversationen	2.667	2.256	347	64
ACE WL	Weblogs	1.716	756	411	549
MUC 7	Nachrichtentexte	552	98	172	282

Tab. 2: Vergleich verschiedener Textkorpora
(Quelle: Komprimierte Darstellung in Anlehnung an Augenstein et al. 2017: 63)

Dadurch, dass in einem Textkorpus das Vorkommen von Personen-, Organisationen- und Ortsnamen unterschiedlich hohe Anteile aufweist, kann auch die Klassifikationsleistung in den jeweiligen *NE*-Klassen uneinheitlich ausfallen (vgl. Maynard et al 2016: 35). Das bedeutet, dass das gleiche *NER*-Verfahren möglicherweise sehr gute Ergebnisse bei der Klassifikation von ‚Orten' in Texten erzielt, aber eine weitaus schlechtere Leistung bei der Erkennung von ‚Personen' aufweisen kann, weil im Trainingskorpus dazu weniger Lernbeispiele enthalten waren.
Es ist auch entscheidend aus welcher Zeit die Textdaten der Korpora stammen und ob viele verschiedene Textsorten darin vorkommen oder nur spezifische Genres, da dies ebenfalls die Art und Vielfalt der vorkommenden *Named Entities* bestimmt (vgl. Augenstein et al. 2017: 80). Textsammlungen von Nachrichtenagenturen stellen dabei eine beliebte Trainingsgrundlage dar, weil sie eine hohe Informationsdichte mit einer großen Proportion an Eigennamennennungen aufweisen (vgl. ebd.).

Doch auch die Themenzusammensetzung der Trainingskorpora kann entscheidend für die Ergebnisqualität des Algorithmus sein. Wenn dieser beispielsweise mit Zeitungsartikeln aus dem Bereich Wirtschaft und Politik trainiert wurde, kann er möglicherweise in Nachrichtentexten der gleichen Sparte präziser Eigennamen identifizieren, als in Interviews oder Beiträgen im Feuilleton.

Zusätzlich verweisen Li et al. darauf, dass auch aufgrund von unterschiedlich annotierten Trainingsdaten Verzerrungen bei der Leistung unterschiedlicher *NER*-Verfahren auftreten können. In ihrem Beispiel weist der Eigenname ‚*Baltimore*' in einem annotierten Trainingsdatensatz die Kennzeichnung ‚*LOC*' auf, da es sich um einen Ort in den USA handelt. Derselbe Eigenname wurde jedoch in einem anderen Trainingskorpus als ‚*ORG*' gekennzeichnet, weil das Wort darin in dem Kontext ‚*Baltimore defeated the Yankees*' einen Sportclub bezeichnet (vgl. Li et al. 2020: 15).

Gegenwärtig kann nicht sichergestellt werden, dass alle verfügbaren Textdaten einer Sprache einheitlich gekennzeichnet wurden (vgl. ebd.). Im Jahre 2014 wurden für die deutsche Sprache *Guidelines* entwickelt, um konsistent zu annotieren (vgl. Benikova et al. 2014: 2524). In diesen Richtlinien ist festgehalten, wie *Named Entities* zu kennzeichnen sind, sowie grundsätzliche Bestimmungen wie die Tatsache, dass Artikel, Titel und Anreden nicht Teil von Eigennamen sind (vgl. ebd.: 2529). Ältere deutsche annotierte Textkorpora können jedoch davon abweichen. Vor allem bei dem Umgang mit ineinander verschachtelten Eigennamen, sogenannten *Nested NEs*. Beispiele hierfür sind Orte oder Personennamen innerhalb von Vereins- und Organisationsbezeichnungen, wie beispielsweise ‚SV Werder Bremen' oder ‚Heinrich-Böll-Stiftung' sowie der dargestellte Eigenname in Abbildung 14.

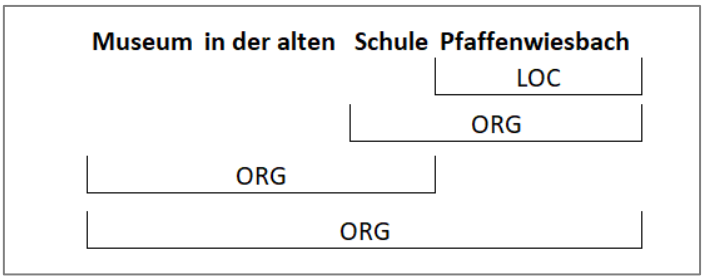


Abb. 14: Verschachtelter Eigenname
(Quelle: Eigene Darstellung in Anlehnung an Rössler 2007: 48)

In Abhängigkeit der erlernten Klassifikation kann das entsprechende *NER*-Verfahren beispielsweise das Museum, nur die Schule oder nur den Ort als *NE* extrahieren (vgl. Rössler 2007: 48). Je nachdem, wie solche eingebetteten *NEs* in den Trainingsdaten berücksichtigt werden, kann dies zu widersprüchlichen Ergebnissen bei der Anwendung unterschiedlicher *NER*-Verfahren führen (vgl. Rössler 2007: 49).

Die Art der Datenannotation sowie die genutzten Trainingstexte von *NER* und zahlreichen anderen *NLP*-Prozessen stellen somit ein elementares Unterscheidungsmerkmal der verschiedenen Verfahren dar.

Neben diesen Variationen innerhalb der verfügbaren Trainingskorpora, unterscheiden sich auch die grundsätzlichen *ML*-Algorithmen, die für *NER* ebenso wie für *tagging* und *parsing* eingesetzt werden, da sie auf verschiedenen statistischen Modellen beruhen. Eine detaillierte Erklärung ihrer Funktionsweise geht über den Rahmen der Arbeit hinaus. Dennoch soll ein verdichteter Überblick darüber gegeben werden, welche Arten von Algorithmen existieren, da sie vielfach in der computerlinguistischen Fachliteratur erwähnt werden und ein Kommunikationswissenschaftler bei der Auswahl eines *NER*-Verfahrens mit diesen Begrifflichkeiten konfrontiert wird.

Die *ML*-Algorithmen unterscheiden sich auf mehreren Ebenen. Eine bereits thematisierte Hauptdifferenzierung ist die Tatsache, ob sie überwacht oder unüberwacht lernen. Zusätzlich wird unterschieden, ob die Lernalgorithmen ihre Entscheidungen mittels Regression, Klassifikation oder Clustering treffen (vgl. Gilch/Schüler 2019: 36). Davon hängt letztlich das Vorhersagemodell ab, welches genutzt wird, um Satzbestandteile korrekt zu identifizieren und klassifizieren. In der Anwendungsliteratur werden häufig Modelle wie *Naive Bayes*, *Support Vector Machines (SVMs)*, *Hidden-Markov-Modelle (HMM)* oder bedingte Zufallsfelder, sogenannte *Conditional Random Fields* (*CRFs*) für *NLP*-Aufgaben eingesetzt (vgl. Song et al. 2018: 22). Abbildung 15 visualisiert die beschriebene Unterscheidung stark simplifiziert und listet in grün mögliche algorithmische Modelle auf.

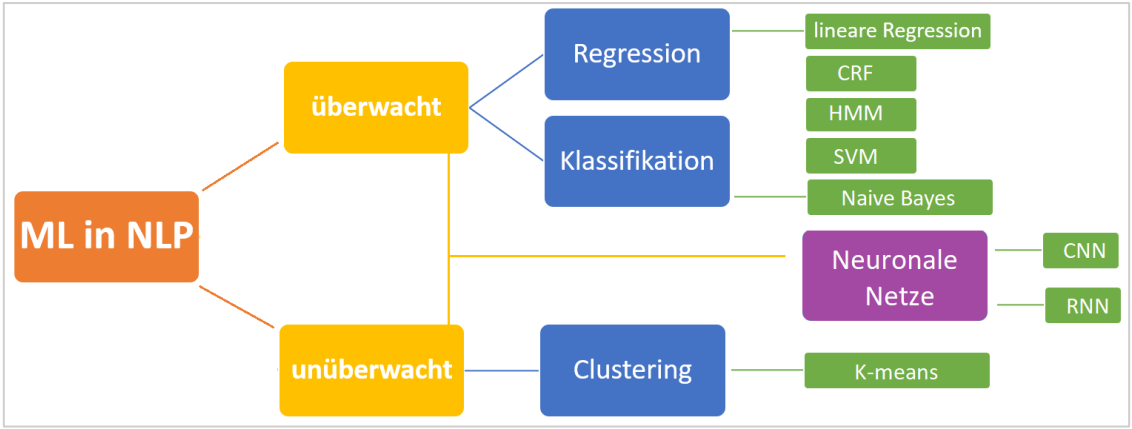


Abb. 15: ML-Algorithmen in NLP-Prozessen
(Quelle: eigene Darstellung in Anlehnung an Galimberti 2017 - https://bit.ly/30sPyXO)

Obwohl *Naive Bayes* oder lineare Regression aus recht simplen algorithmischen Ansätzen bestehen, eignen sich bei bestimmten Verfahren sehr gut. Sie übertreffen bei Klassifikationsauf-

gaben sogar komplexere Algorithmen, wie *SVM*, weil sie weniger zu *overfitting*, der Überanpassung an die Testdaten, neigen (vgl. Stoll et al. 2020: 120).
Die für *NER* häufig eingesetzten bedingten Zufallsfelder (*CRF*) wiederum sind in einigen Bereichen effektiver als beispielsweise *HMM*, da sie nicht von der Unabhängigkeit der einzelnen Token ausgehen, sondern den Kontext miteinbeziehen (vgl. Song et al. 2018: 24).
Kein einzelner Algorithmus eignet sich dabei ideal für alle *NLP*-Einsatzbereiche, da dies immer von den zu bearbeitenden Daten abhängt (vgl. Rudkowsky et al. 2018: 145). Es können jedoch auch mehrere Lernalgorithmen kombiniert werden, um eine bessere Leistung zu erzielen, als dies mit jedem der einzelnen Algorithmen für sich möglich wäre (vgl. Eftimov et al. 2017: 5).
In Abb. 14 sind außerdem künstliche neuronale Netze aufgeführt, diese *Convolutional* und *Recurrent Neural Networks* gelten heutzutage in der Informatik als *state-of-the-art* Werkzeug für die Textverarbeitung und sind aktuell das Kernforschungsgebiet für Lösungsansätze der Computerlinguistik (vgl. Stoll et al. 2020: 130). Es handelt sich dabei um fortgeschrittene, vielschichtige Netze, die, ähnlich wie die Neuronen eines Gehirns, eine Vielzahl an Reizen verarbeiten und unterschiedlich gewichten können (vgl. Lane et al. 2019: 156). Viele der verfügbaren *NLP*-Verfahren setzen solche künstlichen Netzwerke in Kombination mit anderen *ML*-Techniken ein und optimieren dadurch vorhandene Systeme der Informatik und Computerlinguistik wesentlich (vgl. Yadav/Bethard 2019: 1).

Es wird deutlich, dass zahlreiche Herausforderungen bei der maschinellen Verarbeitung menschlicher Sprache sowie der Erkennung und Klassifikation von Eigennamen in Texten bestehen. Nachdem die verschiedenen Einflussgrößen hinter *NER* dargestellt wurden und erkennbar wurde, wie komplex und vielfältig die zugrundeliegenden *ML*-Algorithmen sind, erfolgt im nächsten Kapitel die Auswahl eines konkreten *NER*-Verfahrens.

4. Auswahl eines geeigneten *NER*-Verfahrens

Wie bereits ersichtlich wurde, ist *NLP* ein aktives Forschungs- und Entwicklungsgebiet, in dem diverse Werkzeuge und Technologien existieren, die unterschiedliche Anwendungsfälle abdecken. Für die Auswahl eines geeigneten *NER*-Verfahrens stehen zahlreiche Möglichkeiten zur Verfügung. Es kann mit vorgefertigten Webdiensten, ausführbaren Dateien oder Quellcode gearbeitet werden (vgl. Vychegzhanin/Kotelnikov 2019: 72). In diesem Kapitel werden einige Beispiele dafür vorgestellt, um daraufhin die letztliche Auswahl eines geeigneten Verfahrens zu tätigen. Der Fokus liegt dabei auf der Anwendung bereits trainierter Verfahren, eine gängige Entscheidung, wenn weder genügend Erfahrung, Zeit oder Daten verfügbar sind, um das Verfahren eigenständig für einen bestimmten Zweck zu trainieren (vgl. Pinto et al. 2016: 14).

4.1 Einsatzbereite *NER*-Tools

Für deutschsprachige Texte gibt es verschiedene kommerzielle *Tools* von Anbietern wie beispielsweise *TXTWerk, Dandelion, Microsoft* oder *Google*, die unmittelbar einsatzbereit sind. Diese *Tools* erlauben das direkte Einfügen von Textdaten oder Hochladen bestimmter Dateiformate und die anschließende Erkennung von Eigennamen per Knopfdruck. Die Mehrheit dieser Dienste ist gebührenpflichtig, doch der Zugriff über eine Anwendungsschnittstelle kann meist während eines Probemonats unentgeltlich getestet werden (vgl. Gaus 2018: o.S.).
Es existieren auch kostenfreie Dienste, die im akademischen Umfeld entwickelt wurden, wie beispielsweise der Webservice ‚*WebLicht*' der Universität Tübingen. Abbildung 16 zeigt die recht nutzerfreundliche Bedienoberfläche, in welche die Textdaten eingefügt werden. Nach Angabe der Sprache des Textes und der gewünschten *Processing*-Schritte, wie *POS-tagging*, *parsing* oder *NER*, ist eine satzweise Analyse erhältlich.

Abb. 16: NLP-Tool WebLicht
(Quelle: Screenshot aus dem Interface - weblicht.sfs.uni-tuebingen.de)

Wenn keine tabellarische Ausgabe notwendig ist, aber eine schnelle Analyse des Textdatensatzes erfolgen soll, ist die Nutzung solch eines kostenfreien Tools weitaus zeiteffizienter als die eigene Erstellung einer *Processing Pipeline*. Der initiale Einrichtungsaufwand ist geringer, doch Heuss und Humm stellen bei einem Vergleich zahlreicher *NER-Tools* heraus, dass eine benutzerdefinierte *Pipeline* weitaus bessere Ergebnisse liefert als die anwendungsfertigen Dienste (vgl. Heuss/Humm 2014: 6).

Auch die Untersuchung von Alexander Gaus zeigt auf, dass die kommerziellen Webdienste sehr unterschiedliche Leistungen bei der Identifikation von Eigennamen aufweisen und im Vergleich zu manuell erhobenen Daten keine zufriedenstellenden Ergebnisse liefern (vgl. Gaus 2018: o. S.; s. Anhang [2], S. 100). Der Vergleich von Gaus weist darüber hinaus eine unterschiedliche Identifikationsleistung der *Tools* je nach Textsorte auf, bei der die Eigennamenerkennung in der Rubrik ‚Sport' die besten Ergebnisse erzielte (vgl. ebd.). Die gute Analyseleistung dieser Textsorte verspricht jedoch für die anvisierten Untersuchungen der Nachrichtentexte dieser Arbeit kaum einen Mehrwert.

Ein weiterer Nachteil solcher sofort betriebsbereiten Tools ist, dass selten transparent dargelegt wird, welche Prozesse darin ablaufen und welche Algorithmen und Trainingsdaten für ihre Erstellung genutzt wurden. Außerdem kann häufig nicht definiert werden, welche konkreten Textabschnitte der hochgeladenen Daten analysiert oder ignoriert werden sollen, was bei der Untersuchung unstrukturierter Daten aus unterschiedlichen Quellen und Medienformaten häufig Probleme verursachen kann. Genauso kann selten konkretisiert werden, in welchem Format die Ausgabe erstellt werden soll oder ob nur der Erhalt einer ausgewählten *NE*-Klasse von Interesse ist. Aus diesen Gründen wurde im Rahmen dieser Arbeit kein Webdienst getestet.

4.2 Modifizierbare *NER*-Bibliotheken

Alternativ zu den kommerziellen *Tools* kann mit Programmiercode eine *Pipeline,* wie in Kapitel 3.2 beschrieben, aufgesetzt werden. Der dafür verfügbare Code ist auf *Open-Source*-Plattformen wie *GitHub* oder *kaggle* erhältlich und weitaus flexibler an verschiedene Input- und Output-Formate von Textdaten adaptierbar. Immer mehr inhaltsanalytische Kommunikations- und Journalismusforschung stützt sich auf solche benutzerdefinierten, anpassbaren Programme, welche die große Vielfalt der verfügbaren Code-*Packages* sowie die dafür vorab trainierten Algorithmen ausnutzen (Boumans/Trilling 2016: 17).

Hierbei betonen Boumans und Trilling: "While we do not believe that every journalism scholar has to be a programmer, we deem some code literacy to be more and more useful" (Boumans/Trilling 2016: 17). Damit gemeint ist, dass, obwohl ein Kommunikationswissenschaftler

keinen Code selbstständig programmieren muss, gewisse Informatikkenntnisse durchaus notwendig und hilfreich sind, um erfassen zu können, welche Schritte darin ablaufen und wie sie gegebenenfalls angepasst werden können.

Entscheidend ist im ersten Schritt die Wahl eines Code-*Packages* in einer Programmiersprache, wie *Java*, *Python* oder *R*, sowie die Installation und Einrichtung der jeweiligen Entwicklungsumgebung, um den Code ausführen zu können. In den Fachartikeln und Aufsätzen zu automatisierten Inhaltsanalysen in der Sozial- und Kommunikationswissenschaft wurde eine stärkere Verbreitung bei der Nutzung von *Python* erkannt und daher für die Anwendung in dieser Arbeit gewählt (vgl. Burscher et al. 2014: 196; vgl. Lewis et al. 2014: 8; vgl. van der Meer 2016: 954; vgl. Rudkowsky et al. 2018: 143; vgl. Stoll et al. 2020: 131; vgl. Burggraaff/Trilling 2020: 118; vgl. Pinto 2016: 7).

Für das Gebiet der natürlichen Sprachverarbeitung gilt *Python* als führende Anwendungssprache und verspricht leicht erlernbar und gleichzeitig leistungsfähig einsetzbar zu sein (vgl. Severance 2015: 7). Für diese Programmiersprache gibt es wiederum verschiedene verfügbare *NLP*-Bibliotheken, die für *Named Entity Recognition* genutzt werden können. Solche Bibliotheken umfassen unterschiedliche Unterprogramme mit abrufbaren Funktionalitäten und bereits programmierten Methoden, um die entsprechenden *NLP*-Aufgaben durchführen zu können (vgl. Pinto et al. 2016: 38).

Angesichts der Vielfalt dieser Bibliotheken ist die Auswahl des zu verwendenden Verfahrens jedoch nicht einfach (vgl. ebd.: 2). Die Bibliotheken unterscheiden sich in ihrer Komplexität und Implementierungsschwierigkeit, ebenso wie in der Schnelligkeit und Effizienz ihrer genutzten Algorithmen. Einige Bibliotheken setzen voraus, dass der eingegebene Text mit anderen *Tools* segmentiert, mit Anmerkungen versehen oder in numerische Daten umgewandelt wird, da der Rohtext sonst nicht verarbeitet werden kann (vgl. Qi et al. 2020: 1). Dies, ebenso wie der Aspekt, welche Sprachen sie verarbeiten und *NE*-Klassen sie ermitteln können, schränkt ihre breite Anwendbarkeit auf Texte verschiedener Quellen ein (vgl. Vychegzhanin/Kotelnikov 2019: 72).

In der Forschungsliteratur existieren bereits vielfache Ausarbeitungen, in denen mehrere Bibliotheken gegenübergestellt werden. Außerdem erscheinen regelmäßig neue Publikationen, die verschiedene *NER*-Verfahren testen oder die Weiterentwicklung und Optimierung vorherrschender Techniken thematisieren. Dies kann dazu führen, dass Empfehlungen aus Untersuchungen, die vor zehn Jahren gemacht wurden, heutzutage als überholt gelten oder die genutzten Systeme und durchgeführten Vergleiche nicht mehr repliziert werden können (vgl. Maynard et al. 2016. 26).

Zur Bewertung der Leistung der Bibliotheken werden dabei unter anderem die benötigte Verarbeitungszeit der Prozesse oder die Qualität der Identifikations- und Klassifikationsleistung gemessen (vgl. Shelar et al. 2020: 324). Letzteres wird mit dem Gütemaß *F1-Score* berechnet, einem Wert, der das harmonische Mittel zwischen den Variablen *Precision* und *Recall* darstellt. Diese bestimmen die Exaktheit und Vollständigkeit der Ergebnisse (vgl. Li et al. 2020: 4).

$$F1\text{-}Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Der *Recall* gibt in diesem Fall an, wie viele der manuell identifizierten Eigennamen auch durch die automatisierte Methode gefunden werden. Die *Precision* misst dagegen den Anteil der korrekt klassifizierten Eigennamen innerhalb aller identifizierten *Named Entities* des automatisierten Verfahrens (vgl. Vychegzhanin/Kotelnikov 2019: 75). Der *F1-Score* liegt dabei in einem Wertebereich zwischen 0 und 1, wobei das Ergebnis 1 eine vollständige Übereinstimmung der manuellen und automatisierten Ergebnisse bedeuten würde (vgl. Derczynski 2016: 261).
Bei deutschen *NER*-Verfahren werden gegenwärtig, je nach genutztem Verfahren, Werte zwischen 0,64 und 0,86 erreicht (vgl. Qi et al. 2020: 6; vgl. Yadav/Bethard 2019: 5). Manuell erhobene Eigennamen stellen dabei den sogenannten ‚Goldstandard' an Vergleichsdaten dar. In zahlreichen Publikationen werden die *NER*-Ergebnisse jedoch nicht mit händisch ermittelten Daten, sondern mit anderen automatisiert identifizierten *NEs* verglichen (‚Silberstandard') oder das F-Maß mittels eines Durchschnittwerts über mehrere Kategorien angegeben (*micro-averaged F1-Score*), wodurch eine Vergleichbarkeit der Ergebnisgüte erschwert ist (vgl. Honnibal/Montani 2017: o. S; vgl. Qi et al. 2020: 6).
Da in vielen Untersuchungen die verglichenen Parameter nicht einheitlich gewählt und die Verfahren ständig angepasst werden, gibt es über alle Publikationen hinweg keinen eindeutig erkennbaren ‚*best performer*' (vgl. Pinto et al. 2016: 14). Aus diesem Grund wird beschlossen, nicht nur eine Bibliothek für die *NER*-Analyse zu testen, sondern drei verschiedene, vermehrt genutzte Bibliotheken, gegenüber zu stellen.

Das *Natural Language Toolkit (NLTK)*, welches häufig im akademischen Kontext genutzt wird und für datenwissenschaftliche Projekte zu den bekanntesten *NLP*-Bibliotheken für *Python* zählt, wird hierbei nicht einbezogen, da es in keiner aktuellen Vergleichsstudie am besten abschneidet und weniger fortgeschrittene Algorithmen in der Implementierung nutzt (vgl. Jiang et al. 2016: 24; vgl. Srinivasa-Desikan 2018: 34).
Die nachstehende tabellarische Übersicht (Tab. 3) listet diese und einige weitere Bibliotheken auf, die in *Java*- oder *Python*-Code programmiert sind und bildet ab, ob sie aus Projekten des industriellen oder akademischen Bereichs stammen.

Kommerziell-Industrieller Hintergrund	Akademischer Hintergrund
AllenNLP	Gimli
FLAIR	NERsuite
Gensim	NLTK
LingPipe	Polyglot
OpenNLP	Stanford CoreNLP
spaCy	Stanza

Tab. 3: Gängige industrielle und akademische NER-Tools
(Quelle: Eigene Darstellung in Anlehnung an Li et al. 2020: 3)

Grün markiert sind die Bibliotheken, die in der anschließenden automatisierten Analyse eingesetzt werden. Die drei ausgewählten *Open-Source*-Bibliotheken zählen in verschiedenen aktuellen Publikationen zu den überlegenen Verfahren bei der Erkennung von Eigennamen. Die Bibliothek *spaCy*, die auf kommerzielle *NLP*-Lösungen spezialisiert ist, glänzt durch die kürzeste benötigte Verarbeitungszeit (vgl. Shelar et al. 2020: 324; vgl. Lane et al. 2019: 353). Andererseits schneidet sie in manchen Fällen, aufgrund ihres Fokus auf die Effizienz, schlechter bei der Genauigkeit in der Identifikation von Eigennamen ab (vgl. Qi et al. 2020: 1). Dort weisen derzeitig *Stanza* und *FLAIR* bessere Ergebnisse auf, da sie komplexere *state-of-the-art* Algorithmen einsetzen (vgl. ebd.: 6; s. Anhang [3], S. 100). Nachfolgend werden alle drei Bibliotheken im Detail vorgestellt.

4.2.1 *spaCy*

spaCy wurde von dem Berliner Unternehmen *Explosion AI* entwickelt und ist eine Bibliothek, die derzeitig *NER* in 16 Sprachen unterstützt und dafür *word embeddings* und *Convolutional Neural Networks* einsetzt (vgl. Vychegzhanin/Kotelnikov 2019: 74). Die Bibliothek bietet für die deutsche Sprache drei verschieden große, bereits trainierte Modelle an, die mit unterschiedlichen Textdaten trainiert wurden (vgl. Honnibal/Montani 2017: o. S.). Das kleine *spaCy*-Modell wurde mit zwei Korpora trainiert, einem *Wikipedia*-Datensatz und dem *TIGER-Corpus*, bestehend aus Artikeln der *Frankfurter Rundschau*. Im Gegensatz dazu nutzt das große Modell, neben den ebengenannten annotierten Textkorpora, noch zwei weitere Webdatensätze als Trainingsgrundlage und setzt zusätzlich 500.000 Wortvektoren für die Eigennamenerkennung ein (vgl. ebd.).

Um zu ermitteln, ob sich die *NER*-Leistung der *spaCy*-Modelle stark unterscheidet, wurde probehalber mit einem Textdatensatz analysiert und ein Vergleich der erhaltenen Ergebnisse durchgeführt. Tabelle 4 zeigt auf, dass bei der Bestimmung der zehn meistgenannten Personen, die identifizierten *NEs* und die ermittelte Anzahl der Namensnennungen stark variieren. In Klammern ist zusätzlich aufgeführt, wie häufig die identifizierten Personennamen bei einer manuellen Prüfung des Textdokuments tatsächlich vorgefunden wurden. Es wird ersichtlich, dass

keines der beiden Modelle diese Werte replizieren kann, wobei das große Modell sich der tatsächlich vorgefundenen Anzahl stärker annähert.

SpaCy (small model)	Anzahl		SpaCy (large model)	Anzahl
Cioloș	**12** (21)		**Cioloș**	**13** (21)
Lieberman	**10** (14)		**Lieberman**	**12** (14)
Elke von Grabowski	**6** (10)		**Michiels**	**11** (13)
Meinolf	**5** (5)		**Tönnies**	**10** (13)
Michiels	**4** (13)		**Nielsen**	**8** (13)
Eberl	**4** (6)		**Künast**	**8** (14)
Henne	**4** (6)		**Schröder**	**8** (11)
Jany	**4** (7)		**Elke von Grabowski**	**7** (10)
Thomas Blaha	**3** (3)		**Hoddle**	**6** (8)
Olivia Judson	**3** (3)		**Eberl**	**5** (6)

Tab. 4: Gegenüberstellung identifizierter ‚PER' mittels kleinem und großem spaCy Modell (Quelle: Eigene Darstellung aus Auswertung in Anhang [H])

Das kleine Modell, welches mit weniger Trainingsdaten und ohne Vektoren arbeitet, erkennt nicht immer grundsätzlich weniger *Named Entities*, sondern nimmt an einigen Stellen schlicht falsche Eigennamenidentifikationen und -klassifikationen vor. Erkennbar ist dies in der nachstehenden Abbildung, bei der mit dem Visualisierungstool *displaCy*, welches auch Bestandteil der *spaCy* Bibliothek ist, die identifizierten Eigennamen und deren festgelegte Kategorien anschaulich hervorgehoben werden.

Sichtbar ist hier der bereits bekannte Beispielsatz und der dazugehörige Programmiercode, in den das kleine Modell ‚*de_core_news_sm*' geladen wird. Auch hier werden die Eigennamen ‚Berlin' und ‚Robert-Koch-Institut' identifiziert, doch im Vergleich zu der zuvor erhaltenen Ausgabe (Abb. 11), bei der das große Modell (‚*de_core_news_lg*') genutzt wurde, wird fälschlicherweise auch das Substantiv ‚Antibiotika' als ein Ort klassifiziert.

Abb. 17: Code und Ausgabe mit displaCy Visualisierung (Quelle: Screenshot des Outputs aus eigenem SpaCy Code – small model)

Dieses einfache Beispiel verdeutlicht den Einfluss des gewählten Modells auf die generelle Erkennungsleistung von Eigennamen. Wie in Kapitel 3.4 dargestellt, ist der Umfang ihrer Trainingskorpora und die Art der eingesetzten Algorithmen entscheidend für die erhaltenen Ergebnisse.

Aufgrund der erhaltenen Resultate wurde für die weitere Analyse ausschließlich das große *spaCy*-Modell eingesetzt.

4.2.2 *Stanza*

Die Bibliothek *Stanza* wurde von der *Standford NLP Group* entwickelt, welche bereits mit der *Java* Bibliothek *CoreNLP* langjährige Expertise bei der Entwicklung eines *NER*-Verfahrens erlangt hat (vgl. Qi et al. 2020: 1). Für *NER* unterstützt *Stanza* derzeitig acht Sprachen und nutzt ebenfalls ein statistisches Modell basierend auf bedingten Zufallsfeldern (*CRF*) und neuronalen Netzen sowie der Repräsentation von Wörtern als *word embeddings* (vgl. ebd.: 3).

Für die deutsche Sprache ist ein voreingestelltes Modell verfügbar, welches mit dem Textkorpus *CoNLL03*, bestehend aus Artikeln der *Frankfurter Rundschau* von 1992, trainiert wurde (vgl. Sang/Meulder 2003: 143). Es existiert außerdem ein Modell, welches auf einem Textkorpus namens *GermEval14* basiert, der aus *Wikipedia*-Artikeln und Online-Zeitungsnachrichten besteht (vgl. Qi et al. 2020: 5). Dieses weist in der Forschungsliteratur höhere *F1-Score*-Werte auf (s. Anhang [3], S. 100) und wird nach Anwendung beider Modelle aufgrund einer geringeren Fehlerquote für die finale Gegenüberstellung ausgewählt. Darüber hinaus gibt es biomedizinische und klinische *NER*-Modelle mit Klassifikationskategorien wie ‚Organe', ‚Aminosäuren' oder ‚Chemikalien', die eingesetzt werden können. Im Vergleich zu *spaCy* soll *Stanza* bedeutend mehr Verarbeitungszeit benötigen, jedoch viel höhere *F-Score*-Werte erzielen (vgl. ebd.: 6). Gleichzeitig wirbt die Bibliothek damit bessere Ergebnisse als *FLAIR* zu liefern, obwohl sie bis zu 75% kleinere und komprimiertere Sprachmodelle nutzt (vgl. ebd.).

4.2.3 *FLAIR*

Die Bibliothek *FLAIR* wurde von der Humboldt Universität zu Berlin in Zusammenarbeit mit einer *Open Source Community* und dem *Zalando Research Team* entwickelt. Es handelt sich dabei um eine relativ neue Bibliothek, die aufgrund der Nutzung neuronaler Netzstrukturen und eigens entwickelter *contextual string embedding* eine exaktere Verarbeitung natürlicher Sprache verspricht (Akbik et al. 2019: 54). Die Bibliothek bietet derzeitig trainierte *NER*-Verfahren in vier Sprachen an sowie ein multilinguales Modell und Spezifizierungen für ‚*biomedical NER*' sowie ‚*legal NER*'. Für die englische Sprache stellt *FLAIR* außerdem auch kleinere Modelle zur Verfügung, da die Anwendung der großen Modelle den Hauptprozessor eines Computers (*CPU*) stark beanspruchen kann (vgl. Akbik et al. 2018: 1645). Die zwei verfügbaren deutschen Sprachmodelle wurden zum einen mit dem Textkorpus *CoNLL03* und zum anderen mit dem *Germval18*-Datensatz trainiert (vgl. Akbik et al. 2019: 57). Mit letzterem Trainingssatz erreicht *FLAIR* nach Angabe der Entwickler jedoch keinen so hohen *F1-Score* (0,84) und wurde daher nicht für die nachfolgende Gegenüberstellung gewählt. Getestet wird das mit dem *CoNLL03*-trainierte Modell, welches in aktuellen Publikationen bei deutschsprachigen Textanalysen einen *F1-Score* von 0,88 erreicht (vgl. Akbik et al. 2018: 1645).

5. Durchführung der *NER*-Verfahren

In der Fachliteratur der *Computational Communication Science* existieren kaum Richtlinien zur Orientierung oder Nennungen von Standard-Vorgehensweisen bei der Durchführung von teil- oder vollautomatisierten Prozessen. Es wird jedoch angestrebt, dass bei der Anwendung einer automatisierten Methode das Vorgehen transparent und umfassend dargelegt wird. Nur so können die zahlreich getätigten Auswahlentscheidungen nachvollzogen und überprüft werden:

> "Computational analyses require many choices regarding design, preprocessing and parameter tuning, and transparency are needed to allow scrutiny of these choices. […] In publishing a description of a data set, it should be clear how it was gathered and preprocessed." (van Atteveldt et al. 2019: 3)

Aus diesem Grund werden in diesem Kapitel die durchgeführten Arbeitsschritte von der Bereinigung der Daten bis hin zu der Anwendung der verschiedenen *NER*-Bibliotheken beschrieben und die im Ablauf auftretenden Herausforderungen und Erkenntnisse aufgezeigt. Es wird Bezug auf die in Kapitel 3 erläuterten *NLP*-Verarbeitungsprozesse genommen und konkrete Beispiele aus dem Datensatz dafür geliefert.

5.1 Genutzter Datensatz

Bei den verwendeten Textdaten handelt sich um Online- sowie digitalisierte Print-Nachrichtenbeiträge des *SPIEGEL*, der *deutschen Presseagentur (dpa)*, der *Süddeutschen Zeitung (SZ)* und der *WELT* zu der ‚Corona'-Berichterstattung in dem Zeitraum von Januar bis Juni 2020. Die Materialbeschaffung und Festlegung der Stichprobe erfolgt für eine manuelle Auswertung an dem Lehrstuhl ‚Wissenschaftskommunikation in digitalen Medien' des KIT und beträgt dabei einen Umfang von insgesamt 1.887 Artikeln. Diese Stichprobe wurde durch eine Suche nach Begriffen wie *Corona*, *Covid* oder *SARS* auf den jeweiligen Nachrichtenplattformen erhalten. Dadurch erwies sich ‚Corona' nachträglich in einigen Artikeln nur als ein Randthema, weshalb diese Beiträge in dem manuellen Codier-Prozess zwar erfasst, doch für die weitere Untersuchung aussortiert werden. Dieser Vorgang kann nicht automatisiert nachgestellt werden, weshalb für einen sauberen Vergleich diese Artikel für die *NER*-Analyse händisch aus dem Datensatz aussortiert werden müssen (Artikel mit ‚Corona' als Neben-/Randthema = 685).
Während drei menschliche Codierer den Datensatz manuell in einem Zeitraum von zweieinhalb Monaten bearbeiten, wird überprüft inwieweit sich diese Datengrundlage für die automatisierte Bearbeitung eignet. Da es sich bei dem verwendeten Datensatz um rein deutschsprachige Nachrichtentexte handelt, konzentrieren sich die gewählten Verarbeitungsmethoden auf die formelle deutsche Sprache.

Von Vorteil ist dabei, dass die redigierten Zeitungsartikel grammatikalisch korrekt geschrieben sind, wodurch sich ihre automatisierte Analyse leichter erweist als die maschinelle Verarbeitung von multilingualen, informellen Texten (vgl. van Atteveldt 2008: 8). Für solche Textdaten müssten andere Verarbeitungsalgorithmen gewählt werden, welche mit Rechtschreibüberprüfungen und Synonymlisten unterschiedliche Orthographievarianten abgleichen können und mit den entsprechenden Datensätzen dafür trainiert wurden (vgl. Schneider 2014: 41; vgl. Stoll 2020: 119).

Die Nachrichtenbeiträge des *SPIEGEL* sowie der *dpa* und *WELT* können recht unkompliziert in ein maschinenlesbares Format umgewandelt werden, bei den Artikeln der *SZ* bereitet dies jedoch Probleme (s. Kapitel 5.2). Daher werden die Texte dieses Medientitels aus der letztlichen *NER*-Analyse ausgeschlossen (*SZ*-Artikel = 246). Des Weiteren wird sich auf die Nachrichtenartikel fokussiert, zu denen die manuellen Codierer tatsächlich Akteure erfassen (Artikel ohne Akteurscodierungen = 79). Dass nicht in allen Nachrichtentexten Akteure codiert werden, liegt darin begründet, dass die manuelle Codierung mittels eines bereits etablierten Codebuchs (s. Anhang [4], S. 101) erfolgt, welches klar definiert, dass Akteure nur dann zu codieren sind, wenn sie sich mit konkreten Aussagen zu der untersuchten Thematik äußern. Unterschieden wird dabei zwischen individuellen, institutionellen und generischen Akteuren (s. ebd.).

Dadurch gehen schlussendlich 887 Beiträge zu der Corona-Berichterstattung aus drei Medientiteln in die *NER*-Analyse mit ein (710 *dpa*-, 140 *WELT*- und 37 *SPIEGEL*-Artikel). Sie werden zu einem einzelnen Dokument zusammengeführt und ergeben eine Textdatei bestehend aus über 400.000 Wörtern (2,8 Millionen Token). Dieser Datensatz stellt die Grundlage für den Abgleich der automatisiert erhaltenen Ergebnisse mit den codierten Akteuren aus der manuellen Inhaltsanalyse dar. Ein deckungsgleicher Vergleich ist dabei nicht gegeben, da der *NER*-Algorithmus nicht mit der im Codebuch definierten Einschränkung auf Akteure mit dezidierten Aussagen zum Thema arbeitet. Dennoch ist von Relevanz, ob die manuell erfassten Personen und Organisationen grundsätzlich automatisiert identifiziert werden, welcher Bibliothek dies am besten gelingt und ob die erfasste Häufigkeit der Akteure über alle Artikel hinweg zwischen der manuellen und den automatisierten Methoden ähnlich ausfällt.

Außerdem wird für eine erste Erprobung und Vorbereitung des automatisierten Verfahrens mit einem zusätzlichen Testdatensatz gearbeitet. Dieser basiert auf vergangenen Inhaltsanalysen des Lehrstuhls ‚Wissenschaftskommunikation in digitalen Medien' über die Berichterstattung zu verschiedenen gesundheitlichen Risikothemen. Hierfür liegen bereits manuelle Auswertungen vor, sodass unmittelbar geprüft werden kann, inwieweit die ersten erhaltenen Ergebnisse der erstellten *NLP-Pipeline* mit diesen verfügbaren Codierungen übereinstimmen.

Aus der damaligen Stichprobe, bestehend aus 728 Nachrichtenartikeln, können 159 maschinenlesbare Texte des *SPIEGEL* und der *WELT* zu den Themen Antibiotika-Resistenz, Ebola und Grippepandemien als Testdaten genutzt werden (s. Anhang [C]). Sie dienen zur Übung der Prozesse der Datenbereinigung und -verarbeitung und werden zur Darstellung einiger Hindernisse in der Vorbereitung der Daten und der *NER*-Analyse genutzt. Außerdem helfen sie bei einigen Aspekten der Analyse des großen Datensatzes, die dort erlangten Erkenntnisse zu stützen. Auf Grund ihres geringen, weniger aussagekräftigen Umfangs werden sie jedoch weder im gleichen Detailgrad wie die Corona-Daten ausgewertet noch für den letztlichen Vergleich der drei NLP-Bibliotheken genutzt. Um die Übersichtlichkeit der Befunde der vorliegenden Arbeit sicherzustellen, werden die Ergebnisse dieses Testdatensatzes daher im Anhang ([27], S. 112) mit Erklärungen versehen und aus dem Hauptteil der Arbeit (Kap. 6) exkludiert.

5.2 Vorbereitung und Ablauf der Verfahren

Voraussetzung für die automatisierte Verarbeitung von Texten und Identifikation von Eigennamen ist die Verfügbarkeit von computerlesbarem Material (vgl. Brosius 2016: 174). Dieses kann direkt über große Datenbanken wie *LexisNexis* erhalten werden, die über spezielle Anwendungsschnittstellen einen Zugang zu strukturierten Daten bieten und aktuelle sowie archivierte Nachrichtentexte beinhalten (vgl. ebd.: 173). Die Artikel der *WELT* und des *SPIEGEL* wurden darüber bezogen. Auf die *dpa*-Artikel konnte direkt über die *dpa*-Nachrichtenplattform zugegriffen werden und die *SZ*-Artikel stammen aus der *SZ*-eigenen Datenbank.

Unabhängig davon, ob digitalisierte Hauptausgabeartikel oder genuine Online-Inhalte als Textdaten zur Analyse vorliegen, müssen sie alle in eine Form gebracht werden, die von Algorithmen vektorisiert und somit für alle weiteren *NLP*-Aufgaben verwendet werden kann (vgl. Patel 2020: o. S.). Die zu untersuchenden Daten, die in pdf-Format vorliegen, werden daher in einen einfachen Klartext konvertiert. Dies ist mit speziellen Softwareprogrammen oder auch kostenfrei über browserbasierte Dienste möglich. Alle für diese Arbeit verwendeten Programme und Systeme sind zum Zwecke der Reproduzierbarkeit im Anhang aufgeführt (s. [30], S. 115).
Wichtig ist zu prüfen, ob nach dieser Umwandlung die einzelnen Textbestandteile korrekt dargestellt werden. Die Artikel der *SZ* können nach dem Konvertieren bedauerlicherweise nicht genutzt werden, da bei der Konvertierung die einzelnen Textspalten in falscher und unzusammenhängender Reihenfolge in der txt-Datei zusammengefügt werden und eine manuelle Sortierung aller Textteile einen unverhältnismäßigen Zeit- und Arbeitsaufwand bedeuten würde (s. Anhang [7], S. 102).

Darüber hinaus tritt bei den anderen Artikeln anfangs das Problem auf, dass nach der Konvertierung der Textdaten teilweise Leerzeichen fehlen und die daraus entstehenden zusammengesetzten Wörter die Ergebnisse der *NER*-Analyse deutlich beeinträchtigen. Die entstandenen verbundenen Begriffe werden umgehend als Namen von Personen oder Organisationen klassifiziert und verfälschen die Identifikationsleistung des Verfahrens deutlich (s. Anhang [5], S. 101). Dieses Problem kann schließlich behoben werden, indem die Ursprungsdatei mittels einer anderen Software konvertiert wird, bei der alle Leerzeichen erhalten blieben.
Im Gegensatz dazu resultieren Markierungen von Wörtern in der pdf-Datei bei der Umwandlung in gesonderten Leerzeichen und zusätzlichen Textumbrüchen (s. Anhang [6], S. 101). Dies verzerrt die Testergebnisse nicht nachweislich, nichtsdestotrotz wird sichergestellt, dass in den später zu nutzenden pdf-Dateien der Corona-Nachrichtenartikel alle Markierungen vor der Textumwandlung entfernt werden.
Daraufhin erfolgt der Aufbau und die Prüfung der verschiedenen Elemente der *Processing Pipeline.* Der notwendige Programmcode wird in Zusammenarbeit mit einem Mitarbeiter des Lehrstuhls erstellt (s. Anhang [30], S. 115). Die Nutzung einer einzelnen *Pipeline* für alle drei anzuwendenden *NER*-Verfahren ist möglich, es wird allerdings entschieden pro Bibliothek ein getrenntes Programm zu erstellen und zu speichern. Dadurch kann bei der Durchführung das gesamte Programm ausgeführt werden, statt dass selektiv der jeweilige Code der entsprechenden Bibliothek ausgewählt werden muss.
In jeden der drei erstellten Programme werden eingangs die notwendigen Bibliotheken und Code-*Packages* importiert und anschließend die Textdatei aller zu analysierenden Nachrichtenbeiträge eingelesen (s. Anhang [A]).

Bei dem Einlesen und späteren Abspeichern der Textdaten muss zwingend auf das sogenannte *encoding* geachtet werden. Dies steht für das Format, in dem die Zeichensätze entschlüsselt werden. Damit wird beispielsweise in der deutschen Sprache bestimmt wie die Umlaute dargestellt werden. Bei den *WELT*-Artikeln zu dem Thema Antibiotika-Resistenz, funktioniert die gewählte Entschlüsselung der Sonderzeichen nicht. Dies führt dazu, dass die Umlaute nicht korrekt dargestellt und auch bei der Ausgabe der identifizierten *Named Entities* als ungültige Zeichen wiedergegeben werden. Zunächst wird davon ausgegangen, dass die fehlerhafte Entschlüsselung keinen Einfluss auf die Erkennung von Eigennamen hat und die Umlaute im Nachgang durch ‚Suchen und Ersetzen' bereinigt werden können.
Es fällt jedoch auf, dass einige Wörter und Begriffe als Eigennamen extrahiert werden, bei denen es sich nicht einmal um Nomen handelt. Tabelle 5 bildet einen Auszug der erhaltenen

Ergebnisse ab und zeigt, inwieweit ein falsch verschlüsselter Text, die Identifikation und Klassifikation von Eigenamen beeinflussen kann.

Identifizierte Personen (spaCy)	Korrekt identifiziert?
Heike JÃ¤nz	✔
Silvia von der Weiden	✔
GrÃ¼ne	Keine ‚PER', sondern ‚ORG'
WÃ¤hrend	Kein Eigenname
Schiemann	✔

Tab. 5: Identifizierte Personen bei fehlerhafter Entschlüsselung der Umlaute (Quelle: Eigene Darstellung aus Auswertung in Anhang [H])

Weltweit existieren mehrere Standards, um Zeichensätze darzustellen, daher muss bei der Verarbeitung von Textdaten darauf geachtet werden, welches *encoding* sich für den zu analysierenden Text eignet (vgl. Niekler 2016: 39). Für westeuropäische Sprachen ist die Entschlüsselung per *UTF-8* gängig, alternativ existieren weitere Zeichensatzformate wie *US-ASCII*, *ANSI* oder *Latin-1*, auch als *ISO 8859* bekannt (vgl. ebd.).

Nach der Wahl eines anderen *encoding*-Zeichenformats (‚*Latin-1*') beim Einlesen der Datei und der entsprechenden Anpassung der *Pipeline*, entfallen die Fehlidentifikationen und die gesamte Klassifikationsleistung des gewählten *NER*-Verfahrens verbessert sich. Dies wird im Anhang nochmals anhand einer Gegenüberstellung der am häufigsten identifizierten Personen verdeutlicht (s. Anhang [H]). Aufbauend auf dieser Erkenntnis wird bei allen nachfolgenden Analysen stets darauf geachtet, dass die gewählte Textentschlüsselung beim Einlesen korrekt funktioniert.

Erst nach diesen Schritten wird der eingelesene Text in einzelne Artikel zerlegt. Dabei ist unwesentlich, ob alle Artikel im Vorfeld automatisiert getrennt, separat abgespeichert und nacheinander verarbeitet werden oder, ob eine einzelne Gesamtdatei genutzt wird und diese in der *Pipeline* zerteilt und gespeichert wird. Beide Verfahrensmöglichkeiten werden getestet und funktionieren einwandfrei. Zur Ersparnis von Arbeitsschritten wird sich für die letztgenannte Vorgehensart entschieden.

Um am Ende die automatisiert erhaltenen Ergebnisse mit den manuellen Codierungen auf Artikelebene vergleichen zu können, wird in der *Pipeline* die Überschrift aller Artikel extrahiert. Gespeichert wird alles in tabellarischer Form als sogenanntes *dataframe.* Dies beschreibt eine zweidimensionale Datenstruktur, die idealerweise so aufgebaut ist, dass die einzelnen Beobachtungen eines Datensatzes als Zeilen und die dazugehörigen Variablen in den Spalten erfasst werden (vgl. Wu 2020: 10).

In diesem Fall wird jeder Nachrichtenartikel in einer Zeile abgetragen und alle benötigten Informationen, wie das Medium, die Artikelüberschrift und der Textkörper extrahiert und in den

jeweiligen Spalten festgehalten (s. Abb. 18). Über reguläre Ausdrücke können außerdem ungewollte Umbrüche oder Sonderzeichen in den Spalten des *dataframes* selektiert und entfernt werden, sodass die unten abgebildete Übersicht erhalten wird.

```
text_frame.head()
```

	content	content_clean	source	title	body
0	Di, 07.04.2020, 20:08\n \nedi0431 4 wi 225 cccc...	Di, 07.04.2020, 20:08\nedi0431 4 wi 225 cccce ...	dpa	US-Senat will Konjunkturprogramm um 250 Millia...	Washington (dpa) - Das riesige US-Konjunkturpa...
1	\n\n\nFr, 10.04.2020, 20:28\n \nwap0365 3 pl 15...	Fr, 10.04.2020, 20:28\nwap0365 3 pl 150 lby 14...	dpa	Weiterer Eilantrag gegen bayerische Corona-Bes...	Karlsruhe (dpa) - Das Bundesverfassungsgericht...
2	\n\n\n, 05.04.2020, 10:28\n \nbid0110 4 wi 535 ...	, 05.04.2020, 10:28\nbid0110 4 wi 535 dpa 0405...	dpa	Versicherer kommen Kunden in Corona-Krise entg...	Gegen eine Pandemie sind nur sehr wenige Unter...
3	\n\n\nedi0111 4 pl 290 cccce dpa-euro 0608\n\n...	edi0111 4 pl 290 cccce dpa-euro 0608\nerd0112 ...	dpa	Iran will Corona-Vorschriften leicht lockern	Teheran (dpa) - Der Iran will nach Angaben von...
4	\n\n\nedi0231 4 pl 143 cccce dpa-euro 1144\n\n...	edi0231 4 pl 143 cccce dpa-euro 1144\nerd0232 ...	dpa	Coronavirus auf Jet-Set-Insel Mykonos - Ausgan...	Athen (dpa) - Nachdem bei zwei Menschen auf My...

Abb. 18: Dataframe mit Artikeln pro Zeile und jeweiligen Variablen pro Spalte (Quelle: Screenshot aus dem Code Output der Processing Pipeline)

Einige Nachrichtenartikel weisen in der Überschrift eine andere Formatierung auf, die eine saubere Extraktion beeinträchtigt (s. Anhang [8], S. 103). Dies muss nachträglich manuell angepasst werden, da korrekte Artikelüberschriften für die spätere Auswertung der *NER*-Analyse mit einem Ergebnisabgleich auf Artikelebene essentiell sind.

Des Weiteren ist vor der *NER*-Analyse noch eine Unterscheidung von erwünschten und unerwünschten Inhalten innerhalb der Textdaten notwendig. Damit gemeint ist der Ausschluss von artikelübergreifenden, strukturellen oder textuellen Merkmalen, wie Logos, Grafiken und Werbeanzeigen sowie das Erscheinungsdatum, Fußzeilen, Seitenzahlen oder redaktionelle Hinweise. Ein menschlicher Codierer kann den relevanten Textkörper leicht selbständig erkennen, bei der Nutzung eines automatisierten Verfahrens muss dieser Textbereich hingegen explizit definiert werden. Dadurch soll vermieden werden, dass die Analyse der Daten durch zahlreiche überflüssige Informationen getrübt wird und unter Umständen die Validität der Ergebnisse darunter leidet (vgl. Günther/Scharkow 2014:112).
Die Definition dieses ‚lesbaren Bereichs' der Texte ist bei den Artikeln der *dpa*, *WELT* und des *SPIEGEL* gut möglich, da sie Kennzeichnungen in Form von Metadaten enthalten. Diese Formatierung ist den Datenbanken zu verdanken, aus denen sie bezogen wurden und ist in Abbildung 19 in dem Beispielartikel auf der linken Seite sichtbar. Die Begriffe ‚Body' und ‚Load-Date' ermöglichen eine einheitliche Eingrenzung des Textkörpers nahezu aller Artikel. Sie werden in der *Pipeline* als Schlüsselwörter benutzt, damit nur dieser Abschnitt für die *NER*-Analyse extrahiert und ausgelesen wird.

Unwort der Woche: Coronavirus

Die Welt

Samstag 29. Februar 2020

Copyright 2020 Axel Springer AG Alle Rechte Vorbehalten

DIE WELT

Section: LITERARISCHE-WELT; PROLEGOMENA; S. 26; Ausg. 51

Length: 166 words

Body

Auch die Literatur ist Corona-infiziert, und das nicht nur, weil die Leipziger Buchmesse vielleicht abgesagt werden könnte. Comicfans haben in diesen Tagen ein besonderes Déjà-vu vor Augen: den römischen Bösewicht aus dem vorletzten Asterix-Abenteuer. Im französischen Original von Heft Nummer 37 heißt der Asterix-Gegenspieler mit der fies lächelnden Maske "Coronavirus", und sein Beifahrer hört auf den Namen "Bacillus". Blitzschnell tourt Coronavirus durch Italien und wird durch Sprechchöre gefeiert: "Coronavirus! Coronavirus!". Französische Zeitungen haben den Umstand jetzt anekdotisch aufgegriffen. Der deutsche Asterix-Übersetzer Klaus Jöken, der die beiden Figuren auf Deutsch Caligarius und Bleifuß getauft hatte (vgl. WELT vom 19.10.2017), ist heilfroh: "Asterix in Italien' kam erst neulich als Hörspiel heraus, und gerade im heutigen Kontext möchte bestimmt keiner hören, wie die Norditaliener einen Coronavirus bejubeln", erklärt Jöken auf unsere Nachfrage.

Load-Date: February 29, 2020

End of Document

Finanzen Kompakt: Aktienmarkt: Störung legt Xetra-Handel lahm ++ G7-Staaten: Schuldenstundung für arme Länder ++ Deutsche Bank: Große Andrang auf KfW-Schnellkredite ++ Pharmakonzern: Johnson & Johnson senkt Ausblick

Die Welt

Mittwoch 15. April 2020

Copyright 2020 Axel Springer AG Alle Rechte Vorbehalten

DIE WELT

Section: FINANZEN; KOMPAKT; S. 15; Ausg. 88

Length: 409 words

Body

Aktienmarkt

Störung legt Xetra-Handel lahm

Eine technische Störung hat am Dienstag einen Großteil des Handels an der Deutschen Börse für mehr als vier Stunden lahmgelegt. Nachdem das elektronische Handelssystem Xetra T7 um 09.25 Uhr ausgefallen war, konnte der Handel erst ab 13.45 Uhr wieder aufgenommen werden, wie der Betreiber der Frankfurter Börse mitteilte. Über das System wird der Großteil des deutschen Aktienhandels und der Derivatehandel der Börsen-Tochter Eurex abgewickelt. Es war eine der längsten Störungen in der Geschichte der Frankfurter Börse. Ein Fehler in der internen Kommunikation des Handelssystems habe das Problem ausgelöst. Der Parketthandel in Frankfurt war nicht betroffen.

G7-Staaten

Schuldenstundung für arme Länder

Die führenden Industriestaaten wollen die ärmsten Länder der Welt mit Schuldenerleichterungen unterstützen, damit diese die wirtschaftlichen Folgen der Coronavirus-Pandemie besser abfedern können. Die Staaten der G7-Gruppe seien unter Bedingungen bereit, den Schuldendienst für die ärmeren Länder zeitlich begrenzt auszusetzen, hieß es am Dienstag in einer Mitteilung des US-Finanzministeriums nach einem Telefonat der G7-Minister und Zentralbankchefs. Voraussetzung sei die Beteiligung aller G20-Staaten, hieß es. Damit müsste unter anderem auch China zustimmen, das Kredite an viele Entwicklungsländer vergeben hat.

Deutsche Bank

Große Andrang auf KfW-Schnellkredite

Die Deutsche Bank erwartet einen Ansturm kleinerer und mittlerer Firmen auf den KfW-Schnellkredit in der Corona-Krise. "Wir rechnen ab morgen mit einer großen Nachfrage", sagte Stefan Bender, Leiter Unternehmensbank

Abb. 19: Erschwerte Textkörperbestimmung je nach Artikelart
(Quelle: WELT-Artikel aus dem Corona-Datensatz)

Auf der rechten Seite der Abbildung 19 ist allerdings erkennbar, dass bei Beiträgen bestimmter Rubriken der relevante Textbeitrag in anderen Textdaten eingebettet ist. In solchen Fällen würden zusätzlich die Eigennamen aus irrelevanten Textteilen identifiziert werden, da nicht übergreifend für all diese Artikel standardisiert festgelegt werden kann, welcher Textabschnitt von Interesse ist. Dies würde zwar die Identifikationsleistung des Verfahrens nicht negativ beeinträchtigen oder erheblich mehr Zeitaufwand benötigen, doch die Auswertung der erhaltenen Ergebnisse und der Vergleich mit den manuellen Daten wäre durch die überflüssigen Daten erschwert.

Auch hier ist die manuelle Bereinigung der Texte eine Option und wird als *Unitizing* bezeichnet: „Bei Inhaltsanalysen, in denen Auswahl- und Analyseeinheit nicht identisch sind, muss vor der eigentlichen Codierarbeit zunächst das Untersuchungsmaterial zerlegt werden“ (Scharkow 2013: 294). Dieser zusätzliche Arbeitsschritt der Eliminierung nicht benötigter Textteile muss im ‚Corona‘-Datensatz nur bei sieben Artikeln händisch durchgeführt werden (s. Anhang [8], S. 103). Bei einer größeren Stichprobe mit mehr betroffenen Texten stünde dieser Arbeitsaufwand jedoch nicht im Verhältnis zu den Vorzügen der gewünschten Arbeitserleichterung der automatisierten Methode.

Ein Beispiel hierfür sind erneut die Texte der *SZ*, worin ein weiterer Grund gegen die Einbindung dieser Artikel in die Analyse besteht. Die *SZ*-Texte weisen keinerlei Metadaten auf, dafür jedoch zahlreiche störende, uneinheitliche Elemente, die bei über 400 Artikeln angepasst werden müssten (s. Anhang [9], S. 104).

Die Artikel des *SPIEGEL* dagegen beinhalteten die gleichen Metadaten wie die der *WELT* in Abbildung 17, da sie ebenfalls über die Datenbank *LexisNexis* erhalten wurden. Die Beiträge der *dpa* nutzen ähnliche Meta-Kennzeichnungen mit anderen Benennungen, die für diesen Zweck verwendet und in der *Pipeline* zusätzlich hinterlegt werden.
Festzuhalten ist, dass bei der Nutzung unterschiedlicher Datenquellen, das Textlayout der Medienstichprobe bekannt sein muss, um eine sinnvolle Verarbeitung der Texte zu gewährleisten und die Eingrenzung des zu untersuchenden Textkörpers in der *Pipeline* korrekt definieren zu können. Wenn Nachrichtenartikel von unbekannten, vielen verschiedenen oder schlechtstrukturierten Quellen stammen, werden die algorithmischen Bereinigungsverfahren meist umständlicher und komplexer (vgl. Günther/Scharkow 2014: 114).

Mehrfach fällt erst bei der Sichtung der Ergebnisse der Probedurchläufe auf, dass eine nachträgliche Anpassung der *Pipeline* notwendig ist, da beispielsweise mit der vorgenommenen Texteingrenzung, weiterhin die Redakteursnamen und Bildquellen in die *NER*-Analyse einfließen und die ermittelten Eigennamen verzerren.
Außerdem stellt sich bei der Prüfung der ersten Ausgabedaten heraus, dass Wörter, die in den Nachrichtentexten komplett in Versalien verfasst sind, oft fälschlicherweise als Eigennamen identifiziert und uneinheitlich klassifiziert wurden. Abbildung 20 liefert ein Beispiel hierfür und stellt dar, wie Begriffe wie ‚OLYMPIA' oder ‚BIATHLON' als Eigenname markiert und unterschiedlichen *NE*-Klassen zugehörig gekennzeichnet werden.

Abb. 20: Wörter in Großbuchstaben beeinflussen NER-Identifikationsleistung
(Quelle: Screenshot des Outputs aus eigenem SpaCy Code - large model)

Eine zusätzliche Abbildung im Anhang ([10], S.104) zeigt, dass in den *dpa*-Artikeln häufig gesamte Sätze in Versalien vorkommen, da dies als Stilmittel für Zwischenüberschriften eingesetzt wird. Diese Textteile führen oft zu Fehleinschätzungen von Eigennamen der *NER*-Verfahren. Die gezielte Bereinigung dieser Sätze erweist sich als händische Aufgabe als sehr umfangreich. Alternativ wird daher im *Preprocessing* die Umwandlung des gesamten Datensatzes in Kleinbuchstaben getestet. Dies führt tatsächlich dazu, dass die ehemals in Versalien geschriebenen, fälschlicherweise identifizierten Wörter nicht mehr in den Ergebnissen vorkommen.

Es verursacht jedoch allerlei andere Fehlklassifikationen von Eigennamen. Teilweise fehlen zuvor identifizierte *NEs*, teilweise werden sie mit anderen Kategorien gekennzeichnet (s. Anhang [J]).
Dies belegt die in Kapitel 3.1 beschriebene Tatsache, dass *NER*-Verfahren die morphologischen Eigenschaften von Wörtern für ihre Klassifikation als Eigennamen miteinbeziehen. Aufgrund dessen wird für die schlussendliche Analyse von solch einer Textumwandlung abgesehen und das Vorhandensein von großgeschriebenen Begriffen in den Ergebnissen geduldet und im Nachhinein bereinigt.
An dieser Stelle besteht jedoch Potenzial für weitere Untersuchungen mit mehr *code literacy*, um zu vergleichen wie sich der Einsatz einer *regular expression* auswirken würde, die artikelübergreifend nur solche Wörter, die vollends aus Versalien bestehen, in Kleinbuchstaben umwandelt. Dies könnte dazu führen, dass weniger Wörter falsch als Eigennamen identifiziert werden, es könnte aber auch darin resultieren, dass andere *Named Entities*, wie die Namen von Organisationen und Parteien (WHO, RKI, SPD) schlechter erkannt werden.

Die anderen *Preprocessing*-Schritte, die in Kapitel 3.2 beschrieben wurden, wie die Umwandlung der Wörter in ihre Wortstämme (*lemmatizing)* oder die Beseitigung von Stoppwörtern, sind für die *NER*-Analyse nicht notwendig und zielführend, da die Erkennung von Eigennamen abhängig von einem intakten Satzbau und Kontext ist. Außerdem können Stoppwörter Teile von Eigennamen darstellen, wie beispielsweise ‚Die Linke' oder ‚Zentralinstitut für die kassenärztliche Versorgung', weshalb ihre Beseitigung die Eigennamenidentifikation negativ beeinträchtigen könnte.

Nach erfolgreicher Vorbereitung der Daten und dem Einlesen der Textdatei, kann daher unmittelbar das *Tokenizing* der Textdaten erfolgen. Das Segmentieren der Sätze und Zerteilen in Token läuft bei der Bibliothek *spaCy* als interner Prozess ab, bei *Stanza* muss ein entsprechender Prozessor dafür geladen werden (s. Anhang [11], S. 105), während bei *FLAIR* die Trennung der Sätze (*sentence splitting*) als separater Verarbeitungsschritt notwendig ist. Daraufhin kann die schlussendliche *NER*-Analyse stattfinden. Dafür werden im Code der *Pipeline* die entsprechenden deutschen Modelle der drei ausgewählten Bibliotheken geladen und auf die Spalte mit dem bereinigten Text pro Artikel angewandt. Die dort identifizierten Eigennamen können danach in separaten Spalten des *dataframes* gespeichert werden (s. Anhang [12], S. 105).
Am Ende ist nur noch erforderlich die erzeugte Datenstruktur in csv-Format zu exportieren, um sie über Excel sichten, zusammenfassen und auswerten zu können (s. Anhang [D]).

Für diese Auswertung und Gegenüberstellung der Ergebnisse muss die Output-Datei einen gewissen Aufbau vorweisen. Die identifizierten Eigennamen sollten pro Artikel getrennt und mit Angabe der *NE*-Klasse ausgegeben werden. Dies ist zum einen notwendig, um die *NER*-Leistung getrennt nach Klasse zu untersuchen, aber auch um eine Aggregation der Eigennamen auf Artikelebene durchführen zu können.
Doppelungen der Akteure innerhalb eines Artikels werden zusammengefasst, um ein einmaliges Auftreten einer Person oder Organisation pro Artikel zu erfassen. Bei der Ermittlung der Häufigkeiten kann dadurch sichergestellt werden, dass es sich um die Anzahl der Artikel handelt, die diesen Akteur nennen und nicht um die absolute Häufigkeit des Namens im Datensatz. Ansonsten würden sich vor allem bei Artikeln im Interview-Format Verzerrungen ergeben, da diese Texte den Namen des interviewten Gesprächspartners bei jeder Antwort voranstellen. Die absolute Anzahl der Namen würde daher nicht die übergreifende Häufigkeit der Akteursnennung aller Artikel widerspiegeln und könnte zu einer falschen Ergebnisinterpretation führen.

Für die Nachbearbeitung und Aufbereitung der erhaltenen Daten wird maßgeblich mit Excel gearbeitet und das Excel Tool *Power Query* genutzt. Im Anhang (Abb. [13], S. 105) ist abgebildet, wie die ausgegebenen *NER*-Ergebnisse umgewandelt werden, um die gewünschten Auswertungen durchführen zu können.
Zusätzlich werden fehlerhafte Daten und Symbole entfernt, die völlig irrelevante Inhalte wiedergeben. Ebenso wie falsch ermittelte *Chunks*, bei denen eindeutig erkennbar Teile fehlen. Dies ist identifizierbar an Bindestrichen, denen kein Inhalt folgt oder vorab fehlt, wie zum Beispiel ‚-Institut' oder ‚Frank-'. Diese Fehlleistungen wurden separat gebündelt, um zu evaluieren, welche Bibliothek maßgeblich für die irrelevanten Ergebnisse zuständig ist. Des Weiteren erfolgen die Sichtung, Markierung und Sammlung von Ergebnissen, bei denen es sich nicht um Eigennamen handelt. Dies wird in Kapitel 6 detailliert pro Bibliothek wiedergegeben.
Als letztes wird eine Vereinheitlichung der Lang- und Kurzschreibweisen der Eigennamen durchgeführt (Robert-Koch-Institut = RKI, Bündnis 90/Grüne = Die Grünen). Ebenso werden die flektierten Eigennamen (Merkels, den Grünen, des Zentrums für Virologie) und die bloße Nachnamennennung vereinheitlicht, sodass die Wortformen nur in einer einzigen Verwendungsform im Datensatz vorkommen und zusammengefasst werden können. Mehrere Personennennungen weisen dabei den gleichen Nachnamen auf (‚Hess'/‚Marx'/‚Müller'). In diesen Fällen kann durch den Rückbezug zu dem jeweiligen Nachrichtenartikel die erste vollständige Nennung des Namens nachvollzogen und nachträglich angepasst werden. Ohne Angabe der Artikelzugehörigkeit des identifizierten Eigennamens wäre dies nicht möglich und würde zu Unklarheiten bei der weiteren Auswertung führen.

Insgesamt manifestiert sich die in der Literatur beschriebene Tatsache, dass die Datenaufbereitung als unerlässlicher Teil einer *NLP*-Analyse meist mehr Zeit in Anspruch nimmt als die Auswahl und Implementierung der automatisierten Verfahren selbst (vgl. Gilch/Schüler 2019: 36). Ebenso bestätigt sich der Fakt, dass meist manuelle Teilschritte notwendig sind, welche, wenn sie nicht dokumentiert werden, zu einem Verlust der Reliabilität und Transparenz des Verfahrens führen (vgl. Scharkow 2013: 296).

Es wird sichtbar, dass die Effektivität von automatisierten Verfahren und die letztliche Größe der bearbeitbaren Stichprobe stark davon abhängen, wie zuverlässig sich die Datenerhebung und -bereinigung automatisieren lassen (vgl. ebd.). Liegen die zu analysierenden Daten nicht in digitalisierter Form vor oder können nicht fehlerfrei transformiert werden, stellt eine automatisierte Inhaltsanalyse nicht die geeignete Methode dar.

Es kann auch bestätigt werden, dass die Vorbereitung für die Datenerfassung viel Zeitaufwand erfordert, sodass automatisierte Verfahren nur dann effizienter gegenüber manuellen Methoden sind, wenn die letztlich zu analysierende Stichprobe sehr umfangreich ist (vgl. Graaf/van der Vossen 2013: 440).

Außerdem wird deutlich, dass die Vorverarbeitung der Datensätze maßgeblich den Umfang und die Qualität der resultierenden Ergebnisse der automatisierten Inhaltsanalyse determiniert (vgl. Maier et al. 2018: 106). Die erhaltenen Ergebnisse der *NER*-Analysen werden hierauf gebündelt vorgestellt.

6. Vergleich der Verfahren und Erhebungsergebnisse

Um die Güte der getesteten *NER*-Verfahren bewerten zu können, wird in der vorliegenden Arbeit ein umfassender Vergleich ihrer Anwendungsspezifika und ausgegebenen Ergebnisse durchgeführt. In den Geistes- und Sozialwissenschaften wird die vergleichende Analyse als eine grundlegende wissenschaftliche Tätigkeit angesehen, wenn „der Vergleich nicht unreflektiert durchgeführt wird, sondern ein logisches Vorgehen zugrunde liegt“ (Nordbeck 2013: 110). Hierfür werden in dem ersten Teil des Kapitels anhand von festgelegten Kategorien die Erkennungs- und Klassifikationsleistung der drei untersuchten *NER*-Verfahren detailliert dargelegt und im Anschluss den manuellen Codierungen gegenübergestellt. Alle Befunde werden am Ende des Kapitels (6.4) komprimiert wiedergegeben und tabellarisch zusammengefasst.

6.1 Angewandte Methodik

Zunächst wird ein fallorientierter Vergleich durchgeführt, welcher mehrere Merkmalsdimensionen an wenigen Fällen untersucht (vgl. Klimek/Müller 2015: 60). Diese Fälle, hier die drei *NLP*-Bibliotheken, sind bewusst ausgewählt worden (Kap. 4.2) und werden umfassend analysiert. Durch die Anwendung der Methodik des fallorientierten Vergleichs können Ähnlichkeiten und Unterschiede ermittelt, bestimmte Spezifika hervorgehoben (vgl. ebd.: 62) und dadurch Erkenntnisse über die *NER*-Verfahren gewonnen werden.
Diese Methode eignet sich für die vorliegende Untersuchung, da bei den fallorientierten Forschungsstrategien die Gewinnung von Wissen über den jeweiligen Fall im Vordergrund steht (vgl. Nordbeck 2013: 117f.). „Sie sind auf eine reiche Beschreibung ausgelegt und weisen eher eine entdeckende Funktion auf“ (Klimek/Müller 2015: 60). Dadurch werden solche qualitativen Vergleichsstudien der Komplexität einiger Einzelfälle gerecht.
Dennoch lassen sie aufgrund ihrer beschränkten Fallanzahl kaum Verallgemeinerungen zu (vgl. ebd.). Die hier ermittelten Befunde können deshalb nicht auf alle automatisierten *NER*-Verfahren übertragen werden, geben jedoch einen umfassenden Einblick in die Leistung der Bibliotheken *spaCy*, *Stanza* und *FLAIR*.

Ergänzend werden die umfangreichen Ergebnisse, die nach der Anwendung der drei *NER*-Verfahren erhalten wurden, quantitativ ausgewertet und die in Kapitel 4.2 vorgestellten Leistungskennzahlen *Precision* und *Recall* ermittelt. Um diese Kennzahlen zu berechnen und damit die automatisiert extrahierten Ergebnisse zu beurteilen, werden üblicherweise manuell erhobene Vergleichsdaten als Bewertungsmaßstab für ‚richtige‘ oder ‚falsche‘ Ergebnisse genutzt (vgl. Schwotzer 2014: 55; Nunez-Mir et al. 2016: 126).

Dementsprechend werden bei einer Auswertung die automatisiert extrahierten Ergebnisse, die manuell nicht erhoben wurden, als *False Positives* betitelt (vgl. Pinto et al. 2016: 10). Doch da die für diese Untersuchung verfügbaren, manuell selektierten Akteure nicht alle Personen- und Organisationsnamen des Datensatzes abdecken, weil nur diejenigen Akteure erfasst wurden, die konkrete Aussagen tätigen, wird die Definition der *False Positives* hier weiter gefasst. Es werden alle extrahierten Eigennamen eigenständig geprüft und dabei bewertet, ob es sich bei dem ausgegebenen Ergebnis tatsächlich um einen Eigennamen handelt. Auf diesem Weg erfolgt die Berechnung der *Precision* der Verfahren, oft auch als ‚Exaktheit' oder ‚Verlässlichkeit' bezeichnet (vgl. Rössler 2007: 92; vgl. Ketschik et al. 2020: 204).

Um die Vollständigkeit (*Recall*) der Ergebnisse zu evaluieren, wird im Anschluss untersucht wie viele der manuell erfassten individuellen und institutionellen Akteure auch maschinell identifiziert wurden. Bei diesem Vergleich werden die korrekt identifizierten Eigennamen als *True Positives* bezeichnet (vgl. ebd.).

Die Matrix in Abbildung 21 illustriert die Begriffsgruppen, in welche die erhaltenen *NER*-Ergebnisse eingeteilt werden können.

		Akteur manuell extrahiert?	
		JA	**NEIN**
Akteur automatisiert extrahiert?	**JA**	***True Positives*** Extrahiert (manuell identifizierte Akteure, die auch automatisiert erkannt wurden)	***False Positives*** Extrahiert, aber kein Akteur (automatisiert extrahierte Ergebnisse, bei denen es sich nicht um Eigennamen handelt)
	NEIN	***False Negatives*** Nicht extrahiert (manuell selektierter Akteure wurden nicht von *NER*-Verfahren erkannt)	***True Negatives*** Nicht extrahiert und nicht relevant (korrekt ignorierte Wörter)

Abb. 21: An die Untersuchung angepasste Wahrheitsmatrix
(Quelle: Eigene Darstellung in Anlehnung an Schulte 2019: o. S. - http://bit.ly/3byRXXm)

Bei sogenannten *False Negatives* handelt es sich um Akteure im Datensatz, die manuell erhoben wurden, aber nicht von dem *NER*-Verfahren erkannt wurden. Diese *False Negatives*, ebenso wie die zuvor erläuterten *True* und *False Positives*, werden nachfolgend für alle drei *NER*-Verfahren ermittelt und untersucht.

Die in der Abbildung aufgeführten *True Negatives*, welche all jene Wörter im Datensatz darstellten die keine Eigennamen sind und nicht von den *NER*-Verfahren extrahiert wurden, werden in der Untersuchung außer Acht gelassen, da sie für die Bewertung der Qualität der Verfahren nicht von Bedeutung sind.

Nach der umfassenden Analyse der erhaltenen Ergebnisse soll beurteilt werden, ob die getesteten Verfahren grundsätzlich für die automatisierte Erkennung von Akteuren in Texten eingesetzt werden können und ob sich eine der untersuchten Bibliotheken besser eignet als die anderen.

6.2 Gegenüberstellung der angewandten Verfahren

Für die Gegenüberstellung der drei *NER*-Verfahren werden vier verschiedene Vergleichsvariablen betrachtet. Es werden die Verarbeitungsgeschwindigkeit der Verfahren sowie die absolute Anzahl und Übereinstimmung der erhaltenen Ergebnisse untersucht. Daraufhin wird überprüft, ob es sich bei den erhaltenen Ergebnissen tatsächlich um *Named Entities* handelt. Um die korrekte Erkennung von Eigennamen zu beurteilen, wird unter anderem ermittelt, ob die Identifikation der gesamten Namenssequenzen richtig erfolgt und die vorgenommene *NE*-Klassifikation korrekt ist. Darauf aufbauend können die jeweiligen Fehlerquoten der Verfahren berechnet und verglichen werden.

6.2.1 Verarbeitungsgeschwindigkeit

Alle drei *NER*-Analysen werden an demselben Endgerät mit dem zentralen Prozessor (*CPU)* durchgeführt, um eine Vergleichbarkeit der Verarbeitungsgeschwindigkeit zu gewährleisten (s. Anhang [30], S. 115). Auch wenn parallel keine anderen Programme während der drei *NER*-Analysen geöffnet sind, ist ungewiss, welche Hintergrundprozesse an dem Computer ablaufen und Einfluss auf die Verarbeitungsleistung nehmen. Die Verarbeitungsdauer kann daher an anderen Geräten variieren oder bei der Möglichkeit der Nutzung eines Graphikprozessor (*GPU)* gezielt verkürzt werden (vgl. Vychegzhanin/Kotelnikov 2019: 76). Um die deutlichen Unterschiede in der Verarbeitungszeit zwischen den drei Bibliotheken aufzuzeigen, eignet sich die gewählte Durchführung auf dem *CPU* jedoch gut.

Der Prozess des Identifizierens und Klassifizierens der Eigennamen in dem Datensatz mit n = 887 Corona-Nachrichtenartikeln dauert je nach genutzter *NER*-Bibliothek unterschiedlich lange. Während *spaCy* für einen Durchlauf des gesamten Datensatzes nur 3 Minuten in Anspruch nimmt, benötigt *Stanza* mit dem *CoNLL03*-Modell 44 Minuten und mit dem *GermEval2014*-Modell 106 Minuten. Verglichen dazu braucht die Bibliothek *FLAIR* mit mehr als 2,5 Stunden am längsten, um den Datensatz zu verarbeiten.

Auffällig ist hierbei, dass die benötigte Verarbeitungszeit konträr zu dem Umfang der ausgegebenen Ergebnisse ist. Das *NER*-Verfahren von *FLAIR* mit der längsten Verarbeitungsdauer liefert die geringste Anzahl an Ergebnissen, während *spaCy* mit der kürzesten Dauer die größte Menge an *NEs* markiert.

Dies liefert ein Indiz für die Komplexität der Algorithmen hinter den Verfahren. Wie in Kapitel 3 erwähnt, benötigt der Einsatz von komplexeren *NLP*-Algorithmen höhere Rechenleistungen und somit längere Verarbeitungszeiten. Dies scheint bei *Stanza* und *FLAIR* der Fall zu sein, daher gilt es zu prüfen, ob diese Bibliotheken im Gegenzug weniger Fehler in ihren Ergebnissen aufweisen als das schnelle *NER*-Verfahren der Bibliothek *spaCy*.

6.2.2 Umfang der erhaltenen Ergebnisse nach *NE*-Klasse

Wie bereits erwähnt, fällt der Umfang der identifizierten *Named Entities* je nach genutzter Bibliothek recht unterschiedlich aus. Die *NER*-Analyse mit *spaCy* gibt bei dem analysierten Datensatz über 25.000 Eigennamen aus, während *Stanza* knapp 22.000 *NEs* erkennt und *FLAIR* ungefähr 18.000 Wörter als Eigennamen kennzeichnet. Dabei handelt es sich um die ungeprüften und unbereinigten Ergebnisse.

Abbildung 22 visualisiert die absolute Anzahl der identifizierten Eigennamen pro *NE*-Klasse vor der Ergebnisbereinigung. Erkennbar ist eine sehr ähnliche Identifikationsleistung bei den Personen, während in den Klassen ORG, LOC und MISC eine deutliche Diskrepanz im Umfang der identifizierten Eigennamen sichtbar ist.

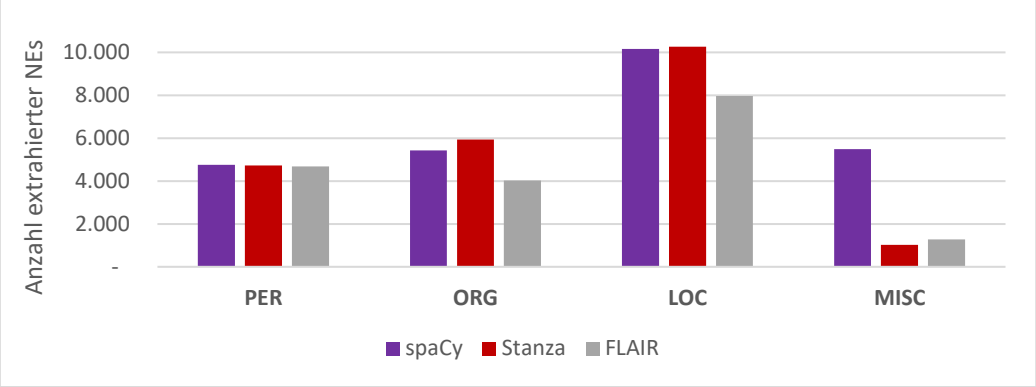


Abb. 22: Unbereinigte absolute Anzahl der NEs pro Klasse im Vergleich (Quelle: Eigene Darstellung aus Anhang [F])

Mittels der in Kapitel 5.2 beschriebenen, manuellen Nachbearbeitung der Ergebnisse werden die Personen-, Organisations- und Ortsnamen vereinheitlicht und Mehrfachnennungen auf Artikelebene zusammengefasst.

Tabelle 6 zeigt, wie sich die Ergebniszusammenfassung auf die Anzahl der erhaltenen Eigennamen auswirkt. Die Gesamtmenge wird um 34% reduziert, während die Mengenunterschiede zwischen den Bibliotheken dabei allerdings nahezu unverändert bleiben (s. Zeile ‚total').

Zusätzlich wird abgebildet, wie hoch die Anteile der Eigennamen pro *NE*-Klasse ausfallen. Diese Anteile decken sich mit den Angaben aus der obigen Abbildung 2 und verschieben sich nach der Zusammenfassung der Eigennamen nur minimal.

Gegenüberstellung der absoluten Anzahl erhaltener NEs												
Unbereinigte Ergebnisse							Ergebnisse nach NE-Zusammenfassung					
	spaCy		Stanza		FLAIR			spaCy		Stanza		FLAIR	
PER	4.759	18,4%	4.722	21,5%	4.683	26,1%	PER	2.833	15,5%	2.480	18,0%	2.503	22,3%
ORG	5.433	21,0%	5.935	27,0%	4.018	22,4%	ORG	4.240	23,2%	3.990	29,0%	2.968	26,5%
LOC	10.162	39,3%	10.263	46,8%	7.959	44,4%	LOC	6.882	37,7%	6.611	48,1%	4.824	43,0%
MISC	5.481	21,2%	1.025	4,7%	1.274	7,1%	MISC	4.319	23,6%	667	4,9%	915	8,2%
Total	**25.835**		**21.945**		**17.934**		Total	**18.274**		**13.748**		**11.210**	
						65.714							43.232
Dauer:	3 min		106 min		148min								**– 34,2%**

Tab. 6: Absolute Anzahl erhaltener Eigennamen vor und nach der Bereinigung (Quelle: Eigene Darstellung aus Anhang [F])

In einigen kommunikationswissenschaftlichen Publikationen, in denen *NER*-Verfahren angewandt werden, sind bei der Zählung der identifizierten Akteure auch der Mittelwert (*M*) pro Nachrichtenartikel und die Standardabweichung (*SD*) angegeben (vgl. Trilling et al. 2017: 48; vgl. Burggraaff/Trilling 2020: 120). Für die drei getesteten Verfahren werden diese Werte ermittelt und tabellarisch zusammengefasst (s. Anhang [14], S. 106). *FLAIR* weist dabei den geringsten Mittelwert mit M = 5,7 auf. Das *NER*-Verfahren dieser Bibliothek extrahiert durchschnittlich einen Akteur weniger pro Artikel als *Stanza* (M = 6,5) oder *spaCy* (M = 6,6).

Aufgeführt ist außerdem, dass bei der Stichprobe von n = 887 Artikeln in einigen Nachrichtentexten keine und in anderen bis zu 52 Personen- und Organisationsnamen extrahiert werden. Dies hängt selbstverständlich von der Länge und Art der untersuchten Artikel ab. Die *dpa*-Artikel sind meist kürzer (Ø-Länge = 278 Wörter) und weisen grundsätzlich weniger Akteure auf (M = 5,1). In den längeren *WELT*-Artikeln (Ø-Länge = 833 Wörter) werden durchschnittlich M = 10 Eigennamen der Klasse ‚ORG' und ‚PER' extrahiert. Die *SPIEGEL*-Artikel der Stichprobe weisen die höchste durchschnittliche Wortanzahl auf (Ø-Länge = 1.200 Wörter) und darin werden M = 11 Akteure identifiziert (s. Anhang [E]). Aufgrund der Zusammensetzung der Stichprobe aus 80% *dpa*-Artikeln ergibt sich der Durchschnittswert von knapp 6 extrahierten Eigennamen pro Nachrichtenartikel.

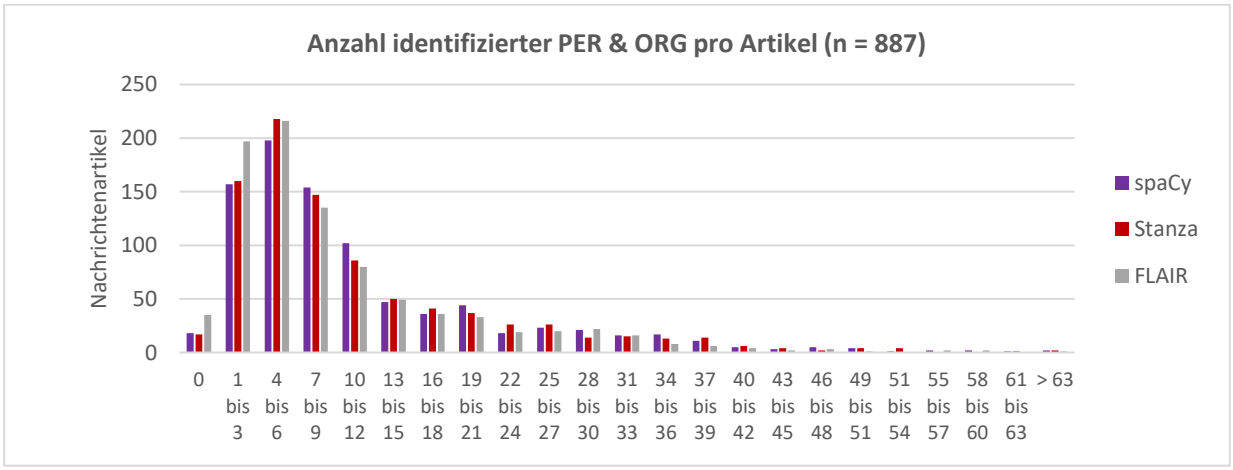


Abb. 23: Histogramm zur Darstellung der extrahierten NEs pro Artikel (Quelle: Eigene Darstellung aus Anhang [E])

Die oben abgebildete Häufigkeitsverteilung hilft klarer zu erkennen, dass in einem Großteil der analysierten Texte zwischen vier und sechs Akteure als ‚PER' oder ‚ORG' klassifiziert werden (s. Anhang [E]). Sichtbar ist auch, dass *FLAIR* in mehr Artikeln eine kleinere Anzahl an Eigennamen extrahiert als *Stanza* und *spaCy*.

Auf den ersten Blick kann nicht beurteilt werden, ob das Identifizieren vieler oder weniger Eigennamen ein gutes oder schlechtes Anzeichen für die Leistung des jeweiligen *NER*-Verfahrens ist. Denn die Ausgabe von vielen Ergebnissen kann viele *False Positives* beinhalten. Während wenig identifizierte *NEs* auf eine schlechtere Erkennungsleistung oder aber auf ein präziseres Verfahren hinweisen können.
Daher ist es wichtig auszuwerten, wie hoch die Fehlerraten der Verfahren sind und zu untersuchen, ob die identifizierten Eigennamen der Verfahren übereinstimmen oder sich stark unterscheiden.

6.2.3 Übereinstimmung der identifizierten Akteure nach *NE*-Klasse

Um beurteilen zu können, inwieweit die erhaltenen Ergebnisse der drei Verfahren sich decken oder voneinander divergieren, bietet sich eine Untersuchung der am häufigsten extrahierten Eigennamen an. Dadurch kann relativ rasch ein Überblick und erster Vergleich der Ergebnisse pro *NE*-Klasse erfolgen.
Eine Auswertung der meistgenannten Akteure kann bei jeder der drei Bibliotheken mittels weniger Code-Zeilen bereits innerhalb der *NLP-Pipeline* erhalten werden (s. Anhang [A]).
Abbildung 24 zeigt exemplarisch die Ergebnisse bei der Ausgabe der häufigsten Personen- und Organisationsnamen.

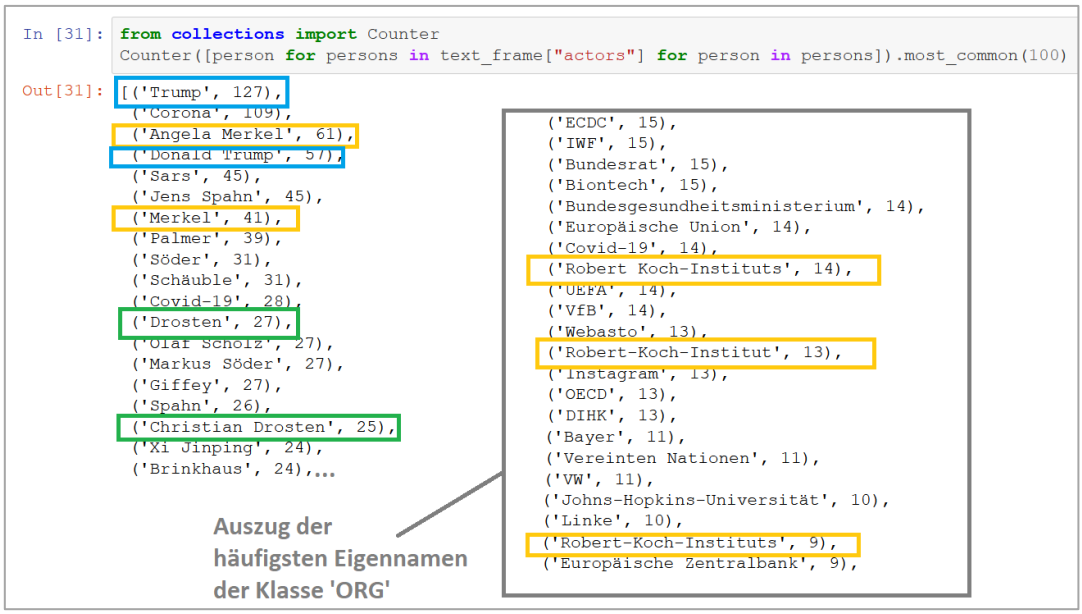


Abb. 24: Auszug der Ausgabe der meistgenannten Akteure im Datensatz (Quelle: Screenshot der Stanza Processing Pipeline s. Anhang [A])

Dieser Schritt in der *Pipeline* ermöglicht einen schnellen ersten Eindruck über die erkannten Akteure des vorliegenden Datensatzes. Doch es wird auch deutlich, dass die verschiedenen Schreibweisen der Eigennamen eine saubere Darstellung der aufsummierten Häufigkeiten verhindern (s. Abb. 24). Außerdem handelt es sich bei den oben erhaltenen Angaben um die absolute Anzahl der identifizierten Namen. Einzelne Interview-Artikel im Datensatz, in denen der Name eines Befragten gehäuft auftaucht (hier ‚Giffey' und ‚Brinkhaus'), verzerren den Gesamteindruck.

Daher ist eine Häufigkeitsermittlung der Akteure nach der manuellen Bereinigung und Zusammenfassung für eine kommunikationswissenschaftliche Akteursanalyse gehaltvoller. Dabei werden die Ergebnisse so komprimiert, dass eine Akteursnennung pro Artikel gezählt wird. Die Häufigkeit der Akteure spiegelt somit die Anzahl an Artikeln wider, in denen sie vorkommen. Die nachfolgende tabellarische Übersicht listet auf, welche 20 Personen und Organisationen auf diesem Wege am häufigsten in dem Datensatz identifiziert werden.

Bei der *NE*-Klasse ‚Personen' ist eine sehr starke Übereinstimmung zwischen den drei Bibliotheken zu erkennen, sowohl bei den Akteursnamen als auch der Anzahl an Artikeln, in denen sie vorkommen.

PER						ORG					
spaCy		stanza		flair		spaCy		stanza		flair	
Donald Trump	57	Donald Trump	57	Donald Trump	57	CDU	123	dpa	159	CDU	126
Angela Merkel	56	Angela Merkel	56	Angela Merkel	56	dpa	98	CDU	144	dpa	95
Jens Spahn	44	Jens Spahn	46	Jens Spahn	45	SPD	68	EU	102	SPD	68
Markus Söder	27	Markus Söder	27	Markus Söder	27	WHO	63	SPD	86	EU	61
Olaf Scholz	26	Olaf Scholz	26	Olaf Scholz	26	EU	61	RKI	62	WHO	57
Christian Drosten	23	Christian Drosten	23	Christian Drosten	23	RKI	52	WHO	59	RKI	53
Peter Altmaier	17	Ursula von der Leyen	19	Peter Altmaier	17	Die Grünen	46	Die Grünen	56	Die Grünen	47
Ursula von der Leyen	15	Peter Altmaier	17	Ursula von der Leyen	17	CSU	36	CSU	46	CSU	37
Winfried Kretschmann	14	Winfried Kretschmann	14	Winfried Kretschmann	14	EU-Kommission	35	Twitter	39	Bundestag	23
Xi Jinping	14	Xi Jinping	14	Xi Jinping	14	Bundestag	31	Bundestag	34	Lufthansa	22
Emmanuel Macron	13	Emmanuel Macron	13	Emmanuel Macron	13	Gesundheitsministerium	30	FDP	28	Berliner Charité	20
Heiko Maas	13	Heiko Maas	13	Heiko Maas	13	Berliner Charité	23	EU-Kommission	26	CDC	18
Horst Seehofer	13	Horst Seehofer	13	Horst Seehofer	13	Lufthansa	22	Berliner Charité	23	FDP	17
Armin Laschet	11	Armin Laschet	11	Armin Laschet	11	CDC	18	Lufthansa	23	EZB	17
Boris Johnson	11	Tedros Ghebreyesus	11	Boris Johnson	11	EZB	17	UN	18	Dax	14
Tedros Ghebreyesus	11	Boris Johnson	10	Tedros Ghebreyesus	11	FDP	17	CDC	18	UN	12
Giuseppe Conte	10	Giuseppe Conte	10	Giuseppe Conte	10	Johns Hopkins Universität	15	EZB	17	VW	12
Franziskus	9	Manne Lucha	10	Manne Lucha	10	VW	13	Johns-Hopkins-Universität	16	IWF	12
Sebastian Kurz	9	Franziskus	9	Franziskus	9	IWF	12	DFL	15	DFL	11
Zhong Nanshan	8	Sebastian Kurz	9	Sebastian Kurz	9	DFL	11	VW	14	Xinhua	10

Tab. 7: 20 häufigste Personen und Organisationen nach Bibliothek
(Quelle: Eigene Darstellung aus Anhang [F])

Bei der Klasse ‚ORG' sind im Vergleich zu der Klasse ‚PER' mehr Abweichungen zwischen den Ergebnissen der Bibliotheken sichtbar. Die erkannten Eigennamen pro Artikel unterscheiden sich stärker, sodass die Rangfolge der vorkommenden Akteure nicht identisch ist. Auffällig ist hier, dass die Ergebnisse von *FLAIR* und *spaCy* sich bei der Anzahl der erkannten Organisationen stärker ähneln, während *Stanza* oft mehr Nennungen erkennt. Bereits in Abbildung 22 war sichtbar, dass *Stanza* mehr Organisationen im Datensatz ermittelt als *spaCy* und *FLAIR*. Zu prüfen ist in den nächsten Unterkapiteln, ob dies an einer besseren Identifikationsleistung dieser Bibliothek oder an mehr falsch ermittelten Eigennamen liegt.

Im Zuge der Untersuchung der meistidentifizierten Organisationen fällt außerdem auf, dass grundsätzlich weder *Stanza* noch *FLAIR* das generische Wort ‚Gesundheitsministerium' als Organisation kennzeichnen. Das *NER*-Verfahren von *FLAIR* extrahiert außerdem die Institutionen ‚Auswärtiges Amt' und ‚EU-Kommission' nicht, während die Verfahren von *spaCy* und *Stanza* diese als Eigennamen identifizieren (s. Anhang [E]). Das andere, eingangs getestete *NER*-Modell von *Stanza* (s. Kap. 4.2.3), welches ebenso wie *FLAIR* mit älteren *CoNLL03*-Texten trainiert wurde, extrahiert die Ausdrücke ‚Auswärtiges Amt' und ‚EU-Kommission' jedoch auch nicht. Dies könnte darauf hinweisen, dass *NER*-Verfahren, die nur mit dem *CoNLL03*-Korpus trainiert wurden, Organisationsnamen, die aus sehr allgemeinen Ausdrücken bestehen, nicht als Eigennamen erkennen.

Die Kategorie ‚LOC' weist beim Vergleich der *NER*-Verfahren eine sehr hohe Übereinstimmung der am häufigsten identifizierten Orten und der Anzahl ihrer Nennungen auf. Die Gegenüberstellung im Anhang zeigt, dass in den erhaltenen Rangfolgen 18 von 20 Eigennamen über alle drei Bibliotheken hinweg gleich sind (s. Anhang [15], S. 106). Bei der Klasse ‚MISC' treten uneinheitliche Ergebnisse auf. Alle drei Bibliotheken klassifizieren Wörter wie ‚Dollar' oder ‚Covid-19' als ‚Sonstiges'. Die Bibliothek *spaCy* ordnet darüber hinaus Nationalitätsbezüge wie ‚deutsche' oder ‚chinesische' dort ein, während *Stanza* diese ebenfalls als Eigennamen extrahiert, sie jedoch der Klasse ‚LOC' zuordnet (s. Anhang [E]).

Um einen besseren Überblick über die Fehlerarten in den Ergebnissen der einzelnen Bibliotheken zu erhalten, werden diese im Anschluss gebündelt vorgestellt und mit konkreten Beispielen illustriert.

6.2.4 Fehlerausprägungen und -quoten der *NER*-Verfahren

Bei der Prüfung der erhaltenen Ergebnisse aller drei Verfahren fallen verschiedene Arten von Fehlern auf. Da die vorliegende Arbeit zum Ziel hat die Güte der getesteten Verfahren zu beurteilen, wird hier im Detail auf die entdeckten Fehlerausprägungen eingegangen. Dies soll die Transparenz und Nachvollziehbarkeit der Evaluation der Ergebnisse sowie der Berechnung der *Precision*- und *Recall*-Werte gewährleisten.

Fehlklassifikationen

Eingangs soll auf die Eigennamen eingegangen werden, die zwar erkannt, aber einer falschen *NE*-Klasse zugeordnet werden. Diese erhalten bei der Ergebnisprüfung nicht die Markierung als ‚falsch', sondern werden als Fehlklassifikation gezählt und ausgewertet.

Tabelle 8 zeigt, dass der Anteil solcher Fehlklassifikationen im Verhältnis zu allen korrekt klassifizierten Eigennamen bei allen drei *NER*-Verfahren zwischen 0,3% und 6% liegt und damit

recht gering ausfällt. Deutlich wird, dass alle drei Bibliotheken am häufigsten Namen von Organisationen fälschlicherweise der Klasse ‚PER' zuordnen. Das Verfahren von *FLAIR* weist dabei einen ähnlich hohen Anteil an Fehlklassifikationen auf wie *spaCy*, während dies bei *Stanza* seltener vorkommt.

spaCY				
ist jedoch -->	PER	ORG	LOC	Anteil
als PER klassifiziert		126	19	5,1%
als ORG klassifiziert	55		63	2,8%
als LOC klassifiziert	47	201		3,6%

Stanza				
ist jedoch -->	PER	ORG	LOC	Anteil
als PER klassifiziert		49	7	2,3%
als ORG klassifiziert	12		13	0,6%
als LOC klassifiziert	5	12		0,3%

FLAIR				
ist jedoch -->	PER	ORG	LOC	Anteil
als PER klassifiziert		143	6	6,0%
als ORG klassifiziert	11		6	0,6%
als LOC klassifiziert	12	18		0,6%

Tab. 8: Fehlklassifikationen pro Bibliothek
(Quelle: Eigene Darstellung aus Anhang [F])

Eine Schwierigkeit der Zuordnung ist, dass es sich beispielsweise bei der Bezeichnung ‚Anne Will' einerseits um die Sendung und andererseits um den Namen der Moderatorin handeln kann. Ebenso ist es nicht unüblich, dass ein Firmenname den Nachnamen des Gründers der Organisation beinhaltet und häufig die Person und Organisation analog in demselben Text genannt werden (s. Anhang [16], S. 106). Eine korrekte Klassifikation scheint in diesen Fällen kaum ohne menschliches Kontextwissen möglich.

Ähnlich kompliziert verhält es sich bei der maschinellen Verarbeitung von mehrdeutigen Wörtern wie Sportvereinsnamen. Diese werden je nach Verfahren als Person, Organisation oder Ortsbezeichnung klassifiziert, wie in Tabelle 9 zu sehen ist.

Artikel	NE-Klasse	Eigenname	Bibliothek
Basketball-Zweitligist Hagen stellt Antrag auf Kurzarbeit	LOC	Basketball-Zweitligist Phoenix Hagen	spaCy
Basketball-Zweitligist Hagen stellt Antrag auf Kurzarbeit	LOC	Hagen	spaCy
Basketball-Zweitligist Hagen stellt Antrag auf Kurzarbeit	LOC	Hagen	Stanza
Basketball-Zweitligist Hagen stellt Antrag auf Kurzarbeit	ORG	Phoenix Hagen	Stanza
Basketball-Zweitligist Hagen stellt Antrag auf Kurzarbeit	ORG	Phoenix Hagen	FLAIR
Basketball-Zweitligist Hagen stellt Antrag auf Kurzarbeit	PER	Hagen	FLAIR

Tab. 9: Fehlklassifikation von Eigennamen aufgrund ihrer Mehrdeutigkeit
(Quelle: Eigene Darstellung aus Anhang [E])

Darüber hinaus zeigt die *NER*-Analyse, dass Rechtschreibfehler zu Fehlklassifikationen führen können. So wird das falschgeschriebene Wort ‚Frankeich' trotz Kontext von keinem Verfahren als Ort erkannt und stattdessen der Klasse ‚PER' zugeordnet (s. Anhang [18], S. 108). Dies ist ein Hinweis dafür, dass alle drei gewählten *NER*-Verfahren bei informellen Texten mit Rechtschreibfehlern durchaus mehr Fehlklassifikationen aufweisen könnten.

Im Anhang sind noch weitere Beispiele aus dem ‚Corona'-Datensatz aufgeführt, die zu Fehlklassifikationen führen (s. Anhang [17], S. 107). In den dort erfassten Fällen klassifizieren die Verfahren der drei Bibliotheken die Eigennamen unterschiedlich. Bei *spaCy* ist jedoch eine Tendenz zu mehr falschen Klassifikationen sichtbar als bei den anderen zwei Bibliotheken, was sich mit den ermittelten Befunden aus Tabelle 8 deckt.

Zur Beurteilung der Qualität der drei *NER*-Verfahren wird die Durchführung von Fehlklassifikationen in der vorliegenden Arbeit allerdings als weniger gravierend eingestuft. Entscheidender ist die Anzahl an falsch selektierten Wörtern, den zuvor beschriebenen *False Positives*. Diese erhöhen die Menge an automatisiert erhaltenen Ergebnisse mit überflüssigen Daten und beeinträchtigen die Übersichtlichkeit und Auswertung der *NER*-Analyseergebnisse stark.

False Positives

Bei der Sichtung der Ergebnisse werden jene Begriffe als *False Positives* markiert und gewertet, bei denen es sich unverkennbar nicht um Eigennamen handelt. Dies scheinen hauptsächlich Neologismen und Fremdwörter zu sein, wie beispielsweise ‚Super-Spreader' und ‚Brexit' oder Wörter wie ‚Hygge' und ‚Homeoffice' (s. Anhang [E]). Es liegt die Vermutung nahe, dass dies für die *NER*-Algorithmen unbekannte Begriffe sind und daher fälschlicherweise als Eigennamen eingeordnet werden. Dies spiegelt eine in Kapitel 3.3 beschriebene Herausforderung wieder. Wenn sich die zu analysierenden Texte inhaltlich stark von den Trainingsdaten unterscheiden, steigt die Wahrscheinlichkeit für Fehlleistungen der Verfahren (vgl. Maynard 2016: 27).

Bei *spaCy* kommen noch zahlreiche allgemeine Ausdrücke wie ‚Land' oder ‚Staaten' hinzu, die für sich alleinstehend nicht als Eigennamen gewertet werden können. Bei der Betrachtung der über 2.000 ermittelten *False Positives* innerhalb der Klassen ‚PER' und ‚ORG' ist diese Bibliothek für den Großteil dieser falschidentifizierten Wörter verantwortlich (s. Abb. 25).

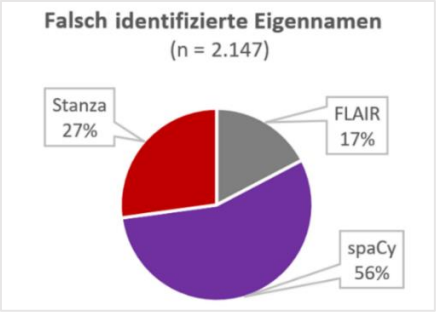


Abb. 25: Anteile an falsch extrahierter ‚PER' und ‚ORG' Ergebnisse
(Quelle: Eigene Darstellung aus Anhang [E])

Sowohl *spaCy* als auch *Stanza* werten Wochentage, Datumsangaben und Emailadressen fälschlicherweise als Eigennamen (s. Anhang [E]). Außerdem ordnet das *NER*-Verfahren von *spaCy* Satzanfänge mit Artikeln fälschlicherweise als Eigennamen ein (s. Anhang [19], S. 108) und

auch falschidentifizierte, mehrdeutige Wörter erhöhen die Anzahl an *False Positives* in den Ergebnissen. So wird zum Beispiel das Wort ‚Ernst' häufig von dem *NER*-Verfahren als Person klassifiziert, obwohl es sich in dem Text nicht um den Namen sondern das Nomen im Ausdruck ‚Ernst der Lage' handelt (s. Tab. 10).

Artikel	NE-Klasse	Eigenname	Bibliothek
»Wir haben günstig gelebt«	MISC	Ernst	spaCy
Diskussion über Ausgangssperren wegen Coronavirus	PER	Ernst	spaCy
Ein großes Experiment	PER	Ernst	spaCy
Ein großes Experiment	PER	Ernst	spaCy
Ein großes Experiment	PER	Ernst	spaCy
Italien-Profi Gosens: «Schlimmste Moment noch nicht überstanden»	PER	Ernst	spaCy
Land bringt «Soforthilfe Corona» für Unternehmen auf den Weg	PER	Ernst	spaCy
Uns bleibt nicht einmal mehr Paris	PER	Ernst	spaCy
Reiseverzicht und schwache Börsen - Coronavirus belastet Wirtschaft	PER	Ernst Kick	FLAIR
Reiseverzicht und schwache Börsen - Coronavirus belastet Wirtschaft	PER	Ernst Kick	spaCy
Reiseverzicht und schwache Börsen - Coronavirus belastet Wirtschaft	PER	Ernst Kick	Stanza

Tab. 10: Auszug der Ergebnisse des NER-Verfahrens von spaCy
(Quelle: Eigene Darstellung aus Anhang [E])

Darüber hinaus extrahiert *Stanza* als einzige Bibliothek irrtümlich die zusammengesetzten Ortsangaben am Anfang der Artikel. Dabei handelt es sich um Angaben wie *(Berlin/Brüssel)*, welche in Nachrichtentexten für die geographische Einordnung des Geschehens genutzt werden (vgl. Rössler 2007: 50).

Überdies weisen alle drei *NER*-Verfahren Schwierigkeiten im Umgang mit den fachspezifischen Begriffen in der Berichterstattung über Krankheitserreger auf. Sowohl bei dem großen Datensatz zu Corona als auch bei dem Probedatensatz mit Texten zu Ebola und Antibiotika-Resistenz, werden Virennamen und andere biomedizinische Fachbegriffe vermehrt für Eigennamen gehalten (s. Anhang [I]).

In Tabelle [20] im Anhang (S. 108) wird gesondert ausgewertet, wie viele Wortkombinationen, welche *Corona, Covid* oder *Sars* beinhalten, als *False Positives* extrahiert werden. Auswertbar ist neben der Menge dieser Wörter, auch die Tatsache, in welche Klasse die jeweiligen Verfahren die Begriffe einordnen. Abbildung 26 stellt dies übersichtlich dar und visualisiert, dass *spaCy* die höchste und *Stanza* die kleinste Anzahl an ‚Corona'-Wortkombinationen extrahiert.

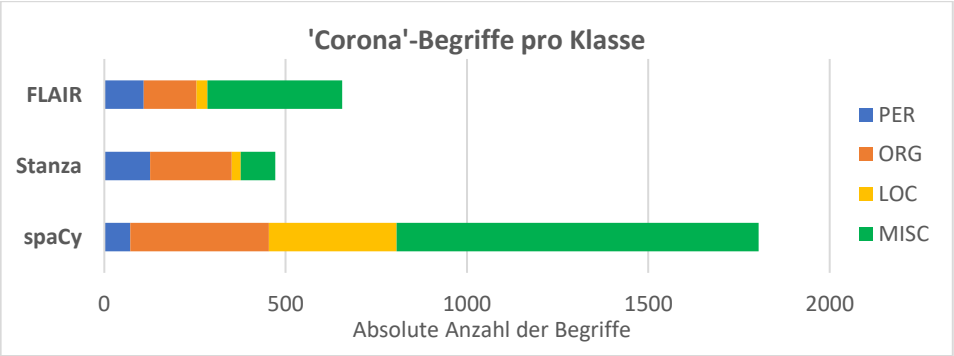


Abb. 26: Menge an 'Corona'-Begriffen in den Ergebnissen je Bibliothek
(Quelle: Eigene Darstellung aus Anhang [E])

Zum Großteil klassifizieren die Bibliotheken die Begriffe als ‚Sonstiges', was nach den deutschen *NER*-Guidelines der korrekte Umgang mit Fachwörtern ist (vgl. Benikova et al. 2014: 2524). Bemerkenswerterweise werden die Wörter aber auch den anderen, unzutreffenden Klassen zugeordnet. Die drei *NER*-Verfahren stufen zwischen 21 und 48% der Begriffe als Organisation ein. Überdies ordnen *Stanza* und *FLAIR* sie im Vergleich zu *spaCy* öfter der Klasse ‚PER' zu (s. Anhang [E]). Beides ist zwar falsch, aber nachvollziehbar, da in vielen Artikeln über die Corona-Pandemie das Virus als Gefahr personifiziert wird und die Ausdrucksweise sowie Satzstruktur davon geprägt sind.

Bei der Untersuchung der erhaltenen *NEs* fällt zusätzlich auf, dass *spaCy* über 200 relativ generische Begriffe wie ‚Abgeordnete' oder ‚DRK-Sprecher' als Eigennamen kennzeichnet, während die anderen zwei Bibliotheken dies nur in seltenen Fällen tun (*Stanza* = 12 Begriffe, *FLAIR* = 26 Begriffe). Obwohl solche allgemeinen Bezeichnungen nach aktuellen *NER-Guidelines* keine Eigennamen darstellen (vgl. Benikova et al. 2014: 2524), kann diskutiert werden, ob diese Ergebnisse nicht trotzdem nützlich sein können. Einige dieser extrahierten Wörter sind ohne Zusatzinformationen recht nutzlos (‚Airlines' oder ‚Streitkräfte'), andere können als alleinstehendes Ergebnis trotzdem aussagekräftig sein (‚Sozialdemokraten' oder ‚Europäische Staaten'). In bestimmten Fällen handelt es sich dabei auch um die Sprecher von Organisationen (‚TUI-Chef' oder ‚Hygiene-Forscher'), welche wichtige Akteure in den Texten darstellen und einen Hinweis zu der Akteursvielfalt der Nachrichtenartikel liefern (s. Anhang [E]).
Es lässt sich eine Parallele dazu erkennen, dass das *NER*-Verfahren von *spaCy* die Institutionsnamen, die aus sehr allgemeinen Begriffen bestehen (‚Auswärtiges Amt' oder ‚Gesundheitsministerium'), als einziges Verfahren als Eigennamen extrahiert. Aus diesem Grund werden die generischen Wörter in den Ergebnissen nicht als Fehler gezählt, sondern extra gekennzeichnet und separat ausgewertet (Kap. 6.3.1).

Chunking

Ein Aspekt, der eine deutliche Unterscheidung zwischen den Leistungen der Bibliotheken zulässt, ist deren Arbeitsweise bei der Extraktion von vollständigen Namenssequenzen. Das Ziel eines *NER*-Verfahrens ist nicht das Erkennen einzelner Bestandteile von Eigennamen, sondern die Identifikation der korrekten Namensgrenzen (s. Kap. 3.3). Bei der Sichtung der erhaltenen Ergebnisse fällt auf, dass dieses *Chunking* bei Doppelnamen von Personen und bei Namenszusätzen in Form von Berufsbezeichnungen zu unterschiedlichen Ergebnissen führt. Tabelle 11 zeigt wie unterschiedlich die Grenzen von gewissen Personennamen identifiziert werden.

NE-Klasse	Eigenname	Bibliothek
PER	Sabine Bätzing-Lichtenthäler	FLAIR
PER	Landesgesundheitsministerin Sabine Bätzing-Lichtenthäler	spaCy
PER	Lichtenthäler	Stanza
PER	Sabine Bätzing	Stanza

Tab. 11: Beispiel für unterschiedliches Chunking der Bibliotheken (Quelle: Eigene Darstellung aus Anhang [E])

Das *NER*-Verfahren von *spaCy* extrahiert häufig zusätzliche Informationen mitsamt den Namen von Personen, wie ‚Charité-Professor Henning Rüden' oder ‚Gesundheitssenator Mario Czaja'. Dies kann einerseits eine hilfreiche, ergänzende Auskunft darstellen, sich jedoch andererseits für die weitere Verarbeitung und Auswertung der Ergebnisse als störender Zusatz erweisen. Eine übergreifende Untersuchung des Vorkommens eines Akteurs kann dadurch beeinträchtigt werden, da der Name durch die uneinheitliche Extraktion möglicherweise nicht überall mitgezählt wird. Wenn ferner mit den erhaltenen Eigennamen weitere Analysen durchgeführt werden sollen, wie die Ermittlung der Prominenz, Reputation oder Publikationen dieser Akteure, ist meist nur der Name von Interesse, um diesen in einer anderen Datenquelle als Suchbegriff zu nutzen. In der hier erhaltenen Form können die Namen jedoch nicht mit einem einfachen Schritt von den Zusatzinformationen getrennt werden, wodurch zusätzlicher manueller Bereinigungsaufwand entsteht, hinter dem sich mögliche Fehlerquellen verbergen.

Das *NER*-Verfahren von *Stanza* weist dagegen bei Wörtern mit Bindestrichen Schwierigkeiten auf, den gesamten Namen als Eigennamen zu erkennen und gibt in zahlreichen Fällen zwei separate Eigennamen aus (s. Tab. 11). Dies stellt bei zusammengesetzten Vor- oder Nachnamen eine große Beeinträchtigung dar. Da bei der *NE*-Extraktion zwei voneinander getrennte Ergebnisse entstehen, bei denen jeweils ein Namensteil fehlt. Eine nachträgliche Zusammenführung der zerteilten Namen kann sich je nach Fall einfacher oder schwieriger gestalten. Insbesondere bei den Namen der Bundesländer fällt auf, dass es einen gravierenden Unterschied macht, wenn fälschlicherweise die Eigennamen ‚Baden' und ‚Sachsen' anstatt der gesamten Eigennamen ‚Baden-Württemberg' und ‚Sachsen-Anhalt' ausgegeben werden. Wenn dieser Fehler unbemerkt bleibt und nicht manuell korrigiert wird, kann er zu falschen Analyseergebnissen führen.

Fehlerquoten

In der Forschungsliteratur wird mehrfach beschrieben, dass die Identifikationsleistung pro *NE*-Klasse unterschiedlich ausfallen kann (vgl. Shelar 2020: 327; vgl. Jiang et al. 2016: 25). Dies bestätigt sich bei der Auswertung der Ergebnisse und wird in Tabelle 12 zusammengefasst. Zu sehen ist, dass die Fehlerraten jeder Bibliothek pro Klasse unterschiedlich stark ausfallen. Die Klasse ‚MISC' ist nicht aufgeführt, da schwierig zu bewerten ist, welche Begriffe von den *NER*-

Verfahren nicht als ‚Sonstiges' eingestuft werden können, während dies bei Personen-, Organisations- und Ortsnamen leichter zu evaluieren ist.

	spaCy			Stanza			FLAIR		
	NEs	Errors	%	NEs	Errors	%	NEs	Errors	%
PER	2.833	417	14,7%	2.480	217	8,8%	2.503	146	5,8%
ORG	4.240	775	18,3%	3.990	366	9,2%	2.968	226	7,6%
LOC	6.882	1.505	21,9%	6.611	1.632	24,7%	4.824	86	1,8%
total	**13.955**	**2.697**	**19,3%**	**13.081**	**2.215**	**16,9%**	**10.295**	**458**	**4,4%**

Tab. 12: NEs und Fehleranteile pro Klasse und Bibliothek
(Quelle: Eigene Darstellung aus Anhang [F])

Aus dem vorherigen Unterkapitel sowie den Ergebnissen des Testdatensatzes zu Antibiotika-Resistenz, Ebola und weiteren Grippepandemien entsteht der allgemeine Eindruck, dass Personennamen besser automatisiert zu erkennen sind als Organisationsnamen (s. Anhang [28], S. 113). Dies kann mit den Ergebnissen aus Tabelle 12 belegt werden. In der Klasse ‚ORG' liegen bei allen drei Bibliotheken die Fehlerquoten höher als in der Klasse ‚PER'.

Außerdem wird in Tabelle 12 deutlich, dass *spaCy* durch alle *NE*-Klassen hinweg die höchsten Fehlerquoten aufweist (19,3%). *Stanza* liegt im Mittelfeld (16,9%) und weist eine auffällig hohe Fehlerrate bei der Identifikation von Orten auf (24,7%). Dies liegt hauptsächlich darin begründet, dass das Verfahren neben Namen von Orten, Gebäuden und Plätzen auch eine Unmenge von Nationalitäten und Regionalbezüge (‚amerikanische' oder ‚baden-württembergischen') extrahiert. Für die vorliegende Untersuchung werden diese alleinstehenden Begriffe in den Ergebnissen als *False Positive* markiert, da sie keine ausreichende Information bei der Suche nach Akteuren im Datensatz bieten.
In Klasse ‚ORG' macht *FLAIR* erstaunlich wenig Fehler (1,8%) und sticht insgesamt mit den geringsten Fehlerquoten positiv hervor (4,4%). Wenn umgekehrt berechnet wird, bei wie vielen der ermittelten *NEs* es sich tatsächlich um Eigennamen handelt, besticht *FLAIR* mit der höchsten Exaktheit (s. Anhang [F]). Die berechneten *Precision*-Werte aller drei Verfahren werden in Tabelle 13 pro *NE*-Klasse dargestellt. Zu sehen ist darin, dass *FLAIR* in allen untersuchten *NE*-Klassen die höchsten Werte aufweist.

Precision pro NE-Klasse			
	spaCy	Stanza	FLAIR
PER	0,85	0,91	0,94
ORG	0,82	0,91	0,92
LOC	0,78	0,75	0,98
Total	**0,81**	**0,83**	**0,96**

Tab. 13: Precision-Werte der drei Verfahren pro NE-Klasse
(Quelle: Eigene Darstellung aus Anhang [F])

Eine allgemein gute *Precision* der *NER*-Verfahren sagt jedoch nichts darüber aus, ob relevante Ergebnisse fehlen, sondern nur wie fehlerarm die vorhandenen Ergebnisse sind.
In den getätigten Auswertungen ist sichtbar, dass das *NER*-Verfahren von *FLAIR* bei der Erkennung von Organisationsnamen beispielsweise knapp ein Viertel weniger Ergebnisse liefert als die anderen zwei Verfahren (s. Abb. 22 und Tab. 12). Auch wenn die fehlerhaften Ergebnisse entfernt werden, weisen *spaCy* und *Stanza* weitaus mehr ermittelte Eigennamen in der Klasse ‚ORG' auf.
Für eine allumfassende Beurteilung der drei Verfahren reicht es daher nicht aus, zu untersuchen, ob es sich bei den ermittelten Wörtern korrekterweise um Eigennamen handelt und wie viele irrelevante Ergebnisse extrahiert werden. Denn damit kann noch keine Aussage darüber getroffen werden, ob womöglich bestimmte Eigennamen überhaupt nicht erkannt wurden. Um diese Leistung bewerten zu können, werden im nächsten Kapitel die automatisiert erhaltenen Ergebnisse der drei Bibliotheken mit den manuell erhobenen Daten abgeglichen. Nur so kann die Vollständigkeit (*Recall*) der Ergebnisse objektiv beurteilt werden.

6.3 Vergleich manueller und automatisierter Erhebungsergebnisse

Nach dem Vergleich der drei *NER*-Verfahren untereinander werden die Ergebnisse der automatisierten Verfahren an den Ergebnissen der manuellen Erhebung desselben Datensatzes gemessen. Hierfür werden auf übergeordneter Ebene zunächst alle Namen der individuellen, generischen und institutionellen Akteure, die manuell selektiert wurden, mit den erhaltenen Eigennamen der *NER*-Verfahren verglichen (s. Anhang [F]).
Zusätzlich wird in einem gesonderten Dokument ein umfangreicher Abgleich auf Artikelebene durchgeführt (s. Anhang [G]). Die ermittelten Befunde beider Analyseebenen stimmen überein und werden im Anschluss pro Akteursgruppe wiedergegeben.

Die durchgeführten Klassifikationen der Eigennamen durch die *NER*-Verfahren werden bei der Ermittlung des *Recalls* nicht bewertet, bleiben jedoch nicht unbeachtet. In 96% der Fälle werden die extrahierten Akteure von den *NER*-Verfahren korrekt klassifiziert (s. Anhang [G]). Dies deckt sich mit den geringen Anteilen an Fehlklassifikationen, die in Kapitel 6.2.4 ermittelt wurden. Es fällt bei der Analyse auf Artikelebene auf, dass vereinzelt der gleiche Akteur in demselben Text verschiedene Klassifikationen erhält. Dies geschieht öfter bei dem *NER*-Verfahren von *spaCy* als bei *Stanza* und *FLAIR*. In dem unten abgebildeten Text zum Beispiel (Abb. 27) wird der Eigenname ‚Ursula von der Leyen' von *spaCy* in einem Satz als ‚ORG' eingeordnet und im nächsten Satz als ‚PER' klassifiziert.

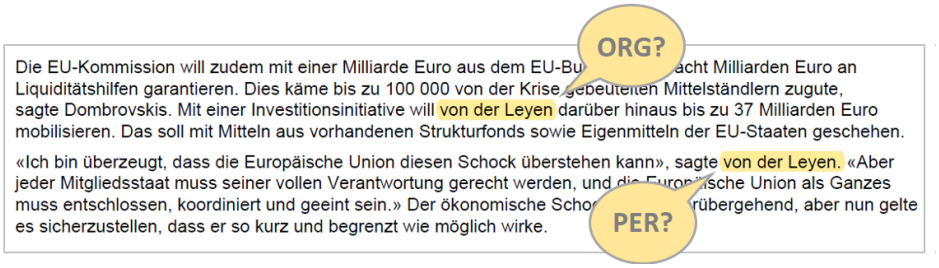


Abb. 27: Text bei dem der gleiche Akteur unterschiedlich klassifiziert wird
(Quelle: dpa-Artikel aus dem Corona-Datensatz – Anhang [B])

Dies zeigt, dass die *NER*-Verfahren auf Satzebene arbeiten und nur darin vorhandene Kontextinformationen werten. Die Klassifikationen werden offenbar nicht übergreifend abgeglichen. Für die Beurteilung der inhaltlichen Vollständigkeit der Ergebnisse kann dies jedoch vernachlässigt werden, da nicht von Bedeutung ist ‚wie', sondern ‚ob' die relevanten Akteure extrahiert werden.

6.3.1 Individuelle Akteure

Die grundsätzliche Menge an erhobenen individuellen Akteuren der manuellen und automatisierten Erhebung unterscheidet sich deutlich, da die Akteure nicht nach denselben inhaltlichen Kriterien aus dem Datensatz selektiert wurden (s. Kap. 5.1).

Doch obwohl die *NER*-Verfahren bis zu 55% zusätzliche Personennamen extrahieren (s. Anhang [21], S. 109), stimmen die meistgenannten Akteure in den Ergebnissen beider Erhebungsmethoden erstaunlich stark überein. Die Auflistung in Tabelle 14 zeigt, dass mit Ausnahme von einem Akteur, die ermittelten Personen des Datensatzes bei der manuellen und der automatisierten Inhaltsanalyse identisch sind.

Vergleich individueller Akteure im Datensatz							
Manuelle Codierung		spaCy		stanza		flair	
Donald Trump	32	Donald Trump	57	Donald Trump	57	Donald Trump	57
Jens Spahn	29	Angela Merkel	56	Angela Merkel	56	Angela Merkel	56
Angela Merkel	23	Jens Spahn	44	Jens Spahn	46	Jens Spahn	45
Christian Drosten	22	Markus Söder	27	Markus Söder	27	Markus Söder	27
Markus Söder	21	Olaf Scholz	26	Olaf Scholz	26	Olaf Scholz	26
Olaf Scholz	15	Christian Drosten	23	Christian Drosten	23	Christian Drosten	23
Winfried Kretschmann	13	Peter Altmaier	17	Ursula von der Leyen	19	Peter Altmaier	17
Heiko Maas	11	Ursula von der Leyen	15	Peter Altmaier	17	Ursula von der Leyen	17
Ursula von der Leyen	11	Winfried Kretschmann	14	Winfried Kretschmann	14	Winfried Kretschmann	14
Manne Lucha	10	Xi Jinping	14	Xi Jinping	14	Xi Jinping	14
Xi Jinping	10	Emmanuel Macron	13	Emmanuel Macron	13	Emmanuel Macron	13
Tedros Ghebreyesus	9	Heiko Maas	13	Heiko Maas	13	Heiko Maas	13
Armin Laschet	8	Horst Seehofer	13	Horst Seehofer	13	Horst Seehofer	13
Emmanuel Macron	8	Armin Laschet	11	Armin Laschet	11	Armin Laschet	11
Horst Seehofer	8	Boris Johnson	11	Tedros Ghebreyesus	11	Boris Johnson	11
Peter Altmaier	8	Tedros Ghebreyesus	11	Boris Johnson	10	Tedros Ghebreyesus	11
Zhong Nanshan	8	Giuseppe Conte	10	Giuseppe Conte	10	Giuseppe Conte	10
Anthony Fauci	7	Franziskus	9	Manne Lucha	10	Manne Lucha	10
Franziskus	7	Sebastian Kurz	9	Franziskus	9	Franziskus	9
Giuseppe Conte	7	Zhong Nanshan	8	Sebastian Kurz	9	Sebastian Kurz	9

Tab. 14: Häufigkeiten manuell und automatisiert erhobener Akteure
(Quelle: Eigene Darstellung aus Anhang [F])

Auch wenn die generelle Menge an Nennungen nicht übereinstimmt, ist durch diese Auswertung sowohl mit der manuellen als der automatisierten Erhebungsmethode ersichtlich, dass die ‚Corona'-Berichterstattung stark von politischen Akteuren geprägt ist. Unter den meist genannten Personen sind zu 80% nationale und internationale Politiker vertreten (s. Tab. 14).
Bei der manuellen Codierung wurde ergänzend die Zugehörigkeit der politischen Akteure erfasst. Diese Information kann mit den *NER*-Verfahren nicht unmittelbar als Ergebniszusatz erhalten werden. Doch die automatisiert extrahierten Nennungen der Parteien im Datensatz weisen hohe Ähnlichkeiten zu den manuell ermittelten Parteienzugehörigkeiten auf. Die Sortierung nach Häufigkeit ergibt eine ähnliche Rangfolge der Parteien, auch wenn die ermittelten Häufigkeiten nicht vergleichbar sind (s. Anhang [22], S. 109).

Für die Berechnung des *Recalls* der *NER*-Verfahren gelten die manuell codierten Akteure als zu erreichender Goldstandard. In dem Datensatz der ‚Corona'-Berichterstattung handelt es sich bei den individuellen Akteuren explizit um 973 verschiedene Personennamen. Der *Recall* gibt somit an, wie viel Prozent der händisch codierten Akteure auch von dem *NER*-Verfahren identifiziert wurden.
Bei der Ermittlung dieses Wertes können zwei Herangehensweisen gewählt werden. Wenn bei dem Vergleich der identifizierten Akteure nur exakt übereinstimmende Personennamen gezählt werden, wird bei der Auswertung von *exact matching* gesprochen (vgl. Jiang et al. 2016: 23). Dieses *exact matching* kann über eine Verweisfunktion in Excel recht schnell erfolgen (s. Anhang [F]).

Entsprechen sich die Namen nicht genau, stimmen aber weitestgehend überein, ist von *loose matching* die Rede (vgl. ebd.). Bei der Auswertung solcher partiellen Treffer werden ergänzend die Eigennamen mitgezählt, bei denen erkennbar ist, dass es sich um den manuell selektierten Akteur handelt, doch ein Namensteil fehlt oder der Name mitsamt einer zusätzlichen Berufsbezeichnung extrahiert wurde. Dies erfordert mehr Arbeitsaufwand, da gezielt nach bestimmten Namensfragmenten innerhalb der erhaltenen *NER*-Ergebnisse gefiltert werden muss. In der vorliegenden Arbeit werden bei dem übergeordneten Namensvergleich beide Auswertungsmethoden durchgeführt, um genau zu erkennen, in welchen Fällen die Personennamen nicht übereinstimmen.

Tabelle 15 zeigt wie viele der 973 individuellen Akteure aus der manuellen Codierung von den jeweiligen *NER*-Verfahren identifiziert wurden. Bei dem *exact matching* weist das *NER*-Verfahren von *FLAIR* die besten Werte auf. Da die *NER*-Verfahren von *spaCy* und *Stanza* einige

Namen fehlerhaft extrahieren (s. Kap. 6.2.4) kommt die in der Tabelle erkennbare größere Differenz zwischen ihren Werten in dem *exact* und *loose matching* zustande.

manuell: 973	spaCy	Stanza	FLAIR
Exact Matching	**925**	**924**	**950**
Recall	0,95	0,95	0,98
Loose Matching	**956**	**955**	**955**
Recall	0,98	0,98	0,98

Tab. 15: Recall-Werte bei der Identifikation von individuellen Akteuren (Quelle: Eigene Darstellung aus Anhang [F])

Werden die partiellen Treffer im *loose matching* mitgezählt, unterscheiden sich die *Recall*-Werte der Verfahren nicht mehr voneinander. Alle drei *NER*-Verfahren finden 98% der manuell erhobenen individuellen Akteure. In Tabelle 15 ist allerdings auch zu sehen, dass trotz *loose matching* keines der drei Verfahren alle 973 manuell extrahierten Personen identifiziert.
Bei der Überprüfung dieser ‚fehlenden' individuellen Akteure wird deutlich, dass es sich um manuell codierte Benamungen handelt, die so nicht im Datensatz vorkommen.
Ein Beispiel ist die manuell extrahierte Person ‚Hildegard Calgéer'. Dieser Name wurde aufgrund menschlicher Abstraktionsfähigkeit aus dem Nachrichtenartikel gezogen, kommt dort jedoch nicht in dieser Form darin vor (s. Abb. 28).

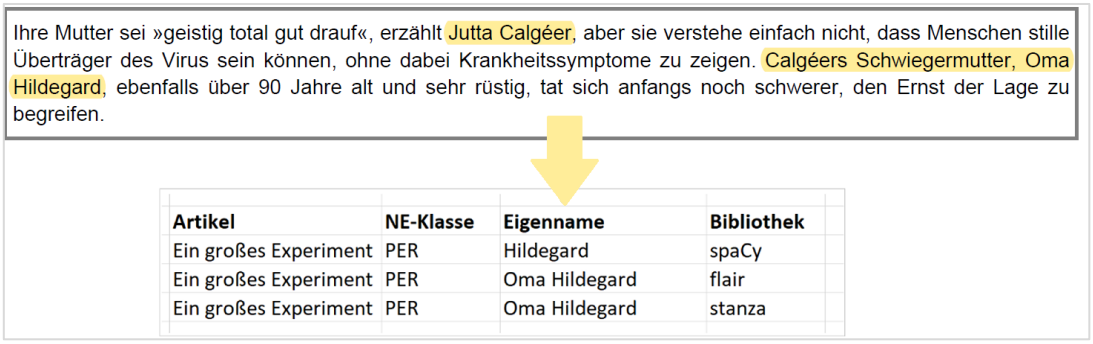


Abb. 28: Textbeispiel mit zugehörigen extrahierten Eigennamen (Quelle: SPIEGEL-Artikel im Datensatz und Screenshot der NER-Ergebnisse aus Anhang [E])

Die *NER*-Verfahren können gewiss nur die im Text vorhandenen Informationen ermitteln und geben daher die in Abbildung 28 aufgeführten Ergebnisse aus. Nur bei detaillierter Prüfung auf Artikelebene kann nachträglich ermittelt werden, dass es sich dabei um den gleichen Akteur handelt.
Ähnlich verhält es sich bei manuell codierten Akteuren, deren Namen nicht in ihrer vollständigen Form automatisiert extrahiert werden, weil sich die Information dazu nicht in dem definierten Analysebereich das *NER*-Verfahrens befinden. Abbildung 29 liefert hierzu ein Beispiel aus dem Datensatz. Zu sehen ist dort, dass in dem linken Artikel der gesamte Personenname nicht in dem Artikeltext vorkommt, ein Phänomen was auch bei sechs weiteren individuellen Akteursnamen festgestellt wird (s. Anhang [F]).

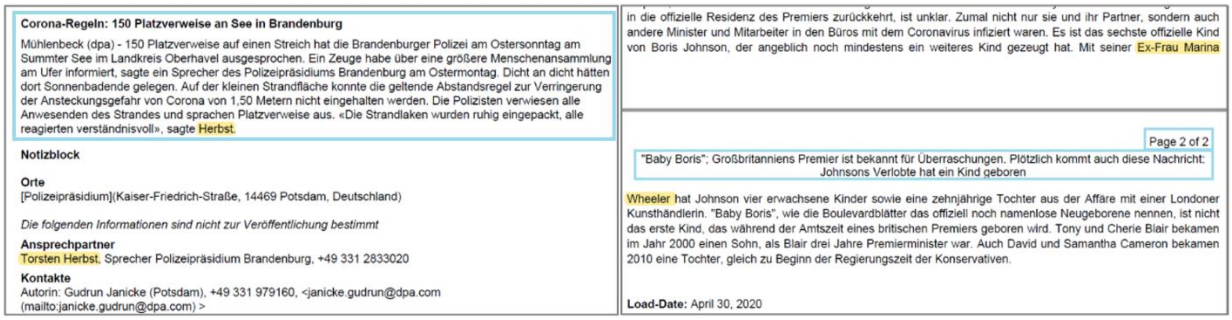


Abb. 29: Eigenname befindet sich nicht im lesbaren Bereich oder wird von Metadaten zerteilt (Quelle: Screenshots eines dpa- und WELT-Artikels aus dem Datensatz – Anhang [B])

Der rechts abgebildete Nachrichtenartikel in dem Beispiel (s. Abb. 29) geht über mehrere Seiten. Die blau markierten Textelemente werden bei der Umwandlung des Artikels in txt-Format in den Fließtext integriert und beeinträchtigen somit die richtige Erkennung der Satzgrenzen und Eigennamen durch die *NER*-Verfahren. Solche Textelemente müssten bei der Umwandlung des Datensatzes in maschinenlesbares Material entfernt oder aus der Definition des zu analysierenden Textbereichs exkludiert werden.

Bei beiden obigen Beispielen hätten menschliche Codierer kaum Probleme bei der Ermittlung der vollständigen Personennamen. Die *NER*-Verfahren können diese jedoch aufgrund von fehlendem Verständnis der Textgrenzen unmöglich identifizieren.

Wertet man diese vorgestellten Problematiken als schwierige Rahmenbedingungen des Vergleichs und bezieht die dennoch erhaltenen Teile der Eigennamen in das *loose matching* mit ein, erzielen alle drei *NER*-Verfahren *Recall*-Werte von 0,99 (s. Anhang [F]; s. Tab. 21).

Bei den übrigen individuellen Akteuren, die als *False Negatives* weder von *Stanza* noch *FLAIR* erkannt werden, handelt es sich um Personen, die nicht mit Vor- und Nachnamen codiert wurden, sondern nur mit ihren Berufsbezeichnungen. Tabelle 16 listet diese auf und lässt erkennen, dass nur das *NER*-Verfahren von *spaCy* zwei dieser Bezeichnungen extrahiert.

Individuelle Akteure	manuell	spaCy	Stanza	FLAIR
Gesundheitsministerin	1	1	#NV	#NV
Französischer Innenminister	1	#NV	#NV	#NV
Präsident	1	#NV	#NV	#NV
Innensenator	1	1	#NV	#NV
Leiter der Medizinischen Fakultät	1	#NV	#NV	#NV
Regionalpräsident	1	#NV	#NV	#NV
Amtsleiter	1	#NV	#NV	#NV
Berliner Senatssprecherin	1	#NV	#NV	#NV
Bezirksbürgermeister	1	#NV	#NV	#NV
Bildungsminister	1	#NV	#NV	#NV
Bürgermeister	1	#NV	#NV	#NV

Tab. 16: Individuelle Akteure, die nicht mit Vor- und Nachnamen codiert wurden (Quelle: Daten aus der Aussagenanalyse des Corona-Datensatzes – Anhang [G])

Die automatisierte Extraktion solcher Akteure ist grundsätzlich schwierig, da die *NER*-Verfahren nicht dafür konzipiert wurden, Berufsbezeichnungen als Eigennamen zu identifizieren (s. Kap. 3.3).

In Kapitel 6.2.3 wurde bereits erläutert, dass *spaCy* solche Begriffe in manchen Fällen nichtsdestotrotz extrahiert. Das Beispiel in Tabelle 16 zeigt jedoch, dass es sich für die hier vorliegende Untersuchung dabei inhaltlich nicht um die relevanten Akteure (*True Positives*) handelt.

Ähnlich verhalt es sich mit der nächsten untersuchten Akteursgruppe, deren Benamungen aus allgemeinen Bezeichnungen bestehen.

6.3.2 Generische Akteure

Zu den händisch codierten, generischen Akteuren zählen in der bearbeiteten Stichprobe 68 Akteure, die weder als Person noch Organisation eingestuft wurden, wie zum Beispiel ‚Wissenschaftler' und ‚Forscher' oder ‚Republikaner'. Von diesen Akteuren identifiziert *FLAIR* keine, *Stanza* extrahiert zwei und *spaCy* zehn (s. Anhang [24], S. 110). Wie bei den zuvor thematisierten Berufsbezeichnungen, werden diese Akteure kaum von den automatisierten Verfahren identifiziert, da es sich dabei nicht um Eigennamen im klassischen Sinne handelt. Tabelle 17 zeigt die aus dem Vergleich resultierenden *Recall*-Werte dieser Akteursgruppe.

***Recall* generischer Akteure**		
spaCy	**stanza**	**flair**
0,15	**0,03**	**0**

Tab. 17: Erzielte Recall-Werte bei generischen Akteuren (Quelle: Eigene Darstellung aus Anhang [F])

Aufgrund der geringen *Recall*-Werte kann an dieser Stelle bereits festgehalten werden, dass sich keines der gewählten *NER*-Verfahren eignet, um verlässlich und vollständig generische Akteure aus Texten zu extrahieren.

6.3.3 Institutionelle Akteure

Bei den manuell codierten institutionellen Akteuren handelt es sich um 862 verschiedene Namen von Organisationen, Institutionen und Behörden. Vereinzelt wurden auch Staaten und deutsche Bundesländer als institutionelle Akteure codiert, weshalb für den Vergleich auch die extrahierten Eigennamen der *NE*-Klasse ‚LOC' einbezogen werden.

Im Vergleich zu den Personennamen (Kap. 6.3.1) erkennen die *NER*-Verfahren einen geringeren Anteil an institutionellen Akteuren, sowohl bei dem direkten Namensvergleich, als auch bei der Analyse auf Artikelebene (s. Anhang [G]).

Die Auswertung des Testdatensatzes über Antibiotika-Resistenz und andere Grippepandemien weist ebenfalls eine schwächere Identifikationsleistung bei institutionellen Akteuren auf. Im Anhang (s. [28], S. 113) findet sich eine detaillierte Auflistung der darin manuell codierten Akteure pro Thema und Medientitel mit einem Vermerk, ob der jeweilige Akteur automatisiert gefunden wurde oder nicht.

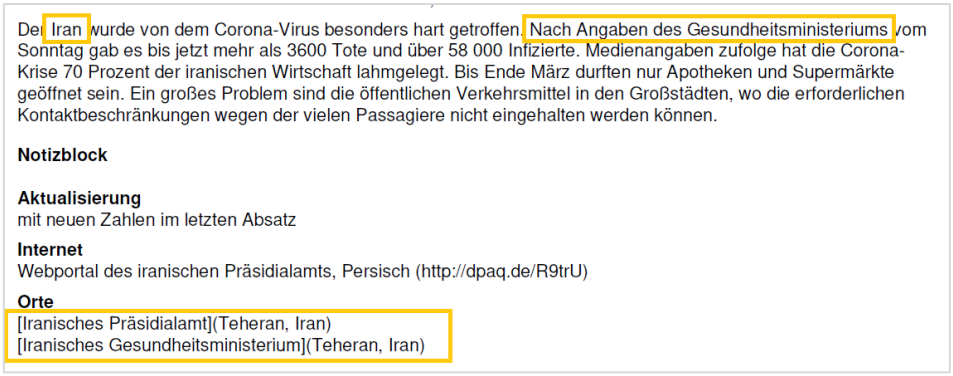
Für den ‚Corona'-Datensatz bildet Tabelle 18 die resultierenden *Recall*-Werte der drei *NER*-Verfahren aus dem Vergleich der manuell und automatisiert extrahierten Akteure ab. In der Tabelle ist zu sehen, dass bei dieser Akteursgruppe *spaCy* die meisten Treffer liefert, während *FLAIR* die schlechteste Leistung erzielt.

manuell: 862	spaCy	Stanza	FLAIR
Exact Matching	**686**	**633**	**581**
Recall	0,80	0,73	0,67
Loose matching	**697**	**644**	**590**
Recall	0,81	0,75	0,68

Tab. 18: Recall-Werte bei der Identifikation institutioneller Akteure (Quelle: Eigene Darstellung aus Anhang [F])

Bei der Ergebnisauswertung wird ebenfalls ein Vergleich mittels *exact* und *loose matching* durchgeführt. Auch hier tritt vermehrt der Fall ein, dass relevante Akteure identifiziert, aber nicht in der Form extrahiert werden, wie die menschlichen Codierer dies durchführen. Dem *NER*-Verfahren von *spaCy* kommt zugute, dass es Eigennamen oft mit Zusatzinformationen ausgibt. Die längeren *Chunks* in den Ergebnissen von *spaCy* erzielen öfter exakte Übereinstimmungen mit den erhobenen Daten der manuellen Inhaltsanalyse (s. Anhang [23], S. 109).

Zusätzlich tritt auch bei den institutionellen Akteuren das Problem auf, dass einige manuell codierte Namen nicht in dieser Form in den analysierten Nachrichtentexten vorkommen. Teilweise ist menschliche Abstraktionsfähigkeit und Kontextwissen erforderlich, um beispielsweise ‚Gesundheitsministerium Iran' als Akteur aus dem Text in Abbildung 30 zu extrahieren.


Der Iran wurde von dem Corona-Virus besonders hart getroffen. Nach Angaben des Gesundheitsministeriums vom Sonntag gab es bis jetzt mehr als 3600 Tote und über 58 000 Infizierte. Medienangaben zufolge hat die Corona-Krise 70 Prozent der iranischen Wirtschaft lahmgelegt. Bis Ende März durften nur Apotheken und Supermärkte geöffnet sein. Ein großes Problem sind die öffentlichen Verkehrsmittel in den Großstädten, wo die erforderlichen Kontaktbeschränkungen wegen der vielen Passagiere nicht eingehalten werden können.

Notizblock

Aktualisierung
mit neuen Zahlen im letzten Absatz

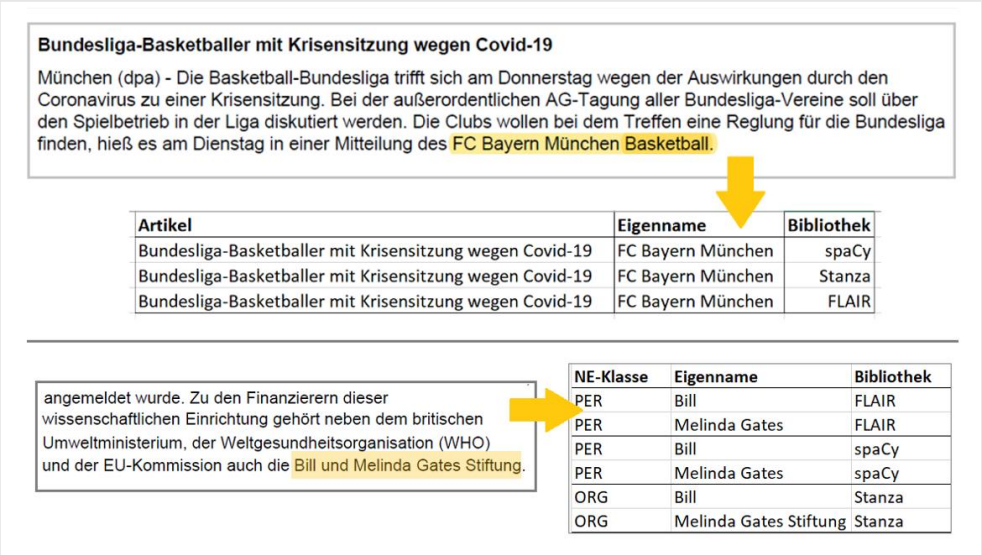
Internet
Webportal des iranischen Präsidialamts, Persisch (http://dpaq.de/R9trU)

Orte
[Iranisches Präsidialamt](Teheran, Iran)
[Iranisches Gesundheitsministerium](Teheran, Iran)

Abb. 30: Institutioneller Akteur nur aus Gesamtkontext ersichtlich (Quelle: dpa-Artikel aus dem Corona-Datensatz)

Der Eigenname ‚Iran' wird von allen drei *NER*-Verfahren identifiziert, doch nur *spaCy* extrahiert das Wort ‚Gesundheitsministerium' (s. Anhang [E]). Als getrennte Ergebnisse sind diese Eigennamen jedoch nur bedingt für eine Akteursanalyse hilfreich.
Zusätzlich stellt der Artikel aus Abb. 30 ein Beispiel dafür dar, dass die Bezeichnung des relevanten Akteurs (‚Iranisches Gesundheitsministerium') nicht im analysierten Textbereich steht. Eine Problematik, die bereits bei dem Vergleich der individuellen Akteure aufgefallen war.

In anderen Fällen ermittelt keines der drei Verfahren den gesamten Eigennamen richtig. Zwei Beispiele hierfür sind der Sportverein ‚FC Bayern München Basketball' und die ‚Bill und Melinda Gates Stiftung', dargestellt in Abbildung 31.



Bundesliga-Basketballer mit Krisensitzung wegen Covid-19

München (dpa) - Die Basketball-Bundesliga trifft sich am Donnerstag wegen der Auswirkungen durch den Coronavirus zu einer Krisensitzung. Bei der außerordentlichen AG-Tagung aller Bundesliga-Vereine soll über den Spielbetrieb in der Liga diskutiert werden. Die Clubs wollen bei dem Treffen eine Reglung für die Bundesliga finden, hieß es am Dienstag in einer Mitteilung des FC Bayern München Basketball.

Artikel	Eigenname	Bibliothek
Bundesliga-Basketballer mit Krisensitzung wegen Covid-19	FC Bayern München	spaCy
Bundesliga-Basketballer mit Krisensitzung wegen Covid-19	FC Bayern München	Stanza
Bundesliga-Basketballer mit Krisensitzung wegen Covid-19	FC Bayern München	FLAIR

angemeldet wurde. Zu den Finanzierern dieser wissenschaftlichen Einrichtung gehört neben dem britischen Umweltministerium, der Weltgesundheitsorganisation (WHO) und der EU-Kommission auch die Bill und Melinda Gates Stiftung.

NE-Klasse	Eigenname	Bibliothek
PER	Bill	FLAIR
PER	Melinda Gates	FLAIR
PER	Bill	spaCy
PER	Melinda Gates	spaCy
ORG	Bill	Stanza
ORG	Melinda Gates Stiftung	Stanza

Abb. 31: Eigennamen von institutionellen Akteuren die fehlerhaft extrahiert werden (Quelle: Eigene Darstellung aus Anhang [E] und dpa-Artikel Anhang [B])

Bei dem obigen Beispiel der extrahierten Ergebnisse (s. Abb. 31) wird schlicht das generische Wort ‚Basketball' nicht zum Eigennamen zugehörig gezählt. In dem darunter abgebildeten Beispiel trennen die *NER*-Verfahren den Organisationsnamen in mehrere *Chunks*. Bei genauer Prüfung der *NER*-Ergebnisse wird sichtbar, dass alle drei Verfahren bei zusammengesetzten Namenskonstellationen von Eheleuten (‚Tony und Cherie Blair') zwei getrennte *Named Entities* ausgeben (‚Tony'/‚Cherie Blair').
In den *NER-Guidelines für* deutschsprachige Texte von Benikova et al. wird nicht definiert, ob es sich dabei um ein korrektes Vorgehen der *NER*-Verfahren handelt. Gleichwohl stellen die dadurch erhaltenen einzelnen Vornamen ohne Zuordnung wenig aussagekräftige Ergebnisse für eine kommunikationswissenschaftliche Akteursanalyse dar. Im oben dargestellten Beispiel führen die getrennt extrahierten *Chunks* außerdem dazu, dass nur das *NER*-Verfahren von *Stanza* den Namen der Stiftung als ‚ORG' klassifiziert.

Werden die zwanzig meistgenannten institutionellen Akteure der manuellen Erhebung des ‚Corona'-Datensatzes mit den zwanzig am häufigsten extrahierten Eigennamen der Klassen ‚ORG' und ‚LOC' verglichen, stimmen jeweils nur die Hälfte der Namen in der Auflistung überein (s. Kap. 6.2.3). In Tabelle 19 sind die meistgenannten institutionellen Akteure aus der manuellen Erhebung abgebildet. Die manuell codierten Organisationen werden für den Vergleich von den codierten Ortsbezeichnungen getrennt aufgelistet. Farblich markiert ist, welche

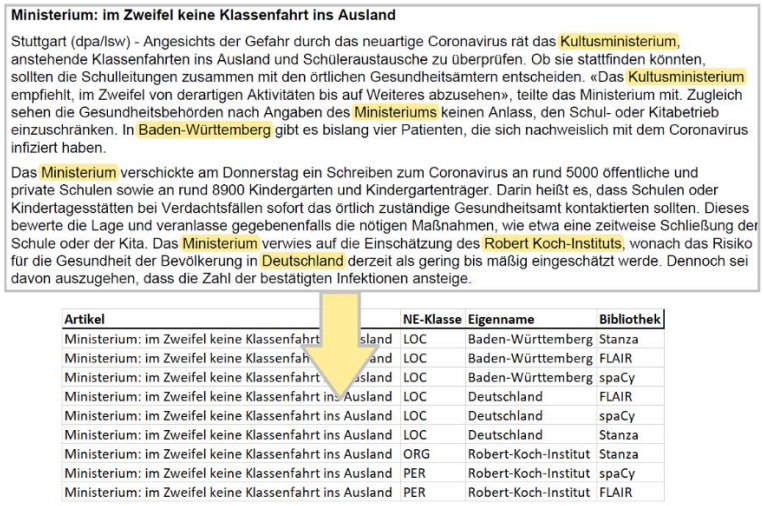
dieser institutionellen Akteure nicht mit den am häufigsten automatisiert extrahierten Ergebnissen übereinstimmen (s. Anhang [F]).

Manuell selektierte institutionelle Akteure			
WHO	39	USA	50
RKI	36	Baden-Württemberg	37
EU	19	Frankreich	27
SPD	19	Bayern	25
Berliner Charité	18	Italien	18
FDP	16	China	16
CDU	15	Nordrhein-Westfalen	16
CDC	12	Großbritannien	11
Die Grünen	12	Österreich	11
Bundesgesundheitsministerium	9	Berlin	9
DFB	9	Japan	8
UN	9	New York	8
AfD	7	Thüringen	8
DIHK	6	Niedersachsen	7
Imperial College London	6	Polen	7
IOC	6	Brasilien	6
CureVac	5	Dänemark	6
Apple	5	Hessen	6
BCG	5	Spanien	6
BDI	5	Hamburg	5

Tab. 19: Die häufigsten manuell selektierten Akteure getrennt nach Organisations- und Ortsnamen (Quelle: Eigene Darstellung aus Anhang [F])

Wie bereits in Kap. 6.2.3 festgestellt wurde, werden von *FLAIR* einige allgemeine Bezeichnungen wie ‚Auswärtiges Amt' oder ‚Hanois Volkskomitee' grundsätzlich nicht als Eigennamen identifiziert (s. Anhang [E]). Doch auch die *NER*-Verfahren von *spaCy* und *Stanza* weisen bei diesen generischen Begriffen Identifikationsschwierigkeiten auf.

Ein letztes Beispiel zeigt einen Nachrichtenartikel, in welchem keines der drei *NER*-Verfahren das darin genannte Ministerium als Eigennamen ermittelt (s. Abb. 32). Die erhaltenen zugehörigen Ergebnisse des Textes sind mitabgebildet.



Ministerium: im Zweifel keine Klassenfahrt ins Ausland

Stuttgart (dpa/lsw) - Angesichts der Gefahr durch das neuartige Coronavirus rät das Kultusministerium, anstehende Klassenfahrten ins Ausland und Schüleraustausche zu überprüfen. Ob sie stattfinden könnten, sollten die Schulleitungen zusammen mit den örtlichen Gesundheitsämtern entscheiden. «Das Kultusministerium empfiehlt, im Zweifel von derartigen Aktivitäten bis auf Weiteres abzusehen», teilte das Ministerium mit. Zugleich sehen die Gesundheitsbehörden nach Angaben des Ministeriums keinen Anlass, den Schul- oder Kitabetrieb einzuschränken. In Baden-Württemberg gibt es bislang vier Patienten, die sich nachweislich mit dem Coronavirus infiziert haben.

Das Ministerium verschickte am Donnerstag ein Schreiben zum Coronavirus an rund 5000 öffentliche und private Schulen sowie an rund 8900 Kindergärten und Kindergartenträger. Darin heißt es, dass Schulen oder Kindertagesstätten bei Verdachtsfällen sofort das örtlich zuständige Gesundheitsamt kontaktierten sollten. Dieses bewerte die Lage und veranlasse gegebenenfalls die nötigen Maßnahmen, wie etwa eine zeitweise Schließung der Schule oder der Kita. Das Ministerium verwies auf die Einschätzung des Robert Koch-Instituts, wonach das Risiko für die Gesundheit der Bevölkerung in Deutschland derzeit als gering bis mäßig eingeschätzt werde. Dennoch sei davon auszugehen, dass die Zahl der bestätigten Infektionen ansteige.

Artikel	NE-Klasse	Eigenname	Bibliothek
Ministerium: im Zweifel keine Klassenfahrt [illegible]sland	LOC	Baden-Württemberg	Stanza
Ministerium: im Zweifel keine Klassenfah[illegible]and	LOC	Baden-Württemberg	FLAIR
Ministerium: im Zweifel keine Klassenfahrt [illegible]sland	LOC	Baden-Württemberg	spaCy
Ministerium: im Zweifel keine Klassenfahrt ins Ausland	LOC	Deutschland	FLAIR
Ministerium: im Zweifel keine Klassenfahrt ins Ausland	LOC	Deutschland	spaCy
Ministerium: im Zweifel keine Klassenfahrt ins Ausland	LOC	Deutschland	Stanza
Ministerium: im Zweifel keine Klassenfahrt ins Ausland	ORG	Robert-Koch-Institut	Stanza
Ministerium: im Zweifel keine Klassenfahrt ins Ausland	PER	Robert-Koch-Institut	spaCy
Ministerium: im Zweifel keine Klassenfahrt ins Ausland	PER	Robert-Koch-Institut	FLAIR

Abb. 32: Artikel mit institutionellen Akteuren und zugehörige NER-Ergebnisse (Quelle: Screenshot dpa-Artikel aus dem Corona-Datensatz – Anhang [B])

Dies illustriert, weshalb in dieser Untersuchung grundsätzlich weniger institutionelle als individuelle Akteure automatisiert identifiziert werden und die etwas geringeren *Recall*-Werte zustande kommen (s. Tab. 18).

Dennoch bleibt unklar, warum in anderen Artikeln Eigennamen wie ‚Gesundheitsministerium' und ‚Außenministerium' extrahiert werden (s. Anhang [E]). Dies könnte zum einen von dem jeweiligen Satzbau oder Kontext beeinflusst worden sein, in dem diese Wörter in dem Datensatz vorkommen. Zum anderen könnten die Begriffe in den Trainingstexten der Verfahren vorgekommen und unterschiedlich annotiert worden sein und deswegen zu den uneinheitlichen Ergebnissen führen. Es kann sich bei den allgemeinen Bezeichnungen der staatlichen Behörden und Ministerien jedoch auch um Begriffe handeln, die in anderen Untersuchungen als *False Positives* gewertet worden wären. Für den vorliegenden Vergleich gelten sie jedoch als Maßstab und beeinflussen die erzielten *Recall*-Werte aller drei Verfahren.

Nach der Berechnung und Erläuterung der *Precision* und *Recall*-Werte der *NER*-Verfahren, wird abschließend eine Zusammenfassung der Befunde gegeben. Darauf aufbauend kann im Anschluss die Gütebeurteilung der Verfahren erfolgen.

6.4 Zusammenfassung der Befunde

Mithilfe der durchgeführten Gegenüberstellung der drei *NER*-Verfahren sowie dem Vergleich der extrahierten Eigennamen mit den manuell erhobenen Daten, wird ein umfassender Überblick über die Stärken und Schwächen der einzelnen Verfahren erhalten. Diese werden in der nachfolgenden Tabelle aufgeführt und kompakt erläutert.

	spaCy	Stanza	FLAIR
Stärken	- schnellstes Verfahren - Extraktion von teilweise relevanten, generischen Organisationsnamen *(Auswärtiges Amt, Gesundheitsministerium)*, dadurch höchster *Recall* bei institutionellen Akteuren	- Beste Klassifikation der Eigennamen in korrekte *NE* -Klasse	- Höchste *Precision* bei der Identifikation von Eigennamen - Höchster *Recall* bei der Erkennung von individuellen Akteuren
Schwächen	- Störende Namenszusätze bei NE-Extraktion *(CDU-Bundestagsabgeordneter Philip Amthor)* - Extrahiert Satzanfänge mit Artikeln, liefert die meisten False Positives - Extrahiert Nationalitätsbezüge (deutsche, chinesische) als NE	- Teils falsches Chunking bei Namen mit Bindestrichen (Erhalt von zwei getrennten NEs) - Extrahiert Nationalitätsbezüge (deutsche, chinesische) als NE	- langsamstes Verfahren - höchste Fehlklassifikation von Organisationen als 'PER' - Geringster *Recall* bei der Erkennung von 'ORG'

Tab. 20: Übersicht der Stärken und Schwächen je Bibliothek (Quelle: Eigene Darstellung)

Die Gegenüberstellung der drei Bibliotheken lässt erkennen, dass *spaCy* das schnellste *NER*-Verfahren ist, welches mit *displaCy* zudem die Möglichkeit von integrierten Ergebnisvisualisierungen bietet. Das *NER*-Verfahren dieser Bibliothek extrahiert als Einziges allgemeine Berufsbezeichnungen aus den Texten, weist gleichzeitig allerdings auch die größte Menge an irrelevanten Ergebnissen und Fehlklassifikationen auf.

Diese hohe Anzahl an *False Positives*, bei denen es sich nicht um Eigennamen handelt, führt dazu, dass das Verfahren insgesamt betrachtet in allen *NE*-Klassen die geringste ‚Verlässlichkeit' birgt und die unübersichtlichsten Ergebnisse liefert.
Die extrahierten generischen Eigennamen (z. B. ‚Abgeordnete' oder ‚EU-Wirtschaftskommissar') erweisen sich nach dem Vergleich mit den Ergebnissen der manuellen Erhebung als nicht relevante generische Akteure. Nichtsdestotrotz scheint die Identifikation von Eigennamen, die aus recht allgemeinen Begriffen bestehen, für die grundsätzliche Extraktion von generischen Organisations- und Institutionsnamen (z.B. ‚Finanzministerium') von Vorteil zu sein.
Denn bei der Auswertung der ‚Vollständigkeit' der gelieferten Ergebnisse übertrifft das Verfahren von *spaCy* bei der Identifikation von institutionellen Akteuren im Datensatz die Leistung der beiden anderen Verfahren.

Die Leistung des *NER*-Verfahrens der Bibliothek *Stanza* liegt bei allen betrachteten Bewertungsvariablen im Mittelfeld zwischen den Ergebnissen von *spaCy* und *FLAIR*. Lediglich bei der Klassifikation der extrahierten Eigennamen sticht dieses Verfahren mit der geringsten Anzahl an Fehlklassifikationen heraus. Aufgrund der Extraktion vieler *False Positives* sowie dem teils fehlerhaften Umgang mit Eigennamen, die Bindestriche beinhalten, erreicht das Verfahren von *Stanza* im *exact matching* keine vergleichbar hohen Werte. Es ist bei der Erkennung von Eigennamen zwar durchweg präziser als *spaCy*, doch erzielt es in keiner *NE*-Klasse bessere *Precision*-Werte als *FLAIR*.

Das *NER*-Verfahren der Bibliothek *FLAIR* benötigt die längte Verarbeitungszeit, doch es erreicht im Gegenzug die höchsten *Recall*- und *Precision*-Werte bei der Identifikation von individuellen Akteuren im Datensatz. Im Vergleich zu *spaCy* und *Stanza* weist es keine erkennbaren systematischen Fehler bei der Extraktion von Eigennamen auf. Außerdem extrahiert es den kleinsten Anteil an *False Positives* und erzielt somit über alle *NE*-Klassen hinweg die geringsten Fehlerquoten. Nichtsdestotrotz weist das Verfahren von *FLAIR* in dieser Untersuchung eine große Schwäche bei der Erkennung von Akteursnamen bestehend aus generischen Bezeichnungen auf. Es erkennt nur 67% der manuell codierten institutionellen Akteure und liegt damit weit hinter den *Recall*-Werten von *spaCy* und *Stanza*.

Tabelle 21 listet die berechneten Kennzahlen gebündelt auf und bildet die resultierenden *F-Scores* als harmonisches Mittel zwischen den *Precision*- und *Recall*-Werten ab. Dieser Durchschnittswert verwässert die herausgearbeiteten Stärken und Schwächen der Verfahren ein wenig, ermöglicht jedoch eine Zusammenfassung der Ergebnisse auf ganzheitlicher Ebene.

		individuelle Akteure				institutionelle Akteure		
		spaCy	Stanza	FLAIR		spaCy	Stanza	FLAIR
Exact Matching	Recall	0,95	0,95	0,98	Recall	0,80	0,73	0,67
	Precision*	0,85	0,91	0,94	Precision**	0,80	0,81	0,96
	F-Score	**0,90**	**0,93**	**0,96**	**F-Score**	**0,80**	**0,77**	**0,79**
Loose Matching	Recall	0,99	0,99	0,99	Recall	0,81	0,75	0,68
	Precision*	0,85	0,91	0,94	Precision**	0,80	0,81	0,96
	F-Score	**0,91**	**0,95**	**0,96**	**F-Score**	**0,80**	**0,78**	**0,80**

*berechnet aus NE-Klasse 'PER'

**berechnet aus NE-Klassen 'ORG' und 'LOC'

Tab. 21: Zusammenfassung der Precision- und Recall-Werte nach Akteursgruppe (Quelle: Eigene Darstellung aus Anhang [F])

Mittels der *F-Scores* ist bei der Bewertung der Identifikation von individuellen Akteuren die Bestleistung des Verfahrens von *FLAIR* bestimmbar (*F-Score*: 0,96). Bei der Erkennung institutioneller Akteure ist kein so eindeutiges Ergebnis zu sehen. Das Verfahren von *spaCy* weist die höchsten *Recall*-Werte auf, während *FLAIR* durch die höchste *Precision* letztlich einen ebenso hohen *F-Score* erzielt.

Die oben ermittelten Werte, getrennt nach Akteursgruppe sowie *Precision* und *Recall*, ermöglichen eine konkretere Evaluation der Qualität der Ergebnisse als eine zusammengefasste Totalansicht der *F-Scores* pro Bibliothek.

Die Angabe eines Gesamtwerts ist jedoch in der Literatur üblich und wird für die abschließende Bewertung der Leistung der *NER*-Verfahren im nächsten Kapitel berechnet. Dadurch können die hier erhaltenen Ergebnisse des Vergleichs final mit den in der Forschungsliteratur aufgeführten *F-Scores* in Bezug gesetzt werden.

7. Gütebeurteilung und Validierung der *NER*-Verfahren

Bei manuellen Inhaltsanalysen werden in den Sozialwissenschaften Reliabilitätstests zur Qualitätssicherung und Sicherstellung der Güte der durchgeführten Messungen eingesetzt (vgl. Dumm/Niekler 2014: 21). Beim Einsatz von automatisierten Analysemethoden muss diese Gütebeurteilung auf andere Art und Weise erfolgen. Auf technischer Ebene ist ein automatisiertes Verfahren vollständig zuverlässig und reproduzierbar (vgl. Scharkow 2013: 290). Auf inhaltlicher Ebene muss jedoch stets geprüft werden, ob die erhaltenen Ergebnisse gültig sind (vgl. ebd.).
Zwar fehlen in den *Computational Social Sciences* gegenwärtig universelle Richtwerte zur Bewertung von automatisierten Verfahren (vgl. Niemann-Lenz et al. 2019: 3892), doch in der Forschungsliteratur wird vermehrt empfohlen die automatisiert erhaltenen Ergebnisse mit manuell durchgeführten Erhebungen zu vergleichen (vgl. Grimmer/Stewart 2013: 271, vgl. Burggraaff/Trilling 2020: 125; vgl. Schwotzer 2014: 55, vgl. Scharkow 2019: 280). Als Gütekriterien werden dabei die Leistungskennzahlen *Precision* und *Recall* genutzt, welche im Fachbereich der angewandten Informatik für die Leistungsevaluation von *NLP*-Aufgaben entwickelt wurden (vgl. Dumm/Niekler 2014: 21; vgl. Derczynski 2016: 262).

Zur Validierung einer analytischen Methode wird geprüft, ob diese für die Aufgabe geeignet ist, die sie erfüllen soll (vgl. Kromidas 2011: 4). Um zu bestätigen, dass sie für den beabsichtigten Gebrauch eingesetzt werden kann, muss eine Untersuchung sowie die Bereitstellung eines Validitätsnachweises erfolgen (vgl. ebd.). Erst nach der Überprüfung der Validität, erlaubt ihr Einsatz in der empirischen Forschung eine gültige Interpretation der erhaltenen Befunde (vgl. Fleischhauer et al. 2008: 326). Die durchgeführte Untersuchung der vorliegenden Arbeit kann als Überprüfung und dokumentierter Nachweis angesehen werden, um zu validieren, ob sich der Einsatz der untersuchten *NER*-Verfahren für die automatisierte Identifikation von Akteuren in journalistischen Texten eignet.

7.1 Eignung der *NER*-Verfahren zur Identifikation von Akteuren

Um die Güte der getesteten Verfahren zu beurteilen und für den Einsatz in zukünftigen inhaltsanalytischen Untersuchungen zu validieren, werden die insgesamt erzielten Leistungskennzahlen auf ganzheitlicher Ebene berechnet.
Gegenwärtig sind in der Literatur keine Maßstäbe angesetzt, welche *F-Score*-Werte anzustreben oder für eine erfolgreiche Validierung zu erfüllen sind. Die aufgeführten Werte anderer Publikationen liegen für deutschsprachige Texte zwischen 0,64 und 0,86 (s. Kap. 4.2).

Die hier ermittelten *F-Scores* der *NER*-Verfahren von *spaCy* und *Stanza* liegen in diesem Wertebereich, während das Verfahren von *FLAIR* diese übertrifft. Tabelle 22 bildet diese Gütemaße für alle untersuchten *NE*-Klassen zusammengefasst ab.

Gesamtleistung (PER, ORG & LOC)		spaCy	Stanza	FLAIR
Exact Matching	Recall	0,88	0,85	0,83
	Precision	0,81	0,83	0,96
	F-Score	**0,84**	**0,84**	**0,89**
Loose Matching	Recall	0,90	0,87	0,84
	Precision	0,81	0,83	0,96
	F-Score	**0,85**	**0,85**	**0,90**

Tab. 22: Darstellung der übergreifenden F-Scores pro Bibliothek (Quelle: Eigene Darstellung aus Anhang [F])

Darüber hinaus sind auf den Plattformen, auf denen die Code-*Packages* der jeweiligen Bibliotheken veröffentlicht werden, aktuelle Angaben zu den höchsterzielten Werten der *NER*-Verfahren für die deutschen Sprachmodelle vermerkt. Die dort angegebenen Werte decken sich gut mit den hier erhaltenen finalen *F-Scores.* Das *NER*-Verfahren von *Stanza* wird mit einem F-*Score* von 0,85 ausgewiesen (vgl. Qi et al. 2020: 6), *spaCy* soll *F-Scores* von 0,86 erreichen (vgl. Honnibal/Montani 2017: o. S.) und *FLAIR* wirbt mit einem Wert von 0,88 (vgl. Akbik et al. 2018: 1645).

Im Vorfeld war ungewiss, wie gut sich vorab trainierte *ML*-Verfahren auf unbekannte Datensätze anwenden lassen. Es wurden zahlreiche Einflussvariablen herausgearbeitet, die die grundsätzliche Leistung der *NER*-Verfahren beeinflussen können (s. Kap. 3.4). Doch die errechneten Kennzahlen der Untersuchung zeigen, dass die getesteten *NER*-Verfahren durchaus in der Lage sind, in unbekannten Textdaten mit hoher ‚Verlässlichkeit' Eigennamen zu identifizierten.

Schwieriger ist die Bewertung, ob die ‚Vollständigkeit' der gelieferten Ergebnisse ausreichend für die Ermittlung aller relevanten Akteure in journalistischen Texten ist. Für diese Beurteilung muss die Art der Akteure spezifiziert werden, die aus den Texten extrahiert werden soll, da sich die Identifikationsleistung pro Akteursgruppe stark unterscheidet (s. Kap. 6.4).

So ist keines der drei *NER*-Verfahren für die Extraktion generischer Akteure geeignet, da es sich bei diesen nicht um Eigennamen von Personen oder Organisationen handelt und keines der Verfahren ausreichend darauf trainiert wurde (s. Kap. 6.3.2).

Bei der Identifikation individueller und institutioneller Akteure hingegen weisen alle getesteten Verfahren hohe Trefferquoten auf (s. Tab. 21). Im Vergleich zu den erzielten Werten bei der Identifikation von Personen lässt die Erkennungsleistung von Institutions- und Organisationsnamen jedoch noch Raum für Verbesserungen.

Vermutlich könnte die Leistung der *NER*-Verfahren bei der Erkennung von institutionellen und generischen Akteuren durch entsprechendes Training mit annotierten Daten optimiert werden.

Würden die Verfahren mit Textkorpora trainiert, welche die herausgearbeiteten Fremdwörter beinhalten (s. Kap. 6.2.4), ließe sich voraussichtlich auch die Menge an ausgegebenen *False Positives* verringern. Die Erstellung solch eines Textkorpus und das Training der *NER*-Verfahren ist allerdings nur mit entsprechenden Fachkenntnissen und geeigneten Trainingsdaten möglich (vgl. Niekler 2016: 8). Für Fachfremde stellt dies keine einfache, in den Forschungsalltag leicht zu integrierende Tätigkeit dar.

Die im Theorieteil vorgestellten Herausforderungen bei der maschinellen Verarbeitung natürlicher Sprache (s. Kap. 3.3) finden sich bei der Sichtung und Fehleranalyse der erhaltenen Ergebnisse wieder. Die unterschiedlichen Fehler, die bei der Extraktion von Eigennamen erkannt werden, beeinträchtigen insbesondere die Arbeit mit den ausgegebenen Analyseergebnissen von *Stanza* und *spaCy* (s. Kap. 6.2.4). Daher kann hier keine uneingeschränkte Empfehlung für deren Einsatz gegeben werden. Von den getesteten Verfahren eignet sich für eine präzise und vollständige Erkennung von Personennamen das *NER*-Verfahren von *FLAIR* am besten. Es extrahiert 99% der relevanten Akteure und liefert wenig irrelevante und kaum unvollständige Ergebnisse (*Precision* = 0,94 und *Recall* = 0,99).
Wenn für eine kommunikationswissenschaftliche Analyse die Identifikation von Personen in der Berichterstattung im Fokus liegt, weil im Anschluss beispielsweise ermittelt werden soll, ob es sich um Wissenschaftler oder Fachexperten handelt, kann diese Bibliothek zweifelsfrei zur Extraktion der individuellen Akteure empfohlen werden.
Wenn hingegen die Identifikation von institutionellen Akteuren Schwerpunkt der kommunikationswissenschaftlichen Untersuchung ist, empfiehlt sich ergänzend mit dem schnellen Verfahren von *spaCy* zu arbeiten. Hier müssen zwar weitaus mehr Begriffe gesichtet und bereinigt werden, doch können dabei voraussichtlich vollständigere Ergebnisse bei der Extraktion von Organisationsnamen erhalten werden.
Je nach Untersuchungsgegenstand und Forschungsfrage muss daher abgewogen werden, welches *NER*-Verfahren genutzt wird. Die hier vorgestellten Stärken und Schwächen der jeweiligen Bibliothek sollen bei dieser Entscheidung unterstützen.

7.2 Replikation manueller Codierungen durch die *NER*-Verfahren

Der durchgeführte Vergleich zwischen den händisch extrahierten Akteuren und den maschinell erhaltenen Ergebnissen zeigt, dass mit einer Trefferquote von 83-88% automatisiert dieselben Akteure im Datensatz ermittelt werden können, die manuell erhoben wurden (s. Tab. 22).

Dabei verspricht der Einsatz eines *NER*-Verfahrens eine grundsätzlich schnellere Bearbeitung und Ausgabe der Ergebnisse als eine manuelle Analyse des Datensatzes.

Ein expliziter Vergleich zwischen der benötigten Bearbeitungszeit beider Erhebungsmethoden ist hier nicht möglich, da nicht dieselben inhaltsanalytischen Aufgaben durchgeführt wurden. Dennoch ist ein deutlicher Unterschied nachweisbar. Im Austausch mit den Codierern, die den ‚Corona'-Datensatz manuell bearbeiteten, wurde die Einschätzung erhalten, dass pro Artikel im Durchschnitt 15 bis 20 Minuten notwendig waren, um die geforderten Kategorien der manuellen Inhaltsanalyse zu erheben. Insgesamt wurde ein Gesamtzeitraum von zweieinhalb Monaten benötigt, um drei Codierer zu schulen, eine ausreichende Intercoderreliabilität zu gewährleisten und den Datensatz zu bearbeiten.
Der Einsatz der automatisierten Verfahren kann je nach vorhandenen Vorkenntnissen beim Aufsetzen einer *Pipeline* und dem Bereinigen der Daten einige Tage Zeit in Anspruch nehmen. Bei dem automatisierten Erhalt der Eigennamen muss damit gerechnet werden, dass dabei sehr umfangreiche Ergebnisse mit verschiedenen Namensvarianten, Abkürzungen und Deklinationen extrahiert werden, die für eine bessere Übersichtlichkeit manuell vereinheitlicht und zusammengefasst werden müssen. Die eigentliche *NER*-Analyse läuft jedoch innerhalb von wenigen Stunden oder gar Minuten ab (s. Kap. 6.2.1).
Auch wenn der hier dokumentierte Einrichtungs- und Auswertungsaufwand groß war, lassen sich nach einmaliger Einrichtung der *Pipelines* künftig gleichartige Analysen in kürzester Zeit durchführen. Bei einer Erweiterung der Stichprobe, um beispielsweise 200 weiteren Nachrichtenartikeln, würde der Unterschied in der Verarbeitungsdauer bei den automatisierten *NER*-Verfahren nur wenige Minuten betragen, während die menschlichen Codierer mehrere zusätzliche Arbeitstage dafür benötigen würden.

Bei einer manuellen Erhebung kann allerdings weitaus flexibler und leichter konkretisiert werden, welche Akteure innerhalb der journalistischen Texte von Interesse sind. So wurde bei der hier thematisierten manuellen Inhaltsanalyse die Funktion eines Handlungsträgers berücksichtigt (vgl. Rössler 2017: 141) und ein Akteur nur dann codiert, wenn er sich zu der untersuchten Thematik in dem Nachrichtenartikel äußerte. Mit einem *NER*-Verfahren ist die Akteursextraktion nach solchen inhaltlichen Kriterien derzeit nicht möglich. Hier könnten maximal die formalen Kriterien, nach denen codiert werden soll, angepasst werden und zum Beispiel die drei zuerst genannten Akteure extrahiert werden (vgl. ebd.).
Inhaltlich können die durchgeführten *NER*-Analysen daher die manuelle inhaltsanalytische Erhebung des Lehrstuhls nicht ersetzen, da sie nicht im Stande sind, die dort getätigten Codierungen spiegelbildlich zu replizieren. Mit den extrahierten Eigennamen ist keine Aussage dazu möglich, ob es sich um Akteure handelt, die eine Aussage zum Untersuchungsgegenstand tätigen oder ob sie selbst Thema der Berichterstattung sind.

Nichtsdestotrotz können mit den erhaltenen Ergebnissen der *NER*-Analyse andere Auswertungen durchgeführt werden, die wertvolle Erkenntnisse über den untersuchten Datensatz ermöglichen. Basierend auf den in Kapitel 2.4 vorgestellten kommunikationswissenschaftlichen Untersuchungen werden im Anhang (s. [26], S. 111) drei verschiedene Beispiele aufgezeigt, wie mit den automatisiert erhaltenen Daten gearbeitet werden kann. Es kann das Vorkommen bestimmter Akteure im Zeitverlauf abgebildet sowie eine Analyse der Kookkurrenz von Akteuren in den Nachrichtenartikeln durchgeführt werden. Wäre die Artikelanzahl pro Medientitel der Stichprobe repräsentativ, ließen die Ergebnisse der *NER*-Analyse außerdem Aussagen über die Akteursvielfalt der Medientitel zu.
Wie eingangs in dieser Arbeit erläutert wurde, zeigt sich, dass automatisierte Methoden nicht zwingend manuelle Methoden ersetzen sollen, stattdessen eignen sie sich häufig gut, um die kommunikationswissenschaftliche Forschung zu ergänzen (s. Kap. 2.2).

In diesem Zusammenhang ist anzubringen, dass bei dem Vergleich der Erhebungsergebnisse in den manuellen Codierungen des ‚Corona'-Datensatzes und des Testdatensatzes vereinzelt Rechtschreib- und Codierungsfehler entdeckt wurden. Innerhalb der manuell codierten individuellen Akteure des ‚Corona'-Datensatzes werden 55 Fehler bei den Personennamen festgestellt (s. Anhang [25], S. 110). Bei der Anzahl an erhobenen Namen macht dies eine Fehlerquote von etwa 5% aus. Meist handelt sich es sich lediglich um fehlende Buchstaben oder Buchstabendreher in den Namen (‚Rosted' statt ‚Rorsted' oder ‚Berger' statt ‚Burger'), die für die manuellen Analysen irrelevant sind und nur den im Rahmen dieser Arbeit durchzuführenden Vergleich der Ergebnisse erschweren.
Doch es wurde auch auf Namen gestoßen, die manuell uneinheitlich codiert wurden und unentdeckt eine Akteurszusammenfassung beeinträchtigen würden. Tippfehler in anderen Kategorien der Inhaltsanalyse können darüber hinaus in der Zuteilung einer inkorrekten Merkmalsausprägung resultieren (‚1' = ‚individueller Akteur' statt ‚3' = ‚generischer Akteur').
Dadurch, dass es sich bei den vorliegenden manuellen Auswertungen um Einzelcodierungen handelt und kein zweiter Codierer denselben Artikel codiert oder auf anderem Wege Unstimmigkeiten kontrolliert werden, bleiben mögliche Tippfehler unentdeckt. Dies zeigt, dass menschliche Codierer auch nach zufriedenstellenden Testcodierungen und Reliabilitätstests dennoch bei der Extraktion von Daten Fehler machen können.
Mit Hilfe eines Vergleichs von automatisiert extrahierten Namen könnten Unstimmigkeiten jedoch schnell erkannt und die manuellen Ergebnisse korrigiert werden. Der parallele Einsatz einer automatisierten Erhebung, kann somit auch als Abgleichs- und Kontrollinstrument genutzt werden.

8. Limitationen

Die durchgeführte Untersuchung zur Gütebeurteilung der *NER*-Verfahren beinhaltet einige Limitationen, welche die Aussagekraft der herausgearbeiteten Befunde einschränken.
Eine grundsätzliche Einschränkung bei dem Einsatz der *NER*-Verfahren ist die Verfügbarkeit maschinenlesbarer Daten. Nur digitale Texte, die problemlos in Klartext umgewandelt werden können, sind mit der hier vorgestellten automatisierten Methode sinnvoll auswertbar. Es ist notwendig, dass ein Zugang zu Nachrichtenarchiven vorliegt, über den die digitalen Artikel mit Metadaten bezogen werden können, da diese wesentlich für die erfolgreiche Textverarbeitung sind.
Außerdem wurden bei der Vorbereitung und Durchführung der *NER*-Analyse zahlreiche Entscheidungen und Eingrenzungen getroffen, die großen Einfluss auf die letztlich erhaltenen Untersuchungsergebnisse hatten. Es musste abgewogen werden, welche *Preprocessing*-Schritte für den Erhalt der Eigennamen durchzuführen und in welchem Detailgrad die extrahierten Daten zu bereinigen und mit den manuellen Daten zu vergleichen sind (s. Kap. 5.2 und 6.2).
Zur Beurteilung der ausgegebenen Ergebnisse der *NER*-Analysen wurde eigenständig entschieden und argumentiert, welche Ergebnisse als richtig und falsch gewertet wurden. Diese Evaluation der Ergebnisse kann jedoch, je nach Auffassung des Betrachters, anders ausfallen und zu abweichenden Berechnungen der Leistungskennzahlen führen.
Um die Befunde der Untersuchung nachvollziehbar zu gestalten, wurden die getätigten Arbeitsschritte detailliert wiedergegeben. Die erhaltenen Ergebnisse und gewonnenen Erkenntnisse sind dabei maßgeblich von den zugrundeliegenden Textdaten der verfügbaren Stichprobe geprägt. Bei der Anwendung der hier durchgeführten Arbeitsschritte auf einen anderen Datensatz kann daher der Erhalt vergleichbarer Resultate nicht gewährleistet werden.

Es muss auch kritisch gewürdigt werden, dass der durchgeführte Vergleich nicht die ideale Bewertungsgrundlage der Ergebnisse der *NER*-Verfahren darstellt. Im Idealfall wären auch die manuellen Codierer angehalten gewesen, alle Akteure des Datensatzes zu extrahieren, um einen exakten Abgleich durchführen zu können. Anderseits zeigt der hier durchgeführte Vergleich dadurch auf, was die jeweiligen Stärken und Schwächen beider Erhebungsmethoden sind. Die manuelle Inhaltsanalyse eignet sich für eine Untersuchung nach inhaltlichen Kriterien, während mit den automatisierten Verfahren der Datensatz effizient nach formalen Kriterien analysiert werden kann. Die Methoden können ergänzend eingesetzt oder für die Untersuchung und Beantwortung unterschiedlicher Forschungsfragen genutzt werden.

Da es sich bei dem Einsatz von automatisierten Verfahren für kommunikationswissenschaftliche Zwecke um eine interdisziplinäre, recht junge Forschungsdisziplin handelt, existieren wenig deutschsprachige Ausarbeitungen und kaum Grundlagenliteratur. Für die Ermittlung des Forschungsstands und den Überblick über die gegenwärtigen Evaluationsmaßnahmen musste hauptsächlich auf Konferenzpapiere, Dissertationen, Fachartikel und Zeitschriftenaufsätze zurückgegriffen werden. Dementsprechend konnte weder ein übergreifender Konsens bei den genutzten Begrifflichkeiten noch ein Qualitätsstandard bei der Anwendung von automatisierten Methoden festgestellt werden. Es bestätigt sich, dass in diesem Fachbereich noch reichlich Aufklärungs- und Standardisierungsbedarf besteht. Neben der notwendigen interdisziplinären Expertise, erschweren die mangelnden Anwendungsrichtlinien sowie fehlenden Ausarbeitungen zur Gütebeurteilung solcher Verfahren, ihre Integration in den kommunikationswissenschaftlichen Forschungsalltag (vgl. Boumans/Trilling 2016: 18).

9. Fazit

Ziel dieser Arbeit war die Anwendung und Validierung eines automatisierten Verfahrens zur Identifikation von Akteuren in journalistischen Texten. Hierfür wurde das Verfahren der *Named Entity Recognition* eingesetzt, welches entwickelt wurde, um Eigennamen in Texten maschinell zu erfassen und auszugeben.

Zunächst wurde ein Überblick darüber gegeben, welche Verfahrensarten in der Kommunikationswissenschaft grundsätzlich für die automatisierte Extraktion von Inhalten aus Textdaten existieren. Daraufhin konnte eingeordnet werden, dass es sich bei *NER* um ein überwachtes, trainiertes Verfahren handelt, welches einen Teilbereich einer Inhaltsanalyse automatisieren kann. Es basiert auf *ML*-Algorithmen, welche mittels speziellen Trainingsdaten erlernen, Personen-, Orts- und Organisationsnamen in Texten zu identifizieren und zu klassifizieren.
Um nachvollziehen zu können, wie diese automatisierte, inhaltsanalytische Aufgabe funktioniert, wurden die Grundlagen der maschinellen Verarbeitung natürlicher Sprache *(NLP)* veranschaulicht. Dabei wurden die Herausforderungen aufgrund der Komplexität menschlicher Sprache herausgearbeitet, ebenso wie die verschiedenen Faktoren, die Einfluss auf die Ergebnisqualität eines *NER*-Verfahrens haben können.
Im Anschluss wurden drei unterschiedliche *NER*-Verfahren ausgewählt, deren Eignung zur Identifikation von Akteuren innerhalb eines Datensatzes bestehend aus 887 Nachrichtenartikeln geprüft wurde. Dafür mussten die Textdaten vorab in maschinenlesbares Material umgewandelt und der Programmiercode zu deren maschinellen Verarbeitung verfasst werden. In diesem Code wurde definiert, welche Bestandteile der Textdaten analysiert und in welcher Form die Daten ausgegeben werden sollten (s. Anhang [A]).

Die daraus erhaltenen Ergebnisse wurden umfassend untersucht und gegenübergestellt, um die jeweiligen Stärken und Schwächen der drei getesteten Verfahren zu ermitteln. Trotz großer Unterschiede in der Verarbeitungszeit, der Menge an extrahierten Eigennamen und der Fehlerquoten erwiesen sich alle drei Verfahren als gut geeignet, um Personennamen aus den Textdaten zu extrahieren. Beurteilt wurde dies basierend auf einem Vergleich der automatisiert erhaltenen Eigennamen mit händisch erhobenen Akteuren einer manuellen Analyse desselben Datensatzes.
Die automatisiert identifizierten Eigennamen deckten 99% der individuellen Akteure ab, die händisch erhoben wurden. Von den manuell selektierten institutionellen Akteuren wurden dagegen nur zwischen 68% und 80% von den *NER*-Verfahren erkannt.

Für die Identifikation von Akteuren mit generischen Namen eignet sich keines der getesteten Verfahren, da es sich bei generischen Akteuren nicht um konkrete Eigennamen handelt, auf deren Erkennung die *NER*-Verfahren trainiert wurden.

Die berechneten Leistungskennzahlen zur Bewertung der Verfahren liegen in den Wertebereichen, die auch in anderen Publikationen beim Test von *NER*-Verfahren erzielt werden. In dieser Forschungsliteratur konnte vorab kein eindeutiger Spitzenreiter unter all den verfügbaren *NER*-Verfahren ausgemacht werden und die vorliegende Untersuchung bestätigt dies. Das Verfahren mit der Bestleistung bei der Identifikation von Personennamen, schneidet am schlechtesten bei der Erkennung von Organisationsnamen ab und umgekehrt.

Festzuhalten ist, dass die Nutzung solcher automatisierten Verfahren sich ideal eignet, um in wenigen Stunden den Großteil der vorkommenden Personen-, Organisations- und Ortsnamen in großen Datensätzen journalistischer Berichterstattung zu extrahieren. Die Relevanz der extrahierten Akteure hängt jedoch von dem konkreten kommunikationswissenschaftlichen Untersuchungsgegenstand ab.
Die Verfahren sind sehr hilfreich bei der Ermittlung der meistgenannten Personen und Institutionen innerhalb umfangreicher Textmengen und dabei weitaus schneller als Menschen. Ihr Einsatz lohnt sich umso mehr, je mehr Daten dabei untersucht werden, da hier ihr zeitlicher und ökonomischer Effizienzvorteil genutzt werden kann (s. Kap. 2.2).

Nicht zu verachten ist beim Einsatz solch einer automatisierten Methode jedoch der Zeitaufwand, der für die Datenaufbereitung vorab und die Bereinigung der erhaltenen Ergebnisse im Anschluss notwendig ist. Diese Anwendungsschritte stellten den größten Arbeitsaufwand in dem hier durchgeführten Erhebungsprozess dar.
Zukünftige Ausarbeitungen, die sich mit dem Einsatz von *NER*-Verfahren befassen, könnten den Fokus insbesondere auf eine integrierte, automatisierte Bereinigung von abgekürzten und deklinierten Akteursnamen innerhalb eines Datensatzes legen. Dadurch würden die nachträglich anfallenden umfangreichen Arbeitsschritte zur Vereinheitlichung der Namensvarianten eingespart und mögliche Fehlerquellen minimiert werden.
Auch der Anwendungsbereich des *Named Entity Linking* (s. Kap. 2.4) verspricht Potenzial für die Erweiterung von *NER*-Analysen. Die automatisierte Extraktion zusätzlicher Informationen über die identifizierten Eigennamen kann dabei weitere manuelle Rechercheschritte ersetzen.

Die in dieser Arbeit durchgeführte Untersuchung eines automatisierten Verfahrens, illustrierte die Herausforderungen bei dessen Anwendung, ebenso wie bei der Beurteilung von dessen Güte. Da gegenwärtig kaum Orientierungsmaßstäbe und standardisierte Vorgehensweisen zur Anwendung automatisierter Verfahren in der Fachliteratur vorliegen, kann die Masterarbeit an dieser Stelle mit den herausgearbeiteten Befunden und gewonnen Erkenntnissen einen Beitrag leisten.

Mit dem Wissen und der Dokumentation aus der vorliegenden Arbeit können Analysen in Zukunft idealerweise schneller und effizienter vollzogen und einige der vorgefundenen Hürden in der Datenaufbereitung und Nachbereinigung umgangen werden. Die hier zur Verfügung gestellten Arbeitsmaterialien können dazu als Einstieg genutzt werden. Ebenso kann die durchgeführte Gütebeurteilung bei der Entscheidung unterstützen, welches *NER*-Verfahren sich für die zu bearbeitende Aufgabe eignet, sodass diese automatisierte Methode in künftigen kommunikationswissenschaftlichen Analysen angewandt und im Forschungsalltag integriert wird.

Literaturverzeichnis

Akbik, Alan/Blythe, Duncan/Vollgraf, Roland (2018): Contextual String Embeddings for Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA: Association for Computational Linguistics, S. 1638-1649.

Akbik, Alan/Bergmann, Tanja/Blythe, Duncan/Rasul, Kashif/Schweter, Stefan/Vollgraf, Roland (2019): FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL, S. 54-59.

Augenstein, Isabelle/Derczynski, Leon/Bontcheva, Kalina (2017): Generalisation in named entity recognition: A quantitative analysis. In: Computer Speech & Language 44, S. 61-83.

Ausserhofer, Julian/Gutounig, Robert/Oppermann, Michael/Matiasek, Sarah/Goldgruber, Eva (2017): The datafication of data journalism scholarship: Focal points, methods, and research propositions for the investigation of data-intensive newswork. In: Journalism, Volume: 21 issue: 7, S. 950-973.

Benikova, Darina/Biemann, Chris/Reznicek, Marc (2014): NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, S. 2524-2531.

Beysolow, Taweh (2018): Applied Natural Language Processing with Python, Berkeley, CA APRESS.

Blei, David M. (2012): Probabilistic topic models. In: Communications of the ACM 55, S. 77-84.

Boberg, Svenja/Quandt, Thorsten/Schatto-Eckrodt, Tim/Frischlich, Lena (2020): Pandemic Populism: Facebook Pages of Alternative News Media and the Corona Crisis - A Computational Content Analysis.

Brosius, Hans-Bernd/Schwer, Katja (2008): Wer kommt in der wissenschaftlichen und in der öffentlichen Debatte zu Wort? In: Brosius, Hans-Bernd/Schwer, Katja (Hrsg.): Die Forschung über Mediengewalt, Reihe: Schriftenreihe der Landeszentrale für Medien und Kommunikation, Bd. 26, S.154-163.

Brosius, Hans-Bernd/Haas, Alexander/Koschel, Friederike (2016): Methoden der empirischen Kommunikationsforschung. Eine Einführung (Studienbücher zur Kommunikations- und Medienwissenschaft), Wiesbaden: Springer VS.

Boumans, Jelle./Trilling, Damian (2016): Taking Stock of the Toolkit. In: Digital Journalism 4, S. 8-23.

Burggraaff, Christiaan/Trilling, Damian (2020): Through a different gate: An automated content analysis of how online news and print news differ. In: Journalism 21, S. 112-129.

Burscher, Björn/Odijk, Daan/Vliegenthart, Rens/Rijke, Maarten de/Vreese, Claes H. de (2014): Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis. In: Communication Methods and Measures 8, S. 190-206.

Derczynski, Leon (2016): Complementarity, F-score, and NLP Evaluation. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), S. 261-266.

Didakowski, Jörg/Geyken, Alexander/Hanneforth, Thomas (2007): Eigennamenerkennung zwischen morphologischer Analyse und Part-of-Speech Tagging: ein automatentheoriebasierter Ansatz. In: Zeitschrift für Sprachwissenschaft 26, S. 157-186.

Domahidi, Emese/Yang, JungHwan/Niemann-Lenz, Julia/Reinecke, Leonard (2019): Computational Communication Science | Outlining the Way Ahead in Computational Communication Science: An Introduction to the IJoC Special Section on "Computational Methods for Communication Science: Toward a Strategic Roadmap". In: International Journal of Communication 13, S. 3876-3884.

Dumm, Sebastian/Niekler, Andreas (2014): Methoden und Gütekriterien. Computergestützte Diskurs- und Inhaltsanalysen zwischen Sozialwissenschaft und Automatischer Sprachverarbeitung. Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus. Discussion Paper Nr. 4, Helmut-Schmidt-Universität Hamburg (UniBw) und Universität Leipzig.

Eftimov, Tome/Korousić Seljak, Barbara/Korosec, Peter (2017): A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PLoS ONE 12(6), e0179488, verfügbar unter: https://doi.org/10.1371/journal.pone.0179488 (zuletzt abgerufen am: 02.10.2020).

Ehrmann, Maud/Romanello, Matteo/Bircher, Stefan/Clematide, Simon (2020): Introducing the CLEF 2020 HIPE Shared Task: Named Entity Recognition and Linking on Historical Newspapers. In: Jose J. et al. (Hrsg.): Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol. 12036. Springer, Cham.

Eisenegger, Mark/Oehmer, Franziska/Udris, Linards/Vogler, Daniel (2020): Die Qualität der Medienberichterstattung zur Corona-Pandemie. Qualität der Medien Studie 1/2020, Forschungszentrum Öffentlichkeit und Gesellschaft (fög) der Universität Zürich.

Evans, Michael/McIntosh, Wayne/Lin, Jimmy/Cates, Cynthia (2006): Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research. Journal of Empirical Legal Studies, Volume 4, Issue 4, S. 1007–1039.

Fleischhauer, Monika/ Enge, Sören/ Donsbach, Wolfgang (2008): Multimodale Konstruktvalidierung: Ein Mehrmethodenansatz am Beispiel des Persönlichkeitsmerkmals ›Need for Cognition‹. In: Matthes, Jörg/ Wirth, Werner/Daschmann, Gregor/ Fahr, Andreas (Hrsg.): Die Brücke zwischen Theorie und Empirie. Operationalisierung, Messung und Validierung in der Kommunikationswissenschaft. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 3, Köln: von Halem, S. 326-354.

Früh, Werner/Früh, Hannah (2015): Empirische Methoden in den Sozialwissenschaften und die Rolle der Inhaltsanalyse. Eine Analyse deutscher und internationaler Fachzeitschriften 2000 bis 2009. In: Wirth, Werner/ Sommer, Katharina/Wettstein, Martin/Matthes, Jörg (Hrsg.): Qualitätskriterien in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 12, Köln: Herbert von Halem, S. 15-77.

Fu, King-Wa/Liang, Hai/Saroha, Nitin/Tse, Zion T. H./Ip, Patrick/Fung, Isaac C.-H. (2016): How people react to Zika virus outbreaks on Twitter? A computational content analysis. In: American journal of infection control 44, S. 1700-1702.

Gaus, Alex (2018): 8 lessons learned about NER. In: Medium - dpa-newslab vom 06.02.2018, https://medium.com/dpa-newslab/8-lessons-learned-about-ner-f40b263490db (zuletzt abgerufen 21.10.2020).

Gasser, Michael/Wanger, Regina/Prada, Ismail (2018): Wenn Algorithmen Zeitschriften lesen. Vom Mehrwert automatisierter Textanreicherung. O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB, 5(4), 181-192.

Gilch, Alexander/Schüler, Theresa (2019): Daten In: Kristian Kersting/Christoph Lampert/Constantin Rothkopf (Hg.), Wie Maschinen lernen. Künstliche Intelligenz verständlich erklärt, Wiesbaden: Springer Fachmedien Wiesbaden; Springer, S. 29-37.

Graaf, Rutger de/van der Vossen, Robert (2013): Bits versus brains in content analysis. Comparing the advantages and disadvantages of manual and automated methods for content analysis. In: Communications 38 (4), deGruyter Mouton, S. 433-443.

Graff, Frederik/Theobald, Elke (2010): User Generated Content als Gegenstand der Kommunikationsforschung – Der Einsatz von Webmonitoring in der Praxis am Beispiel der Bundestagswahl 2009. In: Jackob, Nikolaus/Zerback, Thomas/Jandura, Olaf et al. (Hrsg.): Das Internet als Forschungsinstrument und -gegenstand in der Kommunikationswissenschaft. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 6, Köln: von Halem, S. 194-210.

Grimmer, Justin/Stewart, Brandon M. (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. In: Political Analysis 21, Oxford University Press on behalf of the Society for Political Methodology, S. 267-297.

Günther, Elisabeth/Scharkow, Michael (2014): Automatisierte Datenbereinigung bei Inhalts- und Linkanalysen von Online-Nachrichten. In: Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 111-126.

Hepp, Andreas (2016): Kommunikations- und Medienwissenschaft in datengetriebenen Zeiten. In: Publizistik 61, S. 225-246.

Heuss, Timm/Humm, Bernhard/Henninger, Christian/Rippl, Thomas (2014): A comparison of NER tools w.r.t. a domain-specific vocabular. In: Sack, Harald (Hg.): SEMANTiCS Leipzig 2014, Tranfer/Engineering/Community proceedings of the 10th International Conference on Semantic Systems, Leipzig, New York: Association for Computing Machinery, S. 100-107.

Hirschmann, Hagen (2019): Korpuslinguistik. Eine Einführung, J. B: Metzler Verlag, Berlin.

Honnibal, Mathew/Montani, Ines (2017): spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, https://spacy.io/models/de.

Jiang, Ridong/Banchs, Rafael E./Li, Haizhou (2016): Evaluating and Combining Name Entity Recognition Systems. In: Association for Computational Linguistics, Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, Berlin, Germany, S. 21–27.

Kang, Ning/van Mulligen, Erik M./Kors, Jan A. (2012): Training text chunkers on a silver standard corpus: can silver replace gold? In: BMC bioinformatics 13, S. 17-23.

Kelm, Ole; Gerl, Katharina; Meißner, Florian (2020): Machine Learning. In: Borucki, Isabelle/Kleinen-von Königslöw, Katharina/Marschall, Stefan/Zerback, Thomas. (Hg.): Handbuch Politische Kommunikation. Wiesbaden: Springer Fachmedien Wiesbaden.

Kepplinger, Hans Matthias (1989): Instrumentelle Aktualisierung. In: Kaase M., Schulz W. (Hrsg.): Massenkommunikation. Kölner Zeitschrift für Soziologie und Sozialpsychologie Sonderhefte, vol. 30. VS Verlag für Sozialwissenschaften, Wiesbaden.

Ketschik, Nora/Overbeck, Maximilian/Murr, Sandra/Pichler, Axel/Blessing, André (2020): Interdisziplinäre Annotation von Entitätenreferenzen. Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung. In: Reiter, Nils, Axel Pichler, and Jonas Kuhn (Hrsg.): Reflektierte algorithmische Textanalyse. Berlin, Boston: De Gruyter, S. 203–236.

Klimek, Sonja/Müller, Ralph (2015): Vergleich als Methode? Zur Empirisierung eines philologischen Verfahrens im Zeitalter der Digital Humanities, *Journal of Literary Theory*, *9*(1), 53-78.

Kolb, Steffen (2005): Mediale Thematisierung in Zyklen. Theoretischer Entwurf und empirische Anwendung, Köln, Halem Verlag.

Kossen, Jannik/Müller, Maike E. (2019): Verzerrung-Varianz-Dilemma. In: Kristian Kersting/Christoph Lampert/Constantin Rothkopf (Hg.): Wie Maschinen lernen. Künstliche Intelligenz verständlich erklärt, Wiesbaden: Springer Fachmedien, S. 119-123.

Kovalchuk, Pavlo/Proen¸ Diogo/Borbinha, José/Henriques, Rui (2019): An Unsupervised Method for Concept Association Analysis in Text Collections. In: Doucet, Antoine/Isaac, Antoine/Golub, Koralja/Aalberg, Trond/Jatowt, Adam (Hrsg.): Proceedings - Digital Libraries for Open Knowledge, 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019 Oslo.

Kromidas, Stavros (2011): Handbuch Validierung in der Analytik. 2. Auflage, Weinheim, Wiley-VCH.

Lewis, Seth C./Zamith, Rodrigo/Hermida, Alfred (2013): Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. In: Journal of Broadcasting & Electronic Media, 57(1), 34–52.

Li, Jing/Sun, Aixin/Han, Jianglei/Li, Chenliang (2020): A Survey on Deep Learning for Named Entity Recognition. In: IEEE transactions on knowledge and data engineering, verfügbar unter: https://arxiv.org/pdf/1812.09449.

Marrero, Mónica/Urbano, Julián/Sánchez-Cuadrado, Sonia/Morato, Jorge/Gómez-Berbís, Juan Miguel (2013): Named Entity Recognition: Fallacies, challenges and opportunities. In: Computer Standards & Interfaces 35 (5), S. 482–489.

Maier, Daniel/Waldherr, Annie/Miltner, Peter/Schmid-Petri, Hannah/Häussler, Thomas/Adam, Silke (2014): Stichprobenziehung aus dem Netz – Wie man themenspezifische Online-Inhalte erfassen kann. In: Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 90-110.

Maier, Michaela/Retzbach, Joachim/Glogger, Isabella/Stengel, Karin (2018): Nachrichtenwerttheorie, Baden-Baden: Nomos.

Matthes, Jörg (2008): Medien-Frames inhaltsanalytisch (be)greifen. Eine Analyse von 135 nationalen und internationalen Fachzeitschriftenaufsätzen, 1990-2005. In: In: Matthes, Jörg/ Wirth, Werner/Daschmann, Gregor/ Fahr, Andreas (Hrsg.): Die Brücke zwischen Theorie und Empirie. Operationalisierung, Messung und Validierung in der Kommunikationswissenschaft. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 3, Köln: von Halem, S. 157-177.

Matthes, Jörg/Kohring, Matthias (2008): The Content Analysis of Media Frames: Toward Improving Reliability and Validity. In: Journal of Communication 58, S. 258-279.

Maurer, Marcus/Daxenberger, Johannes /Gurevych, Iryna (2017): Argumentation Mining: Eine neue Methode zur automatisierten Textanalyse und ihre Anwendung in der Kommunikationswissenschaft. In: Jahrestagung der Fachgruppe Methoden der Publizistik- und Kommunikationswissenschaft der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft, Mainz, Konferenzveröffentlichung.

Maynard, Diana/Bontcheva, Kalina/Augenstein, Isabelle (2016): Natural Language Processing for the Semantic Web. In: Synthesis Lectures on the Semantic Web: Theory and Technology 6, S. 1-194.

Naab, Teresa/Sehl, Annika (2014): Inhaltsanalytische Untersuchungen von User-Generated-Content-Angeboten: Eine Bestandsaufnahme zur Anwendung der Methode. In: Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 127-144.

Nadeau, David/Sekine, Satoshi (2007): A survey of named entity recognition and classification. In: Linguisticae Investigationes 30(1): S. 3–26.

Niekler, Andreas (2018): Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 13 Köln: Herbert von Halem Verlag.

Niemann-Lenz, Julia/Bruns, Sophie/Hefner, Dorothée/Knop-Hülß, Katharina/Possler, Daniel/Reich, Sabine/Reinecke, Leonard/Scheper, Jule/Klimmt, Christoph (2019): Computational Communication Science | Crafting a Strategic Roadmap for Computational Methods in Communication Science: Learnings From the CCS 2018 Conference in Hanover – Commentary. In: International Journal of Communication 13, S. 3885–3893.

Nordbeck, Ralf (2013): Die vergleichende Methode als Forschungsansatz. In: Internationaler Politiktransfer und nationaler Politikwandel. Springer VS, Wiesbaden.

Nunez-Mir, Gabriela/Iannone, Basil/Pijanowski, Bryan/Kong, Ningning/Fei, Songlin(2016): Automated content analysis: addressing the big literature challenge in ecology and evolution. In: Methods in Ecology and Evolution 7, S. 1262-1272.

Pinto, Alexandre/Gonçalo Oliveira, Hugo/Oliveira Alves, Ana (2016): Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text. In: Mernik, Marjan/Leal, José Paulo/Oliveira, Hugo Gonçalo: 5th Symposium on Languages, Applications and Technologies (SLATE'16), Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany, Article No. 3, S. 3-16.

Poesio, Massimo/Pradhan, Sameer/Recasens, Marta/Rodriguez, Kepa/Versley, Yannick (2016): Annotated Corpora and Annotation Tools. In: Massimo Poesio/Roland Stuckardt/Yannick Versley (Hrsg.): Anaphora Resolution, Berlin, Heidelberg: Springer, S. 97-140.

Patel, Jay M. (2020): Text preprocessing using scikit-learn and spaCy. In: Towards Data Science vom 06.03.2020, https://towardsdatascience.com/setting-up-text-preprocessing-pipeline-using-scikit-learn-and-spacy-e09b9b76758f (01.08.2020).

Qi, Peng/Zhang, Yuhao/Zhang, Yuhui/Bolton, Jason/Manning, Christopher D. (2020): Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: Association for Computational Linguistics (ACL) System Demonstrations, https://arxiv.org/abs/2003.07082.

Rössler, Patrick/Geise, Stephanie (2013): Standardisierte Inhaltsanalyse: Grundprinzipien, Einsatz und Anwendung. In: Medien & Kommunikationswissenschaft 64, S. 244-269.

Rössler, Marc (2007): Korpus-adaptive Eigennamenerkennung. Dissertation, Universität Duisburg-Essen, Winterthur. Abrufbar unter: https://duepublico2.uni-due.de/receive/duepublico_mods_00014746

Rössler, Patrick (2017): Inhaltsanalyse. utb. basics, Band 2671, Konstanz, München: UVK Verlagsgesellschaft mbH.

Rudkowsky, Elena/Haselmayer, Martin/Wastian, Matthias/Jenny, Marcelo/Emrich, Štefan/Sedlmair, Michael (2018): More than Bags of Words: Sentiment Analysis with Word Embeddings. In: Communication Methods and Measures 12, S. 140-157.

Sang, Erik F. Tjong Kim/Meulder, Fien de (2003): Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of CoNLL-2003, Edmonton, Canada, S. 142-147.

Severance, Charles (2015): Guido van Rossum: The Early Years of Python. In: Computing Conversations 0018- 9 162, IEEE Computer Society, S. 7-9.

Scharkow, Michael (2011): Zur Verknüpfung manueller und automatischer Inhaltsanalyse durch maschinelles Lernen. Reihe „Methodeninnovationen in der Kommunikationswissenschaft", in: Medien & Kommunikationswissenschaft 4/2011, S. 545-560.

Scharkow, Michael (2012): Automatische Inhaltsanalyse und maschinelles Lernen. Dissertation, Universität der Künste Berlin, Berlin: epubli.

Scharkow, Michael (2013). Automatische Inhaltsanalyse. In Möhring, Wiebke/Schlütz, Daniela (Hrsg.), Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft, Springer VS, Wiesbaden, S. 289–306.

Scharkow, Michael (2019): Buchbesprechung - Niekler, Andreas: Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen. Publizistik 64, 279–281.

Schneider, Gerold/Zimmermann, Heinrich (2010): Text-Mining-Methoden im Semantic Web. In: HMD Praxis der Wirtschaftsinformatik 47, S. 35-46.

Schneider, Gerold (2014): Automated Media Content Analysis from the Perspective of Computational Linguistics. In: Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 40-54.

Schumacher, Mareike (2018): Named Entity Recognition (NER). In: forTEXT. Literatur digital erforschen. URL: https://fortext.net/routinen/methoden/named-entity-recognition-ner (letzter Zugriff: 15.10.2020)

Schweiger, Wolfgang (2017): Der (des)informierte Bürger im Netz, Wiesbaden: Springer Fachmedien Wiesbaden.

Schwotzer, Bertil (2014): Automatische Selektion von Beiträgen für themenspezifische Inhaltsanalysen mittels Schlagwortlisten. In: Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 55-72.

Shelar, Hemlata/Kaur, Gagandeep/Heda, Neha/Agrawal, Poorva (2020): Named Entity Recognition Approaches and Their Comparison for Custom NER Model, Science & Technology Libraries, 39 (3), S. 324-337.

Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Ein Vorwort - Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 9-16.

Song, Hye-Jeong/Jo, Byeong-Cheol/Park, Chan-Young/Kim, Jong-Dae/Kim, Yu-Seop (2018): Comparison of named entity recognition methodologies in biomedical documents. In: Biomedical engineering online 17, S. 21-34.

Srinivasa-Desikan, Bhargav (2018): Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, SpaCy and Keras. Birmingham: Packt Publishing Ltd.

Stoll, Anke/Ziegele, Marc/Quiring, Oliver (2020): Detecting Impoliteness and Incivility in Online Discussions Classification Approaches for German User Comments. In: Computational Communication Research 2, S. 109-134.

Strippel, Christian/Bock, Annekatrin/Katzenbach, Christian/Mahrt, Merja/Merten, Lisa/Nuernbergk, Christian/Pentzold, Christian/Puschmann, Cornelius/Waldherr, Annie (2018): Die Zukunft der Kommunikationswissenschaft ist schon da, sie ist nur ungleich verteilt. Eine Kollektivreplik auf Beiträge im „Forum“ - Publizistik, Heft 3 und 4, 2016. In: Publizistik 63, Springer Fachmedien Wiesbaden GmbH, S. 11-27.

Taylor, Ann/Marcus, Mitchell/Santorini Beatrice (2003): The Penn Treebank: An Overview. In: Abeillé Anne (Hrsg.): Treebanks - Building and Using Parsed Corpora. Text, Speech and Language Technology, vol 20, Springer, Dordrecht.

Trilling, Damian (2014): Weg vom manuellen Speichern: RSS -Feeds in der automatisierten Datenerhebung bei Online-Medien. In: Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 73-89.

Vogler, Daniel/Schäfer, Mike (2020): Growing Influence of University PR on Science News Coverage? A Longitudinal Automated Content Analysis of University Media Releases and Newspaper Coverage in Switzerland, 2003–2017. In: International Journal of Communication 14, S. 3143–3164.

Waldherr, Annie/Miltner, Peter/Ostner, Sophia/Stoltenberg, Daniela/Pfetsch, Barbara/Wehden, Lars-Ole (2019): Induktive Kategorienbildung in der Inhaltsanalyse: Kombination automatischer und manueller Verfahren. Forum Qualitative Sozialforschung 20(1), 1-30.

Wettstein, Martin (2014): ›Best of both worlds‹: Die halbautomatische Inhaltsanalyse. In: Sommer, Katharina/Matthes, Jörg/Wettstein, Martin/Wirth, Werner: Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft, Band 11, Köln: Herbert von Halem, S. 16-39.

Wettstein, Martin (2016): Verfahren zur computerunterstützten Inhaltsanalyse in der Kommunikationswissenschaft. Dissertation, Universität Zürich, Institut für Kommunikationswissenschaft und Medienforschung, Zürich.

Wu, Yifan (2020): Is a Dataframe Just a Table? In: 10th Workshop on Evaluation and Usability of Programming Languages and Tools (PLATEAU 2019), Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Article No. 6, S. 1-10.

van Atteveldt, Wouter (2008): Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content. Charleston, S. C.: Book Surge.

van Atteveldt, Wouter/Margolin, Drew /Shen, Cuihua/Trilling, Damian/Weber, René (2019): A Roadmap for Computational Communication Research. In: Computational Communication Research 1, S. 1-5.

van der Meer, Toni G.L.A. (2016): Automated content analysis and crisis communication research. In: Public Relations Review 42, S. 952-961.

Yadav, Vikas/Bethard, Steven (2019): A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In: C18, Computation and Language, arXiv:1910.11470.

Vychegzhanin, Sergey/Kotelnikov, Evgeny (2019): Comparison of Named Entity Recognition Tools Applied to News Articles. In: Ivannikov Ispras Open Conference 2019 (ISPRAS), IEEE, S. 72-77.

Züll, Cornelia/Mohler, Peter (2001): Computerunterstützte Inhaltsanalyse: Codierung und Analyse von Antworten auf offene Fragen. Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim.

Genutzte bit.ly Links in Quellenangaben von Abbildungen:

Abb. 6	https://bit.ly/2S3nEgq	https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e
Abb. 7	https://bit.ly/3h3wRiP	https://developers.google.com/machine-learning/crash-course/embeddings/translating-to-a-lower-dimensional-space
Abb. 12	https://bit.ly/30sPyXO	https://www.teradata.de/Blogs/The-Tree-of-Machine-Learning-Algorithms
Abb. 21	http://bit.ly/3byRXXm	https://blog.codecentric.de/2019/07/machine-learning-modelle-bewerten-die-crux-mit-der-metrik/

Anhang

- Verweise auf andere Erhebungen mit *NER*
- Screenshots: Umwandlung der Medientitel in maschinenlesbare Formate
- Beispiele der Herausforderungen im *Preprocessing*
- *NER*-Ergebnisse - Tabellarische Auswertungen
- Screenshots: Linguistische Herausforderungen bei NER
- Auswertungen des Testdatensatzes (Ebola, AR, Grippe)

Separate Dokumente:

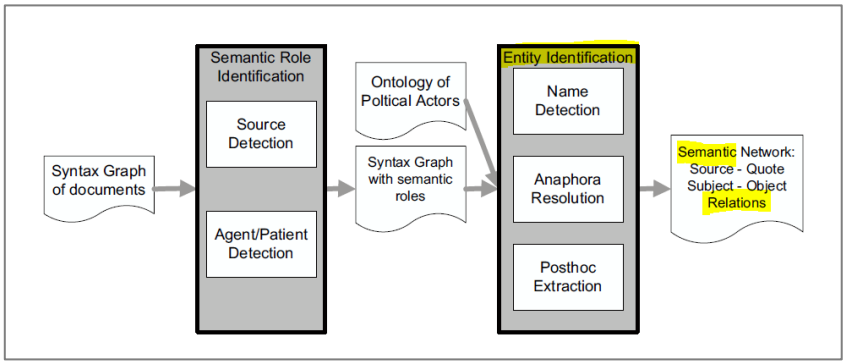
A) 3x Pipeline als HTML und Python-Code Dokument (*spacy, Stanza, FLAIR*)
B) Datensatz Corona kompakt als txt und pdf
C) Datensatz Ebola-AR-Grippe als separate txts

D) 3x Excel-Arbeitsmappen mit *NER*-Ergebnissen pro Bibliothek
E) Excel-Arbeitsmappe ‚NER_Vergleichsdaten.xlsx'
F) Excel-Arbeitsmappe ‚Gegenüberstellung_Bibliotheken.xlsx'
G) Excel-Arbeitsmappe ‚Codierungen_Artikelebene.xlsx'

H) Excel-Arbeitsmappe ‚Test_encoding_spaCy_small_large.xlsx'
I) Excel-Arbeitsmappe ‚Ebola_AR_Grippe_Ergebnisse.xlsx'
J) Excel-Arbeitsmappe ‚Test_Kleingeschriebener_Datensatz.xlsx'

Verweise auf andere Erhebungen mit NER

1) Komplexität einer semantischen Netzwerkanalyse mit Eigennamenerkennung

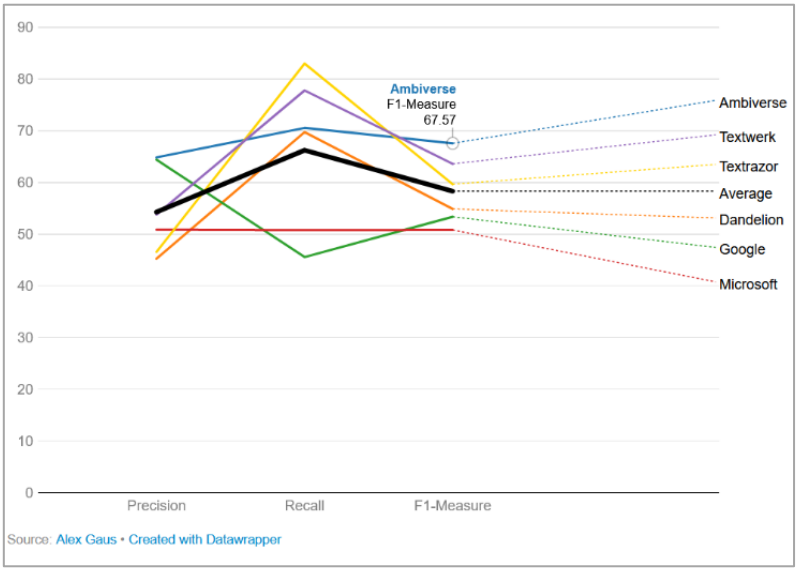


Visuelle Darstellung der benötigten Verarbeitungsschritte (Quelle: van Atteveldt 2008: 63)

2) F-Scores aus Toolanbieter-Vergleich

Das Tool mit den besten Ergebnissen kommt nur auf einen F-Score-Wert von 0,67.

Quelle: https://medium.com/dpa-newslab/8-lessons-learned-about-ner-f40b263490db (Gaus 2018)



3) F1-Scores der Vergleichsstudie von Qi et al.

FLAIR und Stanza weisen bessere F1-Maße als spaCy auf. (Quelle: Qi et al. 2020: 6)

Language	Corpus	# Types	Stanza	FLAIR	spaCy
Arabic	AQMAR	4	**74.3**	74.0	–
Chinese	OntoNotes	18	**79.2**	–	–
Dutch	CoNLL02	4	89.2	**90.3**	73.8
	WikiNER	4	**94.8**	**94.8**	90.9
English	CoNLL03	4	92.1	**92.7**	81.0
	OntoNotes	18	88.8	**89.0**	85.4*
French	WikiNER	4	**92.9**	92.5	88.8*
German	CoNLL03	4	81.9	**82.5**	63.9
	GermEval14	4	85.2	**85.4**	68.4
Russian	WikiNER	4	**92.9**	–	–
Spanish	CoNLL02	4	**88.1**	87.3	77.5
	AnCora	4	**88.6**	88.4	76.1

Table 3: NER performance across different languages and corpora. All scores reported are entity micro-averaged test F_1. For each corpus we also list the number of entity types. $*$ marks results from publicly available pretrained models on the same dataset, while others are from models retrained on our datasets.

4) Auszug des Codebuchs des Lehrstuhls - Wann wird ein Akteur codiert?

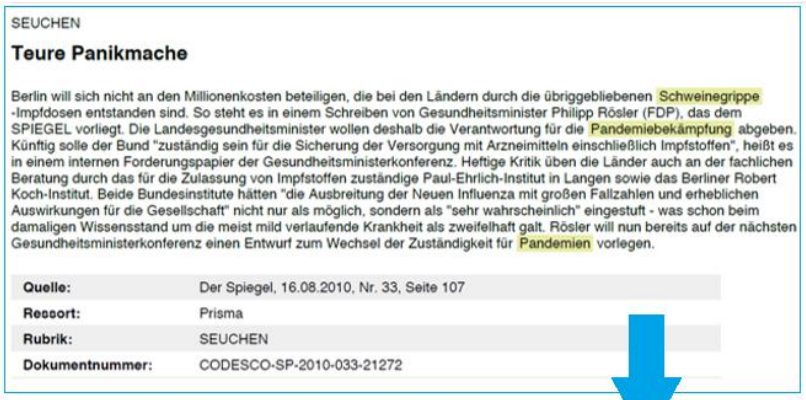
Definition Aussage (S.1):
„Das erste Kriterium für die Bestimmung einer Aussage ist die Präsenz eines Akteurs. Eine Aussage kann nur dann als Aussage codiert werden, wenn sie einem identifizierbaren Akteur (Person, Organisation oder identifizierbare Gemeinschaft wie z.B. „Experten“) zugeschrieben werden kann. Eine Aussage darf nur einem einzelnen Akteur zugerechnet werden. Sobald ein neuer Akteur zu Wort kommt, ist die Selektion als nächste Aussage zu codieren. Das zweite Kriterium für die Bestimmung einer Aussage ist ihr thematischer Bezug. Als Aussagen codiert werden nur solche Aussagen, in denen es um Corona, Antibiotika-Resistenz, Grippe bzw. Ebola geht.“
Spezifizierung des Akteurs (S.8 - AKTSPEZ##):
„Diese Variable beschreibt, wie spezifisch der Akteur, d.h. die Quelle einer Aussage, beschrieben wird. Ein Akteur ist jede namentlich genannte Person, Personengruppe, Organisation/Institution sowie allgemein definierte kollektive Sprecher, die in einem Artikel zu Wort kommen. Jede Aussage muss nur dann als solche codiert werden, wenn sie einem Akteur zurechenbar ist. Wir unterscheiden zwischen vier Kategorien: Individuell, Institutionell (Charité), Generisch (Wissenschaftler, Biologen), Sonstige
Der Akteur wird als ***individuell*** codiert, wenn eine konkrete Person genannt wird. In diesem Fall sind auch die folgende Informationen zu codieren: Name, professionelle Position (e.g. Hygieniker), akademischer Titel (e.g. Professor), professionale Zugehörigkeit (e.g. Uniklinik Mainz, RKI).
Institutionell ist dann zu codieren, wenn eine Organisation oder Institution als Quelle zitiert wird (z.B. WHO, Charité, Uniklinik, Handelskammer, Die Grünen etc.).
Die Kategorie **generisch** meint Akteure, die als umfassend-definierte Allgemeinheiten dargestellt werden. Falls Sammelbegriffe wie „Experten“, „Wissenschaftler“, „Biologen“, „Politiker“ usw. unspezifiziert als Quellen von Aussagen genannt werden, ist generisch zu codieren.
Die ersten drei Kategorien sind hierarchisch zu betrachten. D.h. wenn der gleiche Akteur namentlich genannt wird („Prof. Werner Schmidt“) und auch durch Organisation oder generisch beschrieben („Wissenschaftler der Uniklinik Mainz“, „Biologe“), ist er weiterhin als *individueller* Akteur zu codieren. In der gleichen Art, wenn es in einem Artikel um „Forscher der Institution XY“ geht, muss man den Akteur als institutionell codieren, wenn die Institution/Organisation identifizierbar ist.“

5) Fehlende Leerzeichen beeinträchtigen die Erkennung von Eigennamen

```
Ludwig PER
kannich PER
Ludwig PER
Arzneimittelkommission ORG
Ludwig PER
GlaxoSmithKline ORG
GSK ORG
Gutgekühlt ORG
Ständige Impfkommission ORG
Stiko ORG
Immunsystembesonders ORG
Klaus Theo Schröder PER
Schröder PER
GSK ORG
keineswegsals PER
mansich PER
Stiko-Mitglied Ulrich Heininger PER
Deutschen Gesellschaftfür ORG
Michael Kochen PER
Wolfram Hartmann PER
Bundesverband der Kinder- und Jugendärzte ORG
Hartmann PER
Undzusätzlich PER
Hartmann PER
```

Ergebnisse der Texte zu Grippepandemien

6) Markierungen in der Ursprungsdatei


SEUCHEN

Teure Panikmache

Berlin will sich nicht an den Millionenkosten beteiligen, die bei den Ländern durch die übriggebliebenen Schweinegrippe -Impfdosen entstanden sind. So steht es in einem Schreiben von Gesundheitsminister Philipp Rösler (FDP), das dem SPIEGEL vorliegt. Die Landesgesundheitsminister wollen deshalb die Verantwortung für die Pandemiebekämpfung abgeben. Künftig solle der Bund "zuständig sein für die Sicherung der Versorgung mit Arzneimitteln einschließlich Impfstoffen", heißt es in einem internen Forderungspapier der Gesundheitsministerkonferenz. Heftige Kritik üben die Länder auch an der fachlichen Beratung durch das für die Zulassung von Impfstoffen zuständige Paul-Ehrlich-Institut in Langen sowie das Berliner Robert Koch-Institut. Beide Bundesinstitute hätten "die Ausbreitung der Neuen Influenza mit großen Fallzahlen und erheblichen Auswirkungen für die Gesellschaft" nicht nur als möglich, sondern als "sehr wahrscheinlich" eingestuft - was schon beim damaligen Wissensstand um die meist mild verlaufende Krankheit als zweifelhaft galt. Rösler will nun bereits auf der nächsten Gesundheitsministerkonferenz einen Entwurf zum Wechsel der Zuständigkeit für Pandemien vorlegen.

Quelle:	Der Spiegel, 16.08.2010, Nr. 33, Seite 107
Ressort:	Prisma
Rubrik:	SEUCHEN
Dokumentnummer:	CODESCO-SP-2010-033-21272

```
Der Spiegel, 16.08.2010, Nr. 33, Seite 107 / Prisma

SEUCHEN

Teure Panikmache

Berlin will sich nicht an den Millionenkosten beteiligen, die bei den Ländern
durch die übriggebliebenen Schweinegrippe

-Impfdosen entstanden sind. So steht es in einem Schreiben von
Gesundheitsminister Philipp Rösler (FDP), das dem
SPIEGEL vorliegt. Die Landesgesundheitsminister wollen deshalb die
Verantwortung für die

Pandemiebekämpfung abgeben.
Künftig solle der Bund "zuständig sein für die Sicherung der Versorgung mit
Arzneimitteln einschließlich Impfstoffen", heißt es
```

Sogar Markierungen können nach der Umwandlung Einfluss auf die Textdarstellung haben →

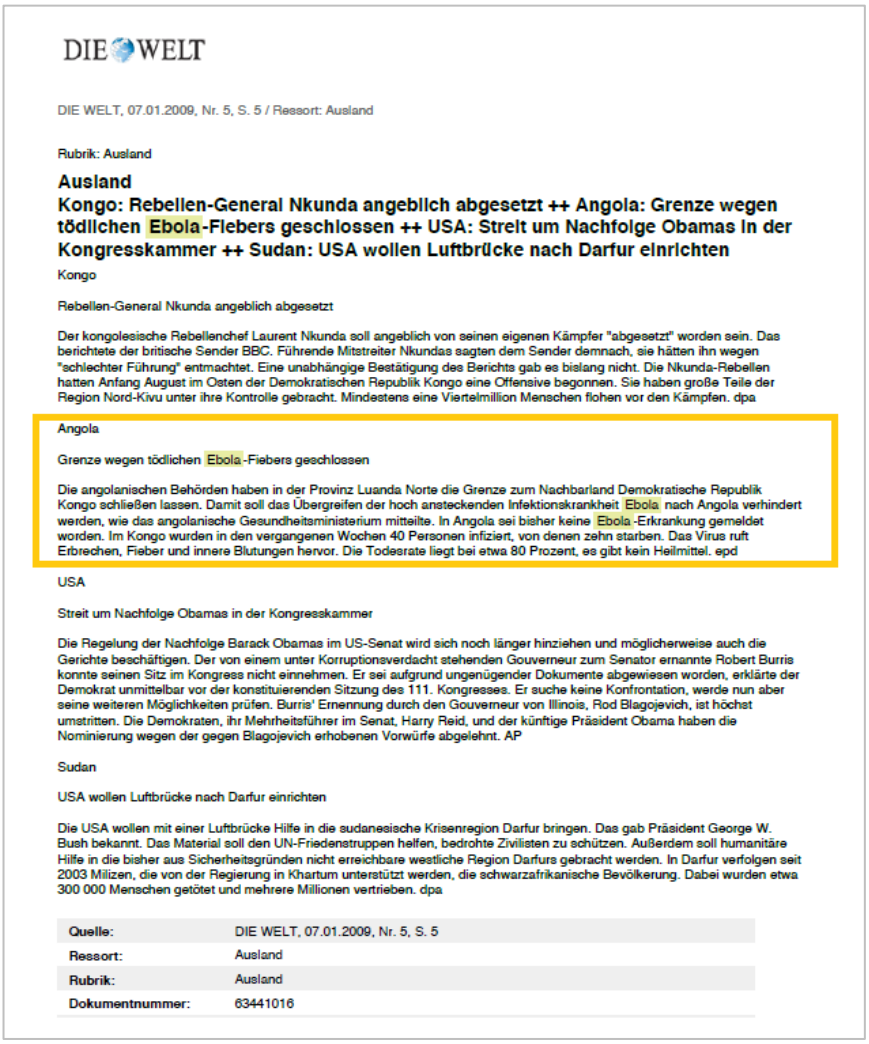
7) Layout der *SZ*-Artikel führt zu fehlerhaften txt-Umwandlung

```
Süddeutsche Zeitung

POLITIK
                                                    Donnerstag, 27. Februar  2020

Deutschland Seite 5, Bayern Seite 5

merhin den zweitgrößten Landesverband
stellt. Ein wenig können sich die Christde-
mokraten  im  Land  zwar  an  Wolfgang
Schäuble aufrichten, der als Bundestags-
präsident das zweithöchste Amt im Staat
bekleidet. Doch die Zeiten, als Volker Kau-
der  die  Bundestagsfraktion  führte,  sind
nur noch nostalgische Erinnerung. Und im
Kabinett findet sich kein einziger Minister
aus Baden-Württemberg.

Um den Weg aus der Krise zu finden, hat
sich  die  Landesspitze  überraschend  ent-
schlossen, am Morgen von Kramp-Karren-
bauers  Auftritt  noch  schnell  Friedrich
Merz zu ihrem Hoffnungsträger zu erklä-
ren.  Als  Kramp-Karrenbauer,  Merz  und
Spahn 2018 um den Vorsitz konkurrierten,
hatte sich der Landesvorsitzende Thomas
Strobl mit Aussagen zum eigenen Wahlver-
halten bewusst zurückgehalten. Mittwoch-
früh kündigt er bei einem Pressegespräch
gemeinsam mit Susanne Eisenmann, Spit-
zenkandidatin bei der Landtagswahl 2021,
und  Generalsekretär  Manuel  Hagel  an,
dass man Merz unterstützen und aktiv für
ihn  werben  werde.  Später  wiederholt
Strobl sein Bekenntnis vor 1500 Parteimit-
gliedern  in  der  Fellbacher  „Alten  Kelter"
und bekommt dafür reichlich Applaus. Für
Merz sprechen aus Strobls Sicht vor allem
zwei Gründe: „Er hat ein sehr, sehr hohes
Ansehen in der deutschen und insbesonde-
re in der baden-württembergischen Wirt-
schaft." Außerdem glaube er, „dass es mit
ihm am besten gelingen kann, Wähler und
Wählerinnen von der AfD zurückzugewin-
nen".  Er  habe  auch  den  Eindruck,  sagt
Strobl,  dass Friedrich Merz bei der  Basis
im Land die Nase vorn habe.

Eisenmann, die sich vor Kurzem für ei-

nen Generationenwechsel und damit indi-


„Der politische Gegner sitzt nicht in den eigenen Reihen": Kramp-Karrenbauer
nach ihrer Rede am Mittwoch in Fellbach bei Stuttgart.                FOTO:
MARIJAN MURAT, DPA

Bei den Freunden von Merz

Annegret Kramp-Karrenbauer, die noch an der Spitze der CDU steht, deutet beim
politischen Aschermittwoch
in Baden-Württemberg an, dass sie für Laschet und Spahn ist. Ihre Gastgeber
sind eindeutig anderer Meinung

rekt für Jens Spahn ausgesprochen hatte,
sagt  nun,  dass  Merz  am  besten  den  An-
spruch  verkörpere,  das  inhaltliche  Profil
der CDU zu schärfen. „Mit seiner klaren Po-
sitionierung, mit seiner klaren Erkennbar-
keit" sei er „der richtige Mann" und auch
kein Kandidat, der die Vergangenheit ver-
körpere.  Sie  traue  ihm „uneingeschränkt
zu", dass er seine Antennen auf die heutige
Gesellschaft einstelle. Vier seiner Themen
haben sie überzeugt: die wirtschaftspoliti-
sche Handschrift, „klare Kante bei der in-


von claudia henzler

Fellbach  –  Die  CDU  in  Baden-Württem-
berg musste sich kurz vor Aschermittwoch
ein bisschen Mitleid gefallen lassen. Da hat-
te sie sich rechtzeitig einen echten Promi
in Person der Parteivorsitzenden Annegret
Kramp-Karrenbauer  gesichert  –  und  hat
letztlich                doch die CDU-Rednerin zu bieten,
für die sich die Leute am wenigsten interes-
```


Süddeutsche Zeitung POLITIK Donnerstag, 27. Februar 2020
Deutschland Seite 5, Bayern Seite 5
#Fellbach20
Politischer Aschermittwoch
Bei den Freunden von Merz
Annegret Kramp-Karrenbauer, die noch an der Spitze der CDU steht, deutet beim politischen Aschermittwoch in Baden-Württemberg an, dass sie für Laschet und Spahn ist. Ihre Gastgeber sind eindeutig anderer Meinung

Die Textspalten werden in falscher Reihenfolge aneinandergereiht

8) Zu beachtende Eigenschaften der Artikel für deren automatisierte Extraktion



DIE WELT

DIE WELT, 07.01.2009, Nr. 5, S. 5 / Ressort: Ausland

Rubrik: Ausland

Ausland

Kongo: Rebellen-General Nkunda angeblich abgesetzt ++ Angola: Grenze wegen tödlichen Ebola-Fiebers geschlossen ++ USA: Streit um Nachfolge Obamas in der Kongresskammer ++ Sudan: USA wollen Luftbrücke nach Darfur einrichten

Kongo

Rebellen-General Nkunda angeblich abgesetzt

Der kongolesische Rebellenchef Laurent Nkunda soll angeblich von seinen eigenen Kämpfer "abgesetzt" worden sein. Das berichtete der britische Sender BBC. Führende Mitstreiter Nkundas sagten dem Sender demnach, sie hätten ihn wegen "schlechter Führung" entmachtet. Eine unabhängige Bestätigung des Berichts gab es bislang nicht. Die Nkunda-Rebellen hatten Anfang August im Osten der Demokratischen Republik Kongo eine Offensive begonnen. Sie haben große Teile der Region Nord-Kivu unter ihre Kontrolle gebracht. Mindestens eine Viertelmillion Menschen flohen vor den Kämpfen. dpa

Angola

Grenze wegen tödlichen Ebola-Fiebers geschlossen

Die angolanischen Behörden haben in der Provinz Luanda Norte die Grenze zum Nachbarland Demokratische Republik Kongo schließen lassen. Damit soll das Übergreifen der hoch ansteckenden Infektionskrankheit Ebola nach Angola verhindert werden, wie das angolanische Gesundheitsministerium mitteilte. In Angola sei bisher keine Ebola-Erkrankung gemeldet worden. Im Kongo wurden in den vergangenen Wochen 40 Personen infiziert, von denen zehn starben. Das Virus ruft Erbrechen, Fieber und innere Blutungen hervor. Die Todesrate liegt bei etwa 80 Prozent, es gibt kein Heilmittel. epd

USA

Streit um Nachfolge Obamas in der Kongresskammer

Die Regelung der Nachfolge Barack Obamas im US-Senat wird sich noch länger hinziehen und möglicherweise auch die Gerichte beschäftigen. Der von einem unter Korruptionsverdacht stehenden Gouverneur zum Senator ernannte Robert Burris konnte seinen Sitz im Kongress nicht einnehmen. Er sei aufgrund ungenügender Dokumente abgewiesen worden, erklärte der Demokrat unmittelbar vor der konstituierenden Sitzung des 111. Kongresses. Er suche keine Konfrontation, werde nun aber seine weiteren Möglichkeiten prüfen. Burris' Ernennung durch den Gouverneur von Illinois, Rod Blagojevich, ist höchst umstritten. Die Demokraten, ihr Mehrheitsführer im Senat, Harry Reid, und der künftige Präsident Obama haben die Nominierung wegen der gegen Blagojevich erhobenen Vorwürfe abgelehnt. AP

Sudan

USA wollen Luftbrücke nach Darfur einrichten

Die USA wollen mit einer Luftbrücke Hilfe in die sudanesische Krisenregion Darfur bringen. Das gab Präsident George W. Bush bekannt. Das Material soll den UN-Friedenstruppen helfen, bedrohte Zivilisten zu schützen. Außerdem soll humanitäre Hilfe in die bisher aus Sicherheitsgründen nicht erreichbare westliche Region Darfurs gebracht werden. In Darfur verfolgen seit 2003 Milizen, die von der Regierung in Khartum unterstützt werden, die schwarzafrikanische Bevölkerung. Dabei wurden etwa 300 000 Menschen getötet und mehrere Millionen vertrieben. dpa

Quelle:	DIE WELT, 07.01.2009, Nr. 5, S. 5
Ressort:	Ausland
Rubrik:	Ausland
Dokumentnummer:	63441016

Bei alten Artikeln des Testdatensatzes fehlten die Metadaten (Body, Load-Date) oder waren anders benannt. Dies muss in der *Pipeline* berücksichtigt werden.

Auch die saubere Extraktion der Überschriften war bei dieser Art von Artikeln erschwert und musste händisch bereinigt werden, um keine Probleme bei der Ausgabe des *dataframes* zu erhalten. Mehrzeilige Überschriften resultierten in der finalen Ausgabe-Datei in Textverschiebungen der zugehörigen Daten pro Zeile.

Auflistung der *WELT*-Artikel aus Datensatz die aufgrund andersartiger Formatierung im Titel sowie Fließtext zusätzlich manuell bereinigt wurden:

I. Finanzen Kompakt; Sentix-Index: Coronavirus-Krise verunsichert leicht ++ Kryptowährungen: Vom Staat nur für Interbankenhandel? ++ Blackrock: Aktivisten dringen in Pariser Zentrale ein

II. Finanzen Kompakt; Heizöl-Preis: Seit zwei Jahren wieder unter 60 Euro ++ EU-Energieverbrauch: Werte weiter deutlich über Ziel für 2020 ++ Passwörter: BSI rückt vom häufigen Wechsel ab ++ Lebensversicherung: Allianz zieht sich von Geschäft zurück

III. Finanzen Kompakt; Coronavirus: Kostenloses Umbuchen bei Lufthansa möglich ++ Arbeit im Homeoffice: Arbeitgeber kann sich an Kosten beteiligen ++ Kindergeld: Kinder müssen aussagen ++ iOS: Video-Gespräche lassen sich aufnehmen

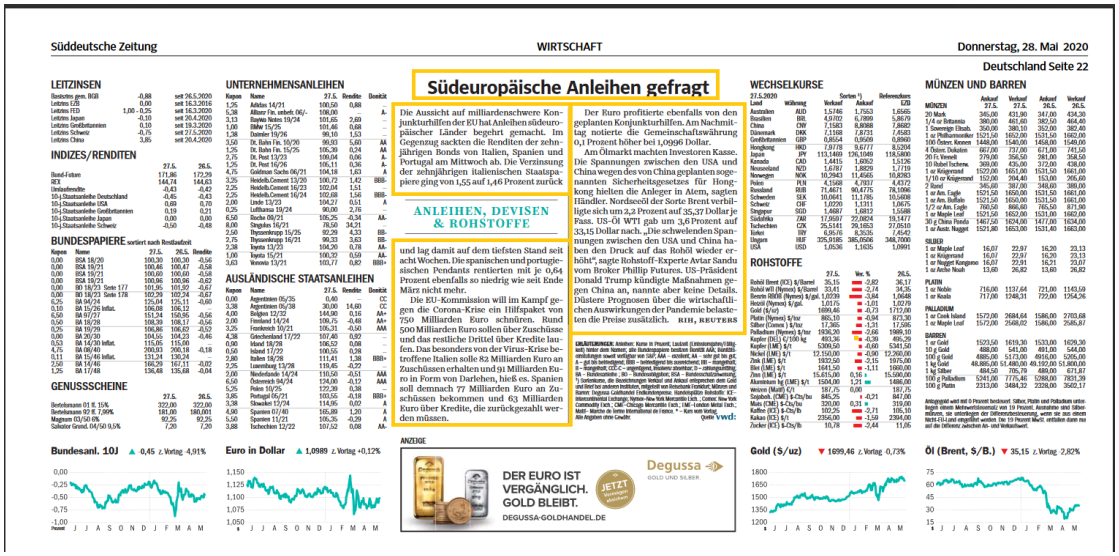
IV. Finanzen Kompakt; Olaf Scholz: Banken sollen Kredite leichter vergeben ++ Fielmann: Optikerkette setzt Dividende aus ++ Bundesbank: Staatsschulden um 16 Milliarden gesunken

V. Finanzen Kompakt; Aktienmarkt: Störung legt Xetra-Handel lahm ++ G7-Staaten: Schuldenstundung für arme Länder ++ Deutsche Bank: Große Andrang auf KfW-Schnellkredite ++ Pharmakonzern: Johnson & Johnson senkt Ausblick

VI. Wirtschaft Kompakt; Wettbewerbsverfahren: EU-Kommission geht gegen Apple vor ++ DER Touristik: Buchungszahlen steigen jede Woche ++ Easyjet: Airbus-Abnahmen um Jahre verschoben

VII. Wissen Kompakt; Covid-19: Malaria-Mittel sind wirkungslos ++ China: Erste Erfolge mit Corona-Impfstoff ++ Nordamerika: Schneemenge schrumpft stark

9) *SZ*-Artikel ohne Metadaten zur Definition des relevanten Textkörpers



Süddeutsche Zeitung — WIRTSCHAFT — Donnerstag, 28. Mai 2020 — Deutschland Seite 22

Südeuropäische Anleihen gefragt

Die Aussicht auf milliardenschwere Konjunkturhilfen der EU hat Anleihen südeuropäischer Länder begehrt gemacht. Im Gegenzug sackten die Renditen der zehnjährigen Bonds von Italien, Spanien und Portugal am Mittwoch ab. Die Verzinsung der zehnjährigen italienischen Staatspapiere ging von 1,55 auf 1,46 Prozent zurück

Der Euro profitierte ebenfalls von den geplanten Konjunkturhilfen. Am Nachmittag notierte die Gemeinschaftswährung 0,1 Prozent höher bei 1,0996 Dollar.

Am Ölmarkt machten Investoren Kasse. Die Spannungen zwischen den USA und China wegen des von China geplanten sogenannten Sicherheitsgesetzes für Hongkong hielten die Anleger in Atem, sagten Händler. Nordseeöl der Sorte Brent verbilligte sich um 2,2 Prozent auf 35,37 Dollar je Fass. US-Öl WTI gab um 3,6 Prozent auf 33,15 Dollar nach. „Die schwelenden Spannungen zwischen den USA und China haben den Druck auf das Rohöl wieder erhöht", sagte Rohstoff-Experte Avtar Sandu vom Broker Phillip Futures. US-Präsident Donald Trump kündigte Maßnahmen gegen China an, nannte aber keine Details. Düstere Prognosen über die wirtschaftlichen Auswirkungen der Pandemie belasten die Preise zusätzlich. RTR, REUTERS

ANLEIHEN, DEVISEN & ROHSTOFFE

und lag damit auf dem tiefsten Stand seit acht Wochen. Die spanischen und portugiesischen Pendants rentierten mit je 0,64 Prozent ebenfalls so niedrig wie seit Ende März nicht mehr.

Die EU-Kommission will im Kampf gegen die Corona-Krise ein Hilfspaket von 750 Milliarden Euro schnüren. Rund 500 Milliarden Euro sollen über Zuschüsse und das restliche Drittel über Kredite laufen. Das besonders von der Virus-Krise betroffene Italien solle 82 Milliarden Euro an Zuschüssen erhalten und 91 Milliarden Euro in Form von Darlehen, hieß es. Spanien soll demnach 77 Milliarden Euro an Zuschüssen bekommen und 63 Milliarden Euro über Kredite, die zurückgezahlt werden müssen.

ANZEIGE

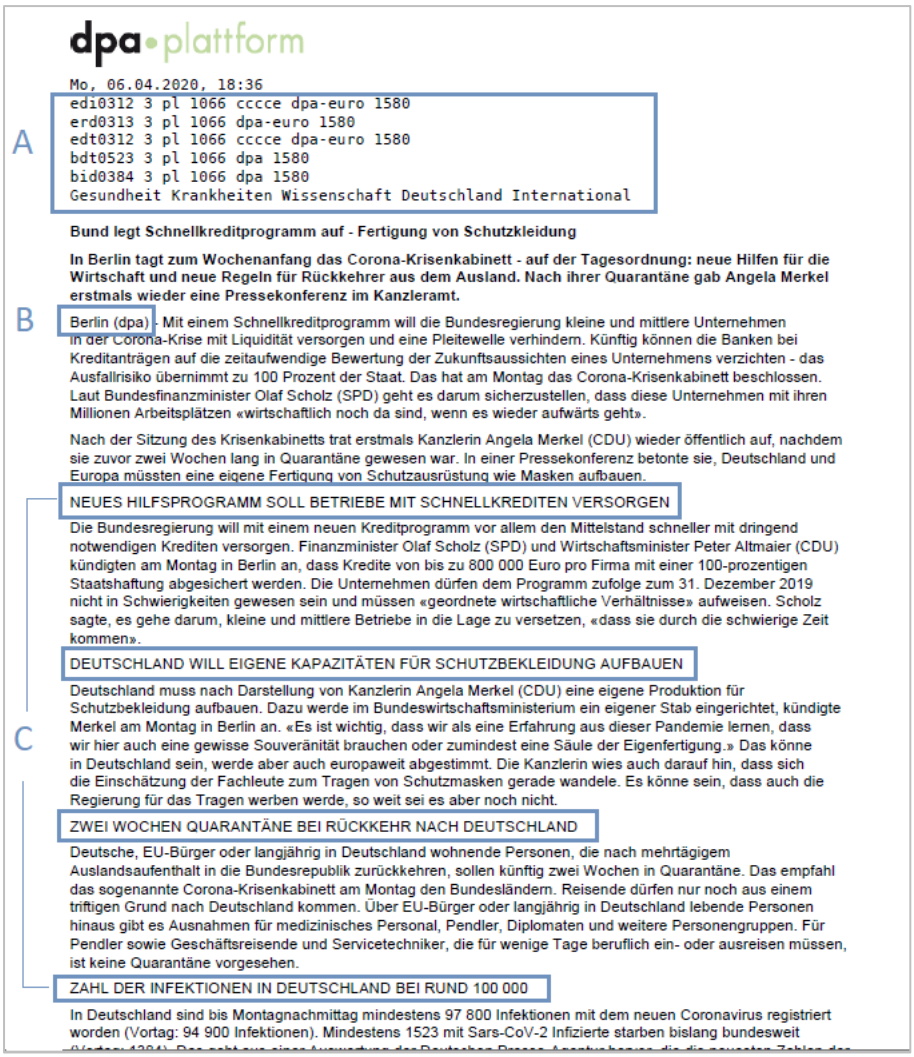
➔ Außerdem viel zusätzlicher, unerwünschter Inhalt (Börseninformationen, Grafiken, Werbeanzeige), der manuell bereinigt werden müsste bzw. die Ergebnisauswertung erschwert.

10) Wörter in Versalien als Einflussfaktor für *NER*-Analyse

A) Unnötige Symbole über *regular expressions* entfernen, um sauber die Überschrift auslesen können

B) Die Nennung der dpa am Anfang jedes Artikels erhöht die Anzahl der NEs künstlich und wurde daher im *Pre-Processing* per *regular expression* entfernt

C) Die großgeschriebenen Zwischenüberschriften beeinträchtigen die Identifikationsleistung der *NER*-Verfahren auch negativ und führen zu zahlreichen Fehlklassifikationen von Begriffen als Akteure



dpa • plattform

```
Mo, 06.04.2020, 18:36
edi0312 3 pl 1066 cccce dpa-euro 1580
erd0313 3 pl 1066 dpa-euro 1580
edt0312 3 pl 1066 cccce dpa-euro 1580
bdt0523 3 pl 1066 dpa 1580
bid0384 3 pl 1066 dpa 1580
Gesundheit Krankheiten Wissenschaft Deutschland International
```

A

Bund legt Schnellkreditprogramm auf - Fertigung von Schutzkleidung

In Berlin tagt zum Wochenanfang das Corona-Krisenkabinett - auf der Tagesordnung: neue Hilfen für die Wirtschaft und neue Regeln für Rückkehrer aus dem Ausland. Nach ihrer Quarantäne gab Angela Merkel erstmals wieder eine Pressekonferenz im Kanzleramt.

B

Berlin (dpa) - Mit einem Schnellkreditprogramm will die Bundesregierung kleine und mittlere Unternehmen in der Corona-Krise mit Liquidität versorgen und eine Pleitewelle verhindern. Künftig können die Banken bei Kreditanträgen auf die zeitaufwendige Bewertung der Zukunftsaussichten eines Unternehmens verzichten - das Ausfallrisiko übernimmt zu 100 Prozent der Staat. Das hat am Montag das Corona-Krisenkabinett beschlossen. Laut Bundesfinanzminister Olaf Scholz (SPD) geht es darum sicherzustellen, dass diese Unternehmen mit ihren Millionen Arbeitsplätzen «wirtschaftlich noch da sind, wenn es wieder aufwärts geht».

Nach der Sitzung des Krisenkabinetts trat erstmals Kanzlerin Angela Merkel (CDU) wieder öffentlich auf, nachdem sie zuvor zwei Wochen lang in Quarantäne gewesen war. In einer Pressekonferenz betonte sie, Deutschland und Europa müssten eine eigene Fertigung von Schutzausrüstung wie Masken aufbauen.

C

NEUES HILFSPROGRAMM SOLL BETRIEBE MIT SCHNELLKREDITEN VERSORGEN

Die Bundesregierung will mit einem neuen Kreditprogramm vor allem den Mittelstand schneller mit dringend notwendigen Krediten versorgen. Finanzminister Olaf Scholz (SPD) und Wirtschaftsminister Peter Altmaier (CDU) kündigten am Montag in Berlin an, dass Kredite von bis zu 800 000 Euro pro Firma mit einer 100-prozentigen Staatshaftung abgesichert werden. Die Unternehmen dürfen dem Programm zufolge zum 31. Dezember 2019 nicht in Schwierigkeiten gewesen sein und müssen «geordnete wirtschaftliche Verhältnisse» aufweisen. Scholz sagte, es gehe darum, kleine und mittlere Betriebe in die Lage zu versetzen, «dass sie durch die schwierige Zeit kommen».

DEUTSCHLAND WILL EIGENE KAPAZITÄTEN FÜR SCHUTZBEKLEIDUNG AUFBAUEN

Deutschland muss nach Darstellung von Kanzlerin Angela Merkel (CDU) eine eigene Produktion für Schutzbekleidung aufbauen. Dazu werde im Bundeswirtschaftsministerium ein eigener Stab eingerichtet, kündigte Merkel am Montag in Berlin an. «Es ist wichtig, dass wir als eine Erfahrung aus dieser Pandemie lernen, dass wir hier auch eine gewisse Souveränität brauchen oder zumindest eine Säule der Eigenfertigung.» Das könne in Deutschland sein, werde aber auch europaweit abgestimmt. Die Kanzlerin wies auch darauf hin, dass sich die Einschätzung der Fachleute zum Tragen von Schutzmasken gerade wandele. Es könne sein, dass auch die Regierung für das Tragen werben werde, so weit sei es aber noch nicht.

ZWEI WOCHEN QUARANTÄNE BEI RÜCKKEHR NACH DEUTSCHLAND

Deutsche, EU-Bürger oder langjährig in Deutschland wohnende Personen, die nach mehrtägigem Auslandsaufenthalt in die Bundesrepublik zurückkehren, sollen künftig zwei Wochen in Quarantäne. Das empfahl das sogenannte Corona-Krisenkabinett am Montag den Bundesländern. Reisende dürfen nur noch aus einem triftigen Grund nach Deutschland kommen. Über EU-Bürger oder langjährig in Deutschland lebende Personen hinaus gibt es Ausnahmen für medizinisches Personal, Pendler, Diplomaten und weitere Personengruppen. Für Pendler sowie Geschäftsreisende und Servicetechniker, die für wenige Tage beruflich ein- oder ausreisen müssen, ist keine Quarantäne vorgesehen.

ZAHL DER INFEKTIONEN IN DEUTSCHLAND BEI RUND 100 000

In Deutschland sind bis Montagnachmittag mindestens 97 800 Infektionen mit dem neuen Coronavirus registriert worden (Vortag: 94 900 Infektionen). Mindestens 1523 mit Sars-CoV-2 Infizierte starben bislang bundesweit

11) Prozessor laden bei *Stanza*

In der NLP-Bibliothek stanza ist für *NER* das mit dem *CoNLL*-Korpus trainierte Modell als *default* vorinstalliert, daher muss explizit das GermEval2014 Modell angegeben und heruntergeladen werden:

```
processor_dict = {
    'tokenize': 'default',
    'ner': 'germeval2014',
}
stanza.download('de', processors=processor_dict, package=None)
nlp = stanza.Pipeline('de', processors=processor_dict, package=None)
```

```
Downloading https://raw.githubusercontent.com/stanfordnlp/stanza-resources/master/resources_1.0.0.json: 120kB [00:00, ?B/s]
2020-12-12 08:54:58 INFO: Downloading these customized packages for language: de (German)...
=================================
| Processor       | Package     |
---------------------------------
| tokenize        | gsd         |
| ner             | germeval2014 |
| forward_charlm  | newswiki    |
| backward_charlm | newswiki    |
=================================

2020-12-12 08:54:59 INFO: File exists: C:\Users\cecil\stanza_resources\de\tokenize\gsd.pt.
Downloading http://nlp.stanford.edu/software/stanza/1.0.0/de/ner/germeval2014.pt:  45%|████      | 159M/351M [00:18<00:23, 8.19MB/s]
```

12) Dataframe mit Spalten für die identifizierten Eigennamen (PER, ORG, LOC)

```
text_frame.head()
```

	source	title	body	entities	entities_persons	entities_organisations	entities_locations
0	dpa	US-Senat will Konjunkturprogramm um 250 Millia...	Washington (dpa) - Das riesige US-Konjunkturpa...	((Washington), (dpa), (US-Konjunkturpaket), (S...	[(Steven, Mnuchin), (Mitch, McConnell), (Nancy...	[(US-Konjunkturpaket), (Demokraten), (Repräsen...	[(Washington), (dpa), (Amerikaner), (den, USA)]
1	dpa	Weiterer Eilantrag gegen bayerische Corona-Bes...	Karlsruhe (dpa) - Das Bundesverfassungsgericht...	((Karlsruhe), (Bundesverfassungsgericht), (bay...	[]	[(Corona-Pandemie)]	[(Bundesverfassungsgericht), (Bayern)]
2	dpa	Coronavirus auf Jet-Set-Insel Mykonos - Ausgan...	Athen (dpa) - Nachdem bei zwei Menschen auf My...	((Athen), (dpa), (Mykonos), (Athen), (Jet-Set-...	[]	[(dpa)]	[(Athen), (Mykonos), (Athen), (Athen), (Haus, ...

13) Umwandlung des erhaltenen Outputs für die Auswertung

Mtitle	Artikel	NE_PER	NE_ORG
dpa	Tübinger Firma soll Impfstoff gegen das Coronavirus finden	[Anja Karliczek, Stephan Becker]	[CDU, DZIF, Instituts für Virologie, Universität Marburg]
WELT	Uneins in der Krise	[Markus Söder, Söder, Peter Altmaier, Altmaier, Ralph Brinkhaus, Olaf Scholz, Altmaier, Brinkhaus, Scholz]	[Deutsche Industrie- und Handelskammertags, DIHK, Bundesverbands der Deutschen Tourismuswirtschaft, SPD, Bundestag, CDU, CSU, SPD, CDU/CSU]
SPIEGEL	Der Corona-Schock	[Jim Reid, Jens Spahn, Nouriel Roubini, Lehman, Lehman, Donald Trump, Roubini, Roubini]	[Deutschen Bank, Apple, Apple, Foxconn, CDU, Lehman Brothers, Universität St. Gallen, Welthandelsorganisation, WTO, WTO]

Mtitle	Artikel	NE	Klasse
dpa	Tübinger Firma soll Impfstoff gegen das Coronavirus finden	Anja Karliczek	PER
dpa	Tübinger Firma soll Impfstoff gegen das Coronavirus finden	Stephan Becker	PER
dpa	Tübinger Firma soll Impfstoff gegen das Coronavirus finden	CDU	PER
dpa	Tübinger Firma soll Impfstoff gegen das Coronavirus finden	DZIF	ORG
dpa	Tübinger Firma soll Impfstoff gegen das Coronavirus finden	Instituts für Virologie	ORG
dpa	Tübinger Firma soll Impfstoff gegen das Coronavirus finden	Universität Marburg	ORG
WELT	Uneins in der Krise	Markus Söder	PER
WELT	Uneins in der Krise	Söder	PER
WELT	Uneins in der Krise	Peter Altmaier	PER

14) Anzahl ermittelter Personen und Organisationen pro Artikel (n = 887)
(→ Berechnungen in separatem Anhang [E])

	Mittelwert	Standard-abweichung	Min.	Max.	Total (PER + ORG)
spaCy	6,6	6,2	0	52	7.073
Stanza	6,5	6,0	0	50	6.470
FLAIR	5,7	5,5	0	46	5.471

Die Standardabweichung gibt an, dass es sich bei den Artikeln mit über 40 Eigennamen, um Ausreißer handelt und die Werte eher im Bereich zwischen 0 und 12 streuen.

15) Zwanzig am häufigsten extrahierte Eigennamen der Klassen ‚LOC' und ‚MISC'

LOC						MISC					
spaCy		**stanza**		**flair**		**spaCy**		**stanza**		**flair**	
Deutschland	273	Deutschland	283	Deutschland	280	Virus	251	Euro	137	Euro	129
Berlin	242	Berlin	232	Berlin	243	Covid-19	124	Dollar	42	Sars-CoV-2	87
China	206	China	214	China	211	Corona-Krise	122	Dax	20	Covid-19	75
Europa	140	Europa	143	Europa	140	Zeit	117	Corona	18	Corona	67
Italien	129	Italien	131	Italien	130	deutschen	114	US	17	Coronavirus	48
USA	117	USA	117	USA	117	deutsche	113	Corona-Krise	17	Dollar	29
Wuhan	81	Wuhan	82	Wuhan	81	chinesischen	88	COVID-19	14	Corona-Krise	26
Frankreich	75	Frankreich	76	Frankreich	76	europäische	66	Zweiter Weltkrieg	11	US-Dollar	19
Baden-Württemberg	65	Bayern	67	Stuttgart	65	Load-Date	59	Eurozone	9	Olympischen Spiele	16
Bayern	62	Stuttgart	66	Baden-Württemberg	64	chinesische	58	deutsch	9	Corona-Pandemie	14
Stuttgart	61	Peking	59	Bayern	64	sozialen	48	MDax	8	Zweiten Weltkrieg	11
Peking	61	Japan	53	Peking	62	Berliner	42	Page 2 of 2	8	Diamond Princess	9
Frankfurt	56	Südkorea	52	Frankfurt	58	Coronavirus-Krise	36	Diamond Princess	8	Formel 1	9
Japan	52	Spanien	50	Südkorea	52	Ostern	34	Corona-Epidemie	7	Neujahrsfest	7
Südkorea	51	München	45	München	52	Internet	33	Corona-Pandemie	7	Champions League	6
München	51	Hubei	45	Japan	52	soziale	31	Cent	6	CoV-2	6
Spanien	50	Nordrhein-Westfalen	43	Spanien	50	Twitter	28	Westerdam	6	Olympischen Spieler	5
Nordrhein-Westfalen	49	Österreich	42	Nordrhein-Westfalen	49	Page 2 of 2	28	EuroStoxx 50	5	Olympia	5
Österreich	42	Baden-Württemberg	40	Hubei	44	Viren	25	iPhone	4	EuroStoxx 50	5
Rom	40	Rom	38	Washington	42	Corona	25	CoV-2	4	Grand Prix	5

16) Schwierigkeit bei Vorkommen von gleichem Organisations- und Personennamen

Fielmann produziert jetzt auch Schutzbrillen für Ärzte

Hamburg (dpa) - Die Optik-Kette Fielmann hat wegen der Corona-Krise die Entwicklung und Produktion von Schutzbrillen für Ärzte und medizinisches Fachpersonal aufgenommen. «Wir rechnen binnen zwei Wochen mit der Zertifizierung und können dann qualitativ hochwertige Schutzbrillen in unterschiedlichen Ausführungen zur Verfügung stellen», erklärte der Vorstandsvorsitzende Marc Fielmann am Donnerstag in Hamburg.

Die Fertigung sei bereits in der vergangenen Woche im brandenburgischen Rathenow aufgenommen worden. Parallel werde Fielmann seine Produktionskapazitäten ausweiten und ab Mitte April große Kontingente Schutzbrillen im Versand anbieten. Ab Ende April werde Fielmann in der Lage sein, Schutzbrillen auch in individueller Sehstärke zu fertigen. Die ersten 20 000 Brillen spendet Fielmann an Krankenhäuser und medizinische Einrichtungen

Artikel	NE-Klasse	Eigenname	Bibliothek
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	ORG	Fielmann	flair
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	PER	Fielmann	flair
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	PER	Fielmann	flair
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	PER	Fielmann	flair
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	PER	Marc Fielmann	flair
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	ORG	Fielmann	spaCy
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	ORG	Fielmann	spaCy
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	ORG	Fielmann	spaCy
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	ORG	Fielmann	spaCy
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	PER	Marc Fielmann	spaCy
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	ORG	Fielmann	stanza
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	PER	Fielmann	stanza
Fielmann produziert jetzt auch Schutzbrillen für Ärzte	PER	Fielmann	stanza

(Quelle: dpa-Artikel im Datensatz und dazu erhaltene NER-Ergebnisse)

17) Beispiele für den Umgang mit mehrdeutigen Eigennamen

Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien

London (dpa) - Die Coronavirus-Pandemie dürfte die weltweite Versicherungsbranche nach Einschätzung des Londoner Versicherungsmarkts Lloyd's so teuer zu stehen kommen wie die verheerenden Hurrikan-Serien von 2005 und 2017. Insgesamt schätzt die Lloyd's-Führung um John Neal die Schadenzahlungen auf 107 Milliarden US-Dollar (98,7 Mrd Euro), wie Lloyd's am Donnerstag in London mitteilte. Hinzu kämen Wertverluste bei Kapitalanlagen, die Lloyd's auf 96 Milliarden Dollar taxiert.

Im Jahr 2005 hatten die Hurrikane «Katrina», «Rita» und «Wilma» versicherte Schäden von 116 Milliarden Dollar angerichtet. 2017 schlugen die Zerstörungen durch «Harvey», «Irma» und «Maria» mit 92 Milliarden Dollar zu Buche.

Artikel	NE Klasse	Eigennname	Bibliothek
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Katrina	flair
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Katrina	spaCy
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Katrina	stanza
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Rita	flair
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Rita	spaCy
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Rita	stanza
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	PER	Wilma	spaCy
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Wilma	flair
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Wilma	stanza
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	PER	Harvey	flair
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	PER	Harvey	spaCy
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	PER	Harvey	stanza
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Irma	flair
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Irma	spaCy
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Irma	stanza
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Maria	flair
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	PER	Maria	spaCy
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	MISC	Maria	stanza

Testbeispiel: ‚Essen' - *spaCy* kennzeichnet irrtümlich drei Eigennamen, während *Stanza* und *FLAIR* nur das richtige Ergebnis ausgeben:

```
import spacy
from spacy import displacy

nlp = spacy.load('de_core_news_lg')
text = "Eines der größten Veranstaltungszentren der Stadt ist die Messe Essen. Man kann auf der Messe Essen und Getränke kaufen und verzehren."
doc = nlp(text)
svg = displacy.render(doc, style='ent', jupyter=True)
```

Eines der größten Veranstaltungszentren der Stadt LOC ist die Messe Essen LOC . Man kann auf der Messe Essen LOC und Getränke kaufen und verzehren.

```
============================
| Processor | Package      |
----------------------------
| tokenize  | gsd          |
| ner       | germeval2014 |
============================

2020-12-16 15:22:09 INFO: Use device: cpu
2020-12-16 15:22:09 INFO: Loading: tokenize
2020-12-16 15:22:09 INFO: Loading: ner
2020-12-16 15:22:11 INFO: Done loading processors!
token: Eines     ner: O
token: der       ner: O
token: größten   ner: O
token: Veranstaltungszentren    ner: O
token: der       ner: O
token: Stadt     ner: O
token: ist       ner: O
token: die       ner: O
token: Messe     ner: B-ORG
token: Essen     ner: E-ORG
token: .         ner: O
token: Man       ner: O
token: kann      ner: O
token: auf       ner: O
token: der       ner: O
token: Messe     ner: O
token: Essen     ner: O
token: kaufen    ner: O
token: und       ner: O
token: verzehren        ner: O
token: .         ner: O
```

```
from flair.data import Sentence
from flair.models import SequenceTagger

# make a sentence
sentence = Sentence('Eines der größten Veranstaltungszentren
                     der Stadt ist die Messe Essen. Man kann
                     auf der Messe Essen und Getränke kaufen ur
# load the NER tagger
tagger = SequenceTagger.load('de-ner')

# run NER over sentence
tagger.predict(sentence)

# iterate over entities and print
for entity in sentence.get_spans('ner'):
    print(entity)

The following NER tags are found:
Span [9,10]: "Messe Essen"   [- Labels: LOC (0.7449)]
```

spaCy klassifiziert beide Male den Begriff ‚Essen' als Ort klassifiziert, während *Stanza* und *FLAIR* nur das relevante Wort als Organisation kennzeichnen → weiterer Beleg für *Stanza* und *FLAIR*s hohe *Precision*

18) Fehlklassifikation aufgrund von Rechtschreibfehlern

FRANKEICH WEITER HART GETROFFEN

Die Maßnahmen zur Eindämmung des Virus beginnen in Frankreich Wirkung zu zeigen, dennoch bleibt die Lage ernst. Fast 14 400 Menschen sind an den Folgen von Covid-19 verstorben, wie das Gesundheitsministerium am Sonntagabend in Paris mitteilte. Wenigstens sank am vierten Tag in Folge die Zahl der Menschen, die auf der Intensivstation behandelt werden, leicht. «Diese Daten bestätigen, dass die Epidemie in unserem Land in dynamischer Weise weitergeht und es weiterhin hart trifft», so das Gesundheitsministerium. Man beobachte den Beginn «eines sehr hohen Plateaus», müsse aber wachsam bleiben.

Medientitel	Artikelüberschrift	NE-Klasse	Name	Bibliothek
dpa	Die Welt im Corona-Dilemma: Vorsicht oder Rückkehr zum Normalbetrieb?	PER	Frankeich	flair
dpa	Die Welt im Corona-Dilemma: Vorsicht oder Rückkehr zum Normalbetrieb?	PER	Frankeich	spaCy
dpa	Die Welt im Corona-Dilemma: Vorsicht oder Rückkehr zum Normalbetrieb?	PER	Frankeich	stanza
Die Welt	Jetzt wird es ernst bei den Steuern	ORG	Landesfinanminister	stanza

Die Ausgaben zur Bekämpfung der Corona-Pandemie bewegen sich in ganz anderen Größenordnungen. Wobei strittig ist, um wie viel Geld es geht. Während Bundeskanzlerin Angela Merkel (CDU) und Bundesfinanzminister Olaf Scholz (SPD) in der Vorwoche von 130 Milliarden Euro für die Jahre 2020 und 2021 sprachen, kam die Zentrale Datenstelle der Landesfinanminister bei einer Bewertung des Konjunkturpakets sogar auf 167 Milliarden Euro. Davon entfallen 124 Milliarden Euro auf das Jahr 2020, 32 Milliarden Euro auf das Jahr 2021 und gut zehn Milliarden Euro auf die Folgejahre.

19) SpaCy Fehlidentifikation (Artikel am Satzanfang sowie Wochentage)

Mtitle	Artikel	NE-Klasse	Eigenname	Bibliothek
Welt	Bücher ohne Bühne?	MISC	Das Berührende	spaCy
dpa	Covid-19: Europaabgeordneter	MISC	Das Chaos	spaCy
dpa	Coronavirus: Gläubige sollen di	MISC	Das edle Heiligtum	spaCy
Welt	Der Politiker als Literatur-Influe	MISC	Das Ende der Illusionen	spaCy
Spiegel	Ein großes Experiment	MISC	Das entscheide ich!	spaCy
dpa	EU-Kommission bereitet Exit-Str	MISC	Das Entscheidende	spaCy
Welt	ITB steht vor dem Aus	MISC	Das Gedächtnis des Urlaubers	spaCy
dpa	Ende der Isolation - China-Rück	MISC	Das Gefühl der Befreiung	spaCy
Welt	Was sich im Juni ändert	MISC	Das Gesetz zur Förderung der beruflic	spaCy
Welt	Die Corona-Charaktere	MISC	Das Glück des Augenblicks	spaCy
Welt	Bücher ohne Bühne?	MISC	Das kleine Glück im großen Unglück	spaCy
dpa	Milliardenschweres Hilfsprogra	MISC	Das Land	spaCy
Welt	Die Corona-Charaktere	MISC	Das Leben wenig	spaCy
Welt	Bücher ohne Bühne?	MISC	Das literarische Leben	spaCy
Welt	Bücher ohne Bühne?	MISC	Das Live-Erlebnis	spaCy
Welt	ITB steht vor dem Aus	MISC	Das löse	spaCy
dpa	Mit «Krötenflummis» und Albat	MISC	Das macht Spaß	spaCy
Welt	"Als Kind wollte man mich töter	MISC	Das macht zusätzliche Angst	spaCy
dpa	Bundesrat beschließt Corona-H	MISC	Das Miteinander	spaCy

PER	Vortag	1	spaCy
PER	Vortag	1	spaCy
PER	Vortag	1	spaCy
PER	Vortag	1	spaCy
PER	Vortag	1	spaCy
ORG	Donnerstag	1	spaCy
ORG	Donnerstag	1	spaCy
ORG	Donnerstag	1	spaCy
ORG	Inter Mailand am Sonntag	1	spaCy
ORG	Samstags-Durchschnittswerten	1	spaCy
LOC	Donnerstagmorgen	1	spaCy
LOC	Freitagmorgen	1	spaCy
LOC	Freitagmorgen	1	spaCy
LOC	Institut am Sonntag	1	spaCy
LOC	Montagfrüh	1	spaCy
LOC	Montagmittag	1	spaCy
LOC	Osterfeiertage	1	spaCy
LOC	Ostermontag	1	spaCy
LOC	Ostermontag	1	spaCy
LOC	Samstagmittag	1	spaCy
LOC	Samstagvormittag	1	spaCy
LOC	Weißen Haus am Sonntag	1	spaCy

20) Klassifikation der Corona-Begriffe pro Bibliothek

'Corona'-Anteile									
	spaCy			Stanza			FLAIR		
	NEs	CoV	%	NEs	CoV	%	NEs	CoV	%
PER	2.833	72	2,5%	2.480	126	5,1%	2.503	109	4,4%
ORG	4.240	382	9,0%	3.990	225	5,6%	2.968	145	4,9%
LOC	6.882	352	5,1%	6.611	25	0,4%	4.824	30	0,6%
MISC	4.319	998	23,1%	667	95	14,2%	915	372	40,7%
	18.274	**1.804**	**9,9%**	**13.748**	**471**	**3,4%**	**11.210**	**656**	**5,9%**

→ Doppelungen pro Artikel zusammengefasst, Begriffe werden nur einmalig gezählt

21) Zusätzlich ermittelte Akteure, die nicht als *True Positives* zählen, da sie nicht manuell codiert wurden:

Unique PER und ORG	Korrekt extrahiert	True Positive	Zusatz		Korrekt extrahiert	True Positives	Zusatz		Korrekt extrahiert	True Positives	Zusatz	
PER	1.480	**956**	+ 55%	524	1.398	**955**	+ 46%	443	1.424	**955**	+49%	469
ORG	1.438	**696**	+107%	742	1.567	**639**	+145%	928	1.221	**584**	+109%	637

*Loose Matching

22) Manuell ermittelte Parteizugehörigkeiten und extrahierte Parteinamensnennungen *(Zugrunde liegende Berechnung in Anhang [F])*

Manuell codierte Zugehörigkeit der Akteure		**Automatisiert extrahierte Nennungen**	
Partei	**Artikelanzahl**	**Partei**	**Artikelanzahl**
CDU/CSU	34	CDU/CSU	159
SPD	31	SPD	68
FDP	29	Die Grünen	46
Die Grünen	20	FDP	17
Die Linke	7	Die Linke	9
AfD	6	AfD	8

23) Unterschiede im Chunking von Eigennamen der Klasse 'ORG'

Artikel	**Eigenname**	**Bibliothek**
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	Versicherungsmarkt Lloyd's	spaCy
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	Lloyd's	Stanza
Versicherungsmarkt Lloyd's: Corona-Krise so teuer wie Hurrikan-Serien	Lloyd's	FLAIR
Füchse-Präsident fordert unbürokratische Ausfallhilfe für Vereine	Handball-Bundesligisten Füchse Berlin	spaCy
Füchse-Präsident fordert unbürokratische Ausfallhilfe für Vereine	Füchse Berlin	Stanza
Füchse-Präsident fordert unbürokratische Ausfallhilfe für Vereine	-	FLAIR
Tragen wir bald alle Schutzmaske?	Ärztegewerkschaft Marburger Bund	spaCy
Tragen wir bald alle Schutzmaske?	Marburger Bund	Stanza
Tragen wir bald alle Schutzmaske?	Marburger Bund	FLAIR
Altmaier stellt Nachbesserungen bei Corona-Hilfen in Aussicht	Autoverband VDA	spaCy
Altmaier stellt Nachbesserungen bei Corona-Hilfen in Aussicht	VDA	Stanza
Altmaier stellt Nachbesserungen bei Corona-Hilfen in Aussicht	VDA	FLAIR

24) Manuell codierte generische Akteure

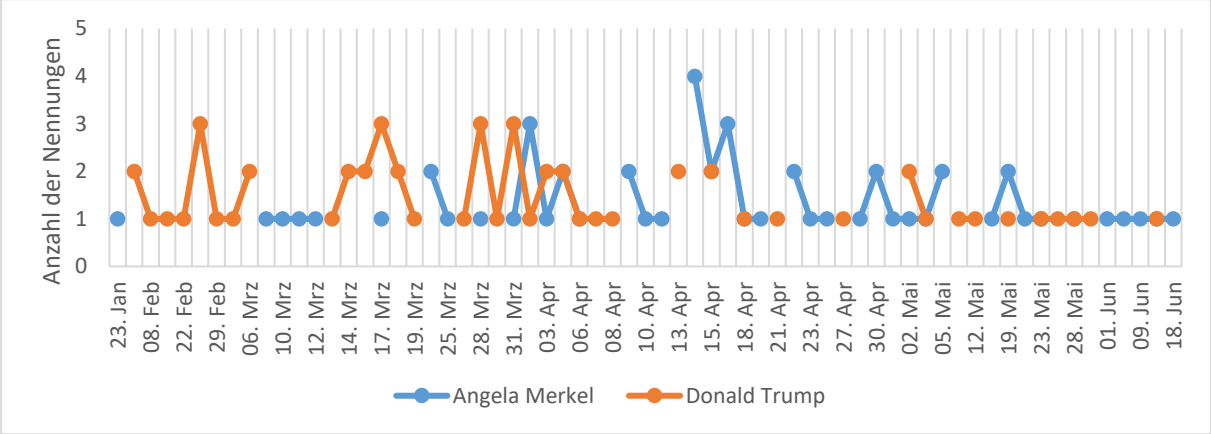
Generische Akteure	spaCy	Stanza	FLAIR
Experten	#NV	#NV	#NV
Nutzer	#NV	#NV	#NV
Forscher	#NV	#NV	#NV
Wissenschaftler	#NV	#NV	#NV
Kritiker	#NV	#NV	#NV
Regierungskreise	#NV	#NV	#NV
Kliniken	#NV	#NV	#NV
italienische Politiker	#NV	#NV	#NV
Mediziner	#NV	#NV	#NV
Techniker	#NV	#NV	#NV
Republikaner	2	#NV	#NV
Berater	#NV	#NV	#NV
Ingenieur	#NV	#NV	#NV
Reisende	#NV	#NV	#NV
Epidemiologen	#NV	#NV	#NV
Fabrikanten	#NV	#NV	#NV
Seuchenexperten	1	#NV	#NV
Chinesen	27	28	#NV
Hongkonger Spitzenforscher	#NV	#NV	#NV
Hausarzt	#NV	#NV	#NV
junger Mann	#NV	#NV	#NV
Krankenschwester	#NV	#NV	#NV
Arzt	#NV	#NV	#NV
Pfleger	#NV	#NV	#NV
Bruder	#NV	#NV	#NV
Freundin	#NV	#NV	#NV
Branchenvertreter	#NV	#NV	#NV
Analysten	1	#NV	#NV
Börsianer	1	#NV	#NV
Händler	1	1	#NV
Preisportale	1	#NV	#NV
Gesundheitsexperten	1	#NV	#NV
Einwohner	#NV	#NV	#NV
Freund	#NV	#NV	#NV
Feriengäste	#NV	#NV	#NV
Junge Frau	#NV	#NV	#NV
Kumpel	#NV	#NV	#NV
Ärzte	#NV	#NV	#NV
Angestellter	#NV	#NV	#NV
Ermittler	#NV	#NV	#NV
Heimleitung	#NV	#NV	#NV
Matrosen	#NV	#NV	#NV
Außenminister	#NV	#NV	#NV
Ministerpräsidenten	#NV	#NV	#NV
Fahrer	#NV	#NV	#NV
Finanzminister	1	#NV	#NV
Schülerin	#NV	#NV	#NV
Expertem	#NV	#NV	#NV
Forscherteams	#NV	#NV	#NV
Soziologen	#NV	#NV	#NV
Juristen	#NV	#NV	#NV
Sportler	#NV	#NV	#NV
Funktionäre Australien	#NV	#NV	#NV
Virologen	#NV	#NV	#NV
Früherer Chef der Staatlichen V	#NV	#NV	#NV
Opposition Brasilien	#NV	#NV	#NV
Menschenrechtsbeobachter	#NV	#NV	#NV
Gymnasiallehrer in Baden-Wü	#NV	#NV	#NV
Beobachter	#NV	#NV	#NV
Politiker	#NV	#NV	#NV
Europäische Staaten	2	#NV	#NV
Autoren	#NV	#NV	#NV
Medizinerkollege	#NV	#NV	#NV
Kläger	#NV	#NV	#NV
Datenschützer	#NV	#NV	#NV
Kultusminister	#NV	#NV	#NV
lokale Behörden	#NV	#NV	#NV
Unionsabgeordnete	#NV	#NV	#NV
Maschinell gefunden	**10**	**2**	**0**

25) Uneinheitlich oder fehlerhaft codierte Personennamen bei der manuellen Erhebung

Artikel-ID	Spezifizier	Vor- und Zuname des Akteurs	Korrektur
20244	individuell	Andrea Blbi	Balbi
10007	individuell	Annette Messanger	Annette Messager
20231	individuell	Bertram Brossadt	Brossardt
41034	individuell	Bodow Ramelow	Bodo
40667	individuell	Christopher Berger	Burger
20209	individuell	Clemes Fuest	Clemens
20006	individuell	Daniel Steller	Stelter
20032	individuell	Deborah Brix	Birx
40253	individuell	Deborah Brix	Birx
40057	individuell	Dietman Woidke	Dietmar
40905	individuell	Dimitri Peskow	Dmitri
20220	individuell	Erk Saarschmidt	Schaarschmidt
40782	individuell	Fraucke Gerlach	Frauke
40857	individuell	Friedmann Schmidt	Friedemann
41036	individuell	Gebharf Fürst	Gebhard
10059	individuell	Gérad Krause	Gérald
40900	individuell	Giulio Gallera	Galleraso
40556	individuell	Gordon Isler	Gorden
40892	individuell	Guiseppe Conte	Giuseppe
40900	individuell	Guiseppe Conte	Giuseppe
40456	individuell	Hakan Samuelsson	Håkan
40551	individuell	Joachim Ruckwied	Rukwied
40303	individuell	Joel Bigayon	Joël
40337	individuell	Kasper Rosted	Rorsted
40176	individuell	Laurence Boon	Boone
20067	individuell	Lena Dupont	Düpont
40615	individuell	Lothar Weiler	Wieler
20232	individuell	Lutz Leichenring	Leichsenring
40051	individuell	Maria Jesús Montero	María
40288	individuell	Maria José Sierra	María
40145	individuell	Nabil al-Asan	Asani
40385	individuell	Norbert Walther-Borjans	Walter-Borjans
20234	individuell	Nura al-Moriari	Motiari
20150	individuell	Peter Altmeier	Altmaier
40212	individuell	Peter Altmeier	Altmaier
40340	individuell	Peter Altmeier	Altmaier
20015	individuell	Péter Gyorkos	Györkös
20020	individuell	Philipp Lee	Phillip
20178	individuell	Phillipe Waechter	Philipe
20229	individuell	Rald Moldenhauer	Ralf
40912	individuell	Roland Döhm	Döhrn
20243	individuell	Rolpf Mützenich	Rolf
20119	individuell	Rudolf Trettenbein	Trettenbrein
40647	individuell	Sabine Bätzing-Lichtenthaler	Lichtenthäler
20232	individuell	Sasche Disselkamp	Sasha
40773	individuell	söJean-Jacques Henchoz	Jean
40819	individuell	Sosanne Johna	Susanne
41067	individuell	Susanne Henning-Wellsow	Hennig-Wellsow
40799	individuell	Ulrich Mäuer	Mäurer
40819	individuell	Ulrich Mäuerer	Mäurer
40982	individuell	Ulrich Mäuerer	Mäurer
40954	individuell	Viktor Orban	Orbán
40212	individuell	Walter-Borjans	Norbert
10060	individuell	Zhong Nan-shan	Nanshan

26) Mögliche Auswertungen des untersuchten Datensatzes:

Vorkommen der Akteure im Zeitverlauf



Hier exemplarisch an zwei Akteuren aus dem Corona-Datensatz abgebildet:
- Zeitachse ergibt sich aus den Veröffentlichungsdatum der analysierten Artikel
- Abgetragen wird die Anzahl der Artikel, in denen der Akteur vorkommt (Y-Achse)
- Auch sichtbar, dass die Akteure nur in einer bestimmten Phase zeitgleich in der Berichterstattung vorkommen

Akteursvorkommen und -vielfalt pro Medientitel

dpa	1488	Spiegel	207	Welt	592
Angela Merkel	34	Donald Trump	9	Angela Merkel	17
Donald Trump	32	Angela Merkel	5	Donald Trump	16
Jens Spahn	28	Jens Spahn	4	Jens Spahn	13
Markus Söder	17	Christian Drosten	4	Olaf Scholz	10
Olaf Scholz	15	Boris Johnson	2	Markus Söder	9
Winfried Kretschmann	13	Georg Marckmann	2	Ursula von der Leyen	8
Peter Altmaier	13	Hitler	2	Christian Drosten	6
Christian Drosten	13	Xi Jinping	2	Armin Laschet	4
Manne Lucha	10	Sebastian Kurz	2	Xi Jinping	4
Heiko Maas	10	Emmanuel Macron	2	Boris Johnson	4
Horst Seehofer	10	Winston Churchill	2	Tedros Adhanom Ghebr	4
Emmanuel Macron	9	Regine Heiland	1	Peter Altmaier	3
Ursula von der Leyen	8	Markus Schinwald	1	Horst Seehofer	3
Xi Jinping	8	Vincent Cheng	1	Franziska Giffey	3
Franziskus	8	Bill Bishop	1	Robin Alexander	3

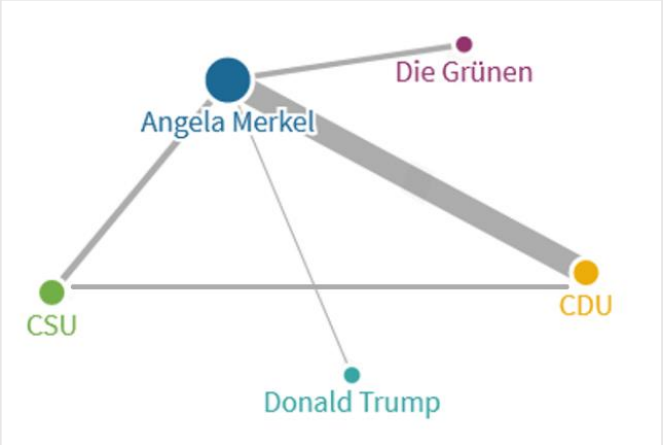
Die erhaltenen Ergebnisse verschiedener Pressetitel untereinander zu vergleichen, kann interessante Rückschlüsse über die redaktionelle Linie oder Vielfalt der Medientitel ermöglichen.

Die Stichprobe des hier untersuchten Datensatzes ist jedoch zu unausgeglichen (Spiegel = 37 Artikel, WELT 142 Artikel, dpa = 710 Artikel), um repräsentative Aussagen über das Medium treffen zu können.

Kookkurrenz und Relevanz der Akteure

Alternativ könnte auch das Vorkommen der Akteure im Datensatz auch als Netzwerk visualisiert werden:
- Größe der Kreise gibt Anzahl der Artikel an, in denen Akteure vorkommen (Indikator für Relevanz)

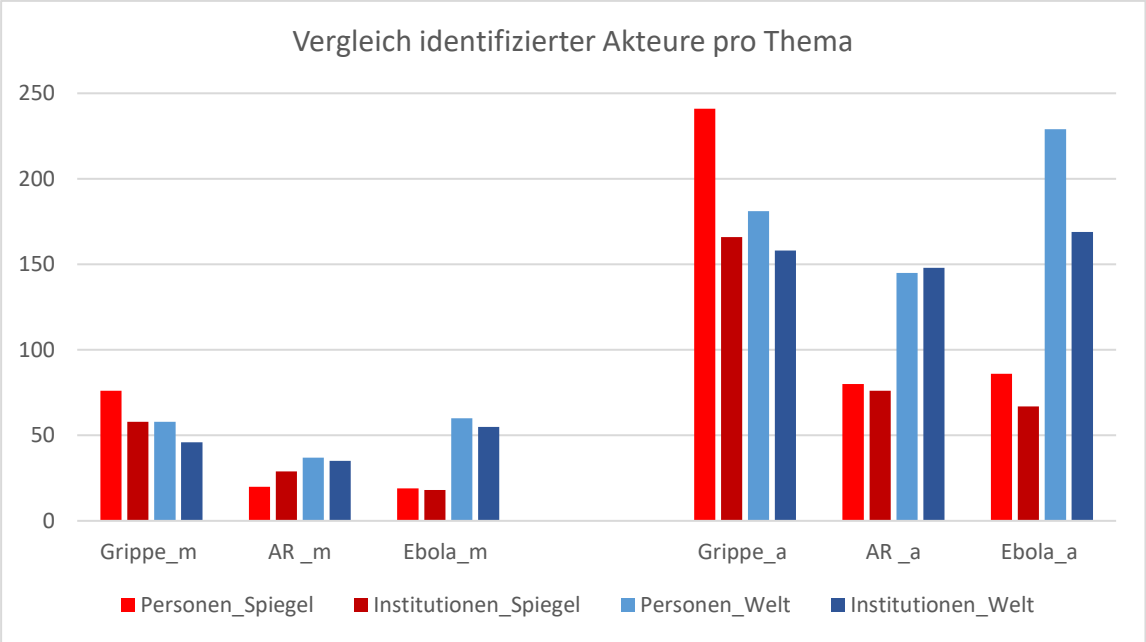
- Linien bilden die Kookkurrenz der Akteure in den Texten ab und die Dicke der Linien stellt die Häufigkeit dieses gemeinsamen Vorkommens dar



27) Auswertungen des Testdatensatzes (Grippe-, AR-, Ebola-Artikel)

	Grippe	AR	Ebola
SPIEGEL	23 Artikel	9 Artikel	14 Artikel
WELT	44 Artikel	29 Artikel	40 Artikel

Vergleich manuell und automatisiert identifizierter Personen und Institutionen:



→ *einmalige Bezeichnungen nicht kumulierte Zusammenfassung der Nennungen*

Berechnung der Anteile biomedizinischer Begriffe und falsch identifizierte NEs:

	Grippe			AR			Ebola		
SPIEGEL	**166**	**ORG**		**76**	**ORG**		**67**	**ORG**	
	50	falsch	**30,1%**	26	falsch	**34,2%**	24	falsch	**35,8%**
	12	biomed.	24,0%	4	biomed.	15,4%	11	biomed.	45,8%
	241	**PER**		**80**	**PER**		**86**	**PER**	
	35	falsch	**14,5%**	12	falsch	**15,0%**	21	falsch	**24,4%**
	12	biomed.	34,3%	8	biomed.	66,7%	4	biomed.	19,0%
WELT	**158**	**ORG**		**148**	**ORG**		**169**	**ORG**	
	33	falsch	**20,9%**	15	falsch	**10,1%**	32	falsch	**18,9%**
	10	biomed.	30,3%	2	biomed.	13,3%	12	biomed.	37,5%
	181	**PER**		**145**	**PER**		**229**	**PER**	
	35	falsch	**19,3%**	36	falsch	**24,8%**	29	falsch	**12,7%**
	18	biomed.	51,4%	25	biomed.	69,4%	11	biomed.	37,9%

→ *biomedizinische Begriffe werden sowohl als PER als auch ORG klassifiziert*
→ *geringere Anteile an falschen Eigennamen bei Per als bei ORG*

28) SPIEGEL Artikel (AR, Ebola, Grippe)

Übersicht - Manuell codierte Akteure (Personen & Institutionen)

Spiegel

Personen

Antibiotika (20)		Ebola (19)		Grippe (76)							
Barbara Stamm	1	Anja Wolz	1	Alejandro López Ruiz	1	Gérard Krause	1	Mark Simmerman	3	Thomas Grein	1
Gary Norskin	1	Charles Haas	1	Andrew Rambaut	1	Hans-Dieter Klenk	2	Martin Meltzer	1	Thorsten Ottlewski	1
Gerhard Kayser	1	David Heymann	2	Angela Spelsberg	1	Heiko Krude	1	Matthias Gruhl	1	Thorsten Wygold	1
Heinz-Wilhelm Selzer	1	Heinz Feldmann	1	Anne Schuchat	1	Heinz Reiniger	1	Matthias Orth	1	Thomas Frieden	1
Jan Dahl	1	Henning Karbach	1	Antonio Lanzavecchia	1	Helmtrud Bisanz	1	Michael Kochen	2	Ulrich Heininger	1
Jens-Michael Schröder	1	Herbert Schmitz	1	Barack Obama	1	Ingegerd Kallings	1	Michael Köllstadt	1	Ulrich Schwabe	1
Margaret Chan	1	Jeremy Farrar	2	Christian Meyer	1	Jakob Cramer	1	Michael Osterholm	1	Virginia Sondia	1
Martin Häusling	1	Jürgen Knobloch	1	Christian Putensen	1	Jeremy Farrar	1	Michele Ginsberg	1	Walter Dowdle	1
Michael Friedrich	1	Klaus Lieb	1	Christoph Hübner	1	Jim Zimmerly	1	Mikhail Matrosovich	1	Wayne Marasco	1
Michael Nielsen	2	Manfred Dietrich	1	Christoph Unger	1	Johannes Löwer	4	Neil Ferguson	2	Wendy Barclay	1
Michael Zasloff	1	Mark Feinberg	1	Damon Choy	1	John Oxford	2	Philipp Rösler	1	Werner Schnappauf	1
Petra Gastmeier	1	Micaela Serafini	1	Daniel Lavanchy	1	Jörg Hacker	3	Reinhard Berner	1	Wolf-Dieter Ludwig	1
Pythagoras	1	Michael Schormann	1	Dietrich Wersich	1	Jürgen Richt	1	Robert Webster	2	Wolfgang Becker-Brüser	2
Renate Künast	1	Noa Freudenthal	1	Donald Henderson	1	Karl Lauterbach	1	Rolf Heckler	3	Wolfgang Schneider-Rathert	1
Sally Davies	1	Peter Jördening	1	Doris Schrage	1	Keiji Fukuda	2	Ruprecht Schmidt-ott	1	Wolfram Hartmann	1
Elke von Grabowski	7	Salome Karwah	8	Eduardo Sada	1	Klaus Stöhr	1	Samuel Ponce de León	1	Yuen Kwok Yung	1
Steven Solomon	1	Stephen Kennedy	1	Frank Ulrich Montgomery	1	Klaus Theo Schröder	1	Stephan Becker	1		
Theodor Mantel	1	Thomas ten Boer	2	Frank von Sonnenburg	1	Margaret Chan	4	Susanne Glasmacher	1		
Thomas Blaha	2	Tino Schwarz	1	Frau Ba*	##	Marie-Paule Kieny	2	Susanne Huggett	1		
Wolfgang Witte	1			Gary Nabel	1	Mark Humphries	1	Susanne Stöcker	1		

Institutionen

Antibiotika (29)		Ebola (18)		Grippe (58)				
Baden-Württembergisches	#	Ärzte ohne Grenzen	1	Abteilung Gesundheitspoli	##	Landesregierung von Ba	##	
Bayerisches Staatsminister	#	Bundespolizeiinspek	1	Arbeitsgemeinschaft Influe	1	London School of Hygie	1	
Bayerisches Staatsminister	#	Deutsche Welthunge	1	Arznei-Telegramm	##	Marburger Institut für V	1	
Berliner Charité	2	Die Zeit	##	Asklepios-Kliniken	1	Marinehospital Stuttgar	##	
Biologische Bundesanstalt	1	Drexel University	1	Bernhard-Nocht-Institut	1	Memorial University of	1	
Bundesamt für Verbrauch	1	Frankfurter Allgemei	1	Bundesamt für Bevölkerun	1	Mexikanische Regierun	##	
Bundestierärztekammer	1	Gesundheitsamt Bre	##	Bundesärztekammer	2	Nationales Referenzzen	1	(--> Nationalen Referenzzentrum für Influenza)
Bundesverband Tierschutz	1	Gesundheitsamt Obe	##	Bundesgesundheitsministe	1	Notfallambulanz am At	##	
CDC	1	Gesundheitsamt Ruc	##	Bundesinnenministerium	1	Novartis	15	
EU	1	Gesundheitsminister	##	Bundesverband der Kinder	1	Operative Intensivmedi	1	(--> Uniklinikum Bonn)
Europäische Arzneimittelb	1	Hamburger Bernhard	1	CDC	10	Paul-Ehrlich-Institut	7	
Europäische Seuchenkontr	#	Universitätsmedizin I	1	CDU	1	Queen Mary Hospital	##	
European Medicines Agen	1	LVR-Klinik Bonn	1	Cochrane Collaboration	2	Regierung der USA	1	(--> US-Regierung)
Gesundheitsamt Oldenbur	1	Merck	2	Dana-Farber Cancer Institu	2	RKI	9	
Gesundheitsministerium v	1	Robert-Koch-Institut	##	Deutsche Gesellschaft für A	1	Robert-Koch-Institut	1	
Grüne im Europäischen Pa	1	Stiftung Wellcome Tr	1	Fachbereich Impfstoffe GS	1	Sächsische Serumwerke	##	
Landestierärztekammer	1	Virologie-Labor	1	Fachblatt Human Vaccines	1	Sanofi	1	(--> Pharmariese Sanofi)
Landesverband praktiziere	#	WHO	11	FDA	3	Smithfield Foods	1	
National Institutes of Heal	1			Hamburger Uni-Klinikum	##	SPD	3	
Regierung (deutsche)	1			Helios Klinikum	1	Stiko	2	
Regierung von Großbritan	#			Justizministerium	1	Transparency Internati	1	(--> Antikorrputionsinitiative)
Robert-Koch-Institut	1			Hongkonger Tourismusbe	##	Uni München	1	
Rockefeller University	1			Imperial College London	1	Universität Hongkong	1	
Tierärztliche Hochschule	1			Influenza Arbeitsgruppe ar	1	Universität Marburg	##	
Universität Freiburg	1			Institut für Biomedizinisch	1	University of Edinburgh	1	
Universität Göttingen	1			Institut für Virologie	2	University of Minnesota	1	
Universitätsklinik Kiel	#			Institut zur Kontrolle von I	1	University of Pittsburgh	1	
Verband der dänischen La	1			Kansas State University	1	WHO	52	*Frau Ba (als ORG identifiziert)
WHO	2			Kinderkrankenhaus Hanno	##	Wissenschaftsmagazin	##	

➔ Übersicht der manuell codierten Akteure aus vorhandenen Daten des Lehrstuhls (abgeleitet aus den SPSS Dateien ‚Aussagenanalyse_all_Ger_USA' + , Artikel_all_Ger_USA')

➔ Als separater Anhang [I] verfügbar: Excel-Arbeitsmappe , Ebola_AR_Grippe_Ergebnisse.xlsx'

➔ Grüne Markierung = Eigenname wurde auch mit *NER*-Verfahren identifiziert

➔ Zahl = Häufigkeit der automatisierten Erkennung des Namens (keine zusammengefasste Angabe der Nennung des vollen Namens + der entsprechenden nachfolgenden Nachnamen-Nennung)

➔ Rote Markierung = Eigenname wurde nicht identifiziert bzw. nicht korrekt als ‚PER' oder ‚ORG'

- Bessere *NER*-Leistung bei Personen verglichen zu Organisationen/Institutionen (unabhängig vom Umfang der Artikel pro Thema)
- Ähnlich gute Erkennungsleistung bei SPIEGEL- und WELT-Artikel
- ORG-Identifikation erschwert, da teilweise Abkürzungen und unterschiedliche Bezeichnungen

29) WELT-Artikel (AR, Ebola, Grippe)

Welt

Personen

Antibiotika (37)				Ebola (60)						Grippe (58)					
Alexander Friedrich	1	Marilyn Roberts	1	Abdullah Ahmed	1	Gabriel Rugalema	1	Martin Lucas	1	Albert Osterhaus	3	John Oxford	1	Peter Sinram	1
Annette Widmann-Mauz	1	Mario Czaja	1	Alfred Palo Conteh	1	Gary Nabel	1	Nina Pham	4	Alexander Kekulé	2	Jörg Debatin	1	Philipp Rösler	2
Axel Kramer	1	Martin Mielke	1	Alinah Oliet Kallon	1	Gordian Schudt	1	Olga Dolnik	1	Androulla Vassiliou	1	Jörg Hacker	1	Ralf Treptow	1
Brian Currie	1	Martin Wernitz	1	Andreas-Christian Tübler	1	Hans-Dieter Klenk	2	Onyebuchi Chukwu	1	Angela Spelsberg	1	José Angel Córdova	1	Regina Kneiding	2
Burkhard Kirchhoff	1	Mel Spigelman	1	Angela Merkel	1	Herbert Schmitz	2	Palo Conteh	1	Anita Tack	##	Julie Hall	1	Reinhard Kurth	2
Chris Hentschel	1	Mischa Moriceau	1	Anthony Banbury	1	Hermann Gröhe	1	Peter Claus Döller	1	Bernd Leidel	1	Karin Wilke	3	René Gottschalk	1
Christian Hanke	1	Norbert Suttorp	1	Ban Ki-Moon	1	Hinrich Sudeck	1	Peter Piot	2	Christoph Unger	2	Katrin Wilke	2	Robert Plum	1
Christine Geffers	1	Otto Cars	1	Barack Obama	3	Jerome Mouton	1	Phil Smith	1	David Byrne	1	Keiji Fukuda	1	Sabine Schirop	##
Elisabeth Meyer	1	Petra Gastmeier	1	Brice de la Vigne	1	Jestina Doe	2	Rajiv Shah	1	Dietrich Wersich	1	Klaus Etzler	1	Sir Francis Drake	##
Glenn Kaatz	1	Reinhard Kurth	1	Bruce Aylward	1	Joachim Gardemann	3	René Gottschalk	1	Elisabeth Neumeier	1	Klaus Stöhr	1	Thijs Kuiken	1
Henning Rüden	1	Renate Künast	2	Christine Lagarde	1	Joanne Liu	1	Robert Davey	1	Fatimah Dawood	1	Klaus Vater	1	Thomas Löscher	3
Joe Cohen	1	Risards Zaleskis	1	Claire Schaffnit-Chatterjee	1	Johanna Wanka	1	Rodolphe Thiébaut	1	Frank Spieth	1	Margaret Chan	1	Thomas Wolfe	1
Jörg Braun	1	Ryland Young	##	David Nabarro	1	John Pryor	1	Rüdiger Langer	1	Frank Ulrich Montgomery	2	Martina Bunge	1	Tom Jefferson	4
Karen Smith	1	Stefan Etgeton	1	David Sanders	1	Joseph Nyumah Boa	1	Stephan Becker	4	Georgius Gouvras	1	Michael Lässig	2	Ulla Schmidt	1
Karl-Max Einhäupl	1	Ulrich Frei	1	Ed Royce	1	Josephine Teah	1	Stephen Kavuma	1	Herbert Schmitz	1	Michael Pfleiderer	2	Ulrich Keil	3
Klaus-Dieter Zastrow	2	Vitaly Vodyanoy	1	Ellen Johnson Sirleaf	1	Julia Duncan-Cassell	1	Tankred Stöbe	1	Ian Wilson	1	Michael Wüstenberg	1	Wolf Dombrowsky	1
Lothar Elling	1	William Gaze	1	Fabian-Carlos Guhl	5	Kaifala Marah	1	Thomas Frieden	1	J. J. Skehell *	1	Milan Brahmbhatt	1	Wolfgang Schimmang	2
Lynn Marks	1			Fadela Chaib	1	Karl Lauterbach	1	Walter Lindner	3	Jan van Lunzen	1	Neil Ferguson	1	Yoshihiro Kawaoka	1
Manuela Zingl	1			Francis Omaswa	1	Klaus Cichutek	1	Wendy Maury	##	Jim Adams	1	Nirundorn Aungtragoolsuk	1		
Margaret Chan	2			Francois Nguessan	1	Marisol Touraine	1	Werner Slenczka	2	Joachim Kühn	1	Peter Cordingley	1	* J.J (identifiziert)	

Institutionen

Antibiotika (35)				Ebola (55)				Grippe (46)				
Asklepios Klinik Hamburg-	1	Universität Washington	1	Ärzte ohne Grenzen	2	Klinik in Bamako	##	Berliner Charité	1	Senatsverwaltung für Gesur	3	
Auburn University	1	Uppsala University	1	Auswärtiger Ausschuss im I	1	liberianische Regieru	##	Berliner Senatsverwaltung	3	Stiko	1	
Berliner Charité	25	Veterans Affairs Medical (	##	BNI	2	London School of Hy	1	Bernhard-Nocht-Institut	1	Technologieunternehmen 3	##	
Bertelsmann-Stiftung	2	Vivantes-Klinikum	1	Bundesforschungsministeri	1	NIAID	1	Brüsseler Generaldirektior	##	ThyssenKrupp Ag	1	
Bundesgesundheitsminist	##	WHO	12	Bundesgesundheitsministe	2	Nationales Ebola-Re	##	Bundesärztekammer	1	Transparency International	1	
Bundesregierung	##	Zentrum für Infektionskra	1	Bundesregierung	##	NIH	1	Bundesamt für Bevölkerun	1	Universität Köln	1	
CDC	1			CDC	5	Non-Profit-Organisa	1	Bundestag	1	Universität Madison	##	
Deutsche Gesellschaft für	2			CDU	3	Paul-Ehrlich-Institut	1	Cochrane Collaboration	1	Universität Münster	1	
Deutsche Gesellschaft für Kra	->(doppelt manuell)			Center for Disease Control	1	Presbyterian Hospita	1	Der Spiegel	2	UKE	3	(Universitätsklinikum Eppend
Deutscher Bauernverbanc	1			CNN	2	Purdue University	1	Erasmus Medical Center R	##	US-Gesundheitsministerium	1	
Fakultät für Biochemie un	##			Deutsche Bundeswehr	4	Regierung der USA	##	EU-Kommission	7	Vereinigung der Oberstudie	1	
Gesundheitsministerium Ch	##			Deutsches Rotes Kreuz	1	Regierung Liberia	##	Gesundheitsministerium	1	Sächsisches Serumwerk Dre	1	
Harvard Universität	1			Ebola-Zentrum Sierra Leon	##	Stadtverwaltung Dal	##	Gesundheitszentrum Atlan	##	Vogelgrippe-Projektgruppe	1	
Helmholtz-Institut für Bio	##			Emerging Markets	1	Street Child	2	GKV	1	Weltbank	4	
Institut für Hygiene der U	1			EU-Kommission	2	Thomas Jefferson U	1	Hamburger Gesundheitsbe	##	Werbellinsee-Grundschule	1	
Institut für Hygiene und U	1	(--> Berliner Vivantes Kliniken)		FAO	2	Tropenklinik Tübinge	1	Imperial College London	##	WHO	37	
Karolinska Institut	1			Frankfurter Gesundheitsam	1	UN	4	Institut für Immunologie	1	Zentrum für hochinfektiöse	1	(und Rheinland-Pfalz)
Medicine for Malaria Ven	1			Gesundheitsämter in Mali	##	Universität von Iowa	1	Lehrgewerkschaft GEW	1			
Montefiore Medical Cente	##			Gesundheitsbehörde	##	Universität von Texa	1	Lessing Gymnasium	1			
Rechtsanwalt	##			Gesundheitsministerium A	##	US Army Medical Re	1	Linke-Fraktion	1			
Regierung von Berlin	##			Gesundheitsministerium Li	##	USAID	1	Londoner St.-Barthikimäus	##			
Robert-Koch-Institut	4			Hamburger Universitätsklin	##	US-Akademie der W	1	Medientitel	##			
Stiftung Warentest	##			Health Presbyterian Hospit	3	Washington Post	1	Münchner Tropeninstitut	##			
Strathclyde University	1			Hochsicherheitslabor Marb	##	WHO	35	Paul-Ehrlich-Institut	2			
Südkorea	##			Impfstoffforschungszentru	##	Wissenschaftszentru	##	RKI	14			
TB Alliance	2			Institut für Virologie der Un	##	Zeitung Daily Monit	##	Robert (?)				
Technische Universität Mü	1			Institut für Virologie Marbu	##			Robert-Koch-Institut	6			
Universität von Seattle	##			IWF	3			Scripps-Forschungsinstitut	1			
Universität von Warwick	1			Johnson & Johnson	1			Senatsgesundheitsverwalt	##			

30) Systeme, Programme und Code-Package-Versionen, die bei der Durchführung der Masterarbeit genutzt wurden:

- Operating System: Windows 10 | 64-Bit Betriebssystem
- Prozessor: AMD Ryzen 3 (3200U) – 2.60 GHz – 8,00 GB RAM
- Python Version Used: Python 3.7.9
- spaCy Version Used: Spacy 2.2
- Stanza Version Used: Stanza 1.1.1
- FLAIR Version Used: FLAIR 0.6.1
- Environment Information: Jupyter Lab
- pdf-file transformation into txt: www.pdf2go.com
- Microsoft Excel Power Query (SSBI)

NLP-Pipeline-Code:

Der für die *Pipelines* genutzte *Python*-Programmcode wurde nicht eigenständig erarbeitet, sondern basierend auf den von Nikolai Promies (Mitarbeiter am Lehrstuhl ‚Wissenschafts-kommunikation in digitalen Medien') verfassten und zur Verfügung gestellten *Jupiter Note-books* angepasst und für die drei *NER*-Analysen genutzt.

Diese Dateien sind als separater Anhang (A) dieser Arbeit beigelegt und können zur Repro-duktion und Kontrolle der Analyseergebnisse genutzt werden.

→ 3 *HTML*- und *ipynb*-Dateien für die jeweilige *spaCy*-, *Stanza*- und *FLAIR-Pipeline*