SCM
Studies in Communication and Media

FULL PAPER

Validierung von NER-Verfahren zur automatisierten Identifikation von Akteuren in deutschsprachigen journalistischen Texten

Validation of NER methods for the automated identification of actors in German journalistic texts

Cecilia Buz, Nikolai Promies, Sarah Kohler & Markus Lehmkuhl

Cecilia Buz (M. A.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunikation, Englerstraße 2, D-76131 Karlsruhe. Contact: cecilia.buz(at)partner.kit.edu.

Nikolai Promies (M. A.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunikation, Englerstraße 2, D-76131 Karlsruhe. Contact: nikolai.promies(at)kit.edu. ORCID: https://orcid.org/0000-0002-4804-4155

Sarah Kohler (Dr.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunikation, Englerstraße 2, D-76131 Karlsruhe. Contact: sarah.kohler(at)kit.edu. ORCID: https://orcid.org/0000-0002-3548-010X

Markus Lehmkuhl (Prof. Dr.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunıkation, Englerstraße 2, D-76131 Karlsruhe. Contact: markus.lehmkuhl(at)kit.edu. ORCID: https://orcid.org/0000-0001-8295-6548

FULL PAPER

Validierung von NER-Verfahren zur automatisierten Identifikation von Akteuren in deutschsprachigen journalistischen Texten

Validation of NER methods for the automated identification of actors in German journalistic texts

Cecilia Buz, Nikolai Promies, Sarah Kohler & Markus Lehmkuhl

Zusammenfassung: Dieser Beitrag befasst sich mit der Validierung von Named Entity Recognition (NER), einem Verfahren, das als Teilschritt der Inhaltsanalyse von umfangreichen Textdaten eingesetzt werden kann und auf die automatisierte Identifikation und Extraktion von Eigennamen (Personen, Organisationen, Orte) in Texten spezialisiert ist. Für diesen Zweck werden oft frei verfügbare NER-Softwarepakete verwendet, die mit spezifischen Textdaten trainiert und optimiert wurden. Dadurch ist jedoch ungewiss, ob diese NER-Pakete bei der Analyse von unbekannten journalistischen Nachrichtentexten richtige und präzise Ergebnisse liefern können. Um dies zu evaluieren, wurden drei in der Programmiersprache Python implementierte NER-Codepackages gegenübergestellt und die Ergebnisse der automatisierten Analyse mit den Ergebnissen einer manuellen Inhaltsanalyse derselben journalistischen Textdaten verglichen. Ziel ist damit, die Eignung und Güte verschiedener NER-Softwarepakete für die Identifikation von Akteuren zu prüfen, denn obwohl in der Kommunikationswissenschaft vermehrt automatisierte Verfahren eingesetzt werden, mangelt es an Studien, die die Validität der erhaltenen Ergebnisse bewerten. Die Ergebnisse zeigen eine hohe Übereinstimmung zwischen den händisch erhobenen und den automatisiert identifizierten Personennamen, lediglich bei der automatisierten Identifikation von Organisationsnamen ist die Übereinstimmungsquote mit den manuellen Codierungen geringer.

Schlagwörter: Automatisierte Inhaltsanalyse, Named Entity Recognition, Akteursanalyse, Validierung.

Abstract: The aim of this paper is the validation of a method that can be used to automate a sub-step in the content analysis of text data. The method under investigation is called Named Entity Recognition (NER) and is specialized in the automated identification and extraction of proper names (persons, organizations, places) in texts. In communication science, automated methods are increasingly used for the analysis of large amounts of text, but there are hardly any studies dealing with the validity of the automatically obtained results. The aim of the study presented here 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 jour-

nalistic news 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 personal names. For the automated identification of the names of organisations, the match rate with the manual codings appears to be lower.

Keywords: Automated content analysis, computational methods, named entity recognition, validation.

1. Einleitung

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 (Domahidi et al., 2019, S. 3877). Dort steht die Nutzung automatisierter Verfahren im Mittelpunkt, um unter anderem die Inhalte großer Textsammlungen mittels Algorithmen zu analysieren, darin Zusammenhänge und Muster zu identifizieren und diese Datenstrukturen zu visualisieren (Grimmer & Stewart, 2013, S. 267). Dies ermöglicht die systematische Auswertung von beispielsweise Nachrichtenbeiträgen oder nutzergenerierten Online-Inhalten, wodurch Erkenntnisse über die Medieninhalte sowie Meinungsbildungsprozesse in der Gesellschaft erlangt werden können (Strippel et al., 2018, S. 16). Da permanent eine kaum zu überblickende Masse an digitalen Inhalten entsteht, steigt die Relevanz von automatisierten Methoden, (Sommer et al., 2014, S. 14), um diese Inhalte zu erschließen. Insbesondere bei quantitativen Analysen von sehr umfangreichen Textkorpora ist eine manuelle Inhaltsanalyse kaum mehr möglich. Umso mehr bietet sich der Einsatz automatisierter Verfahren an, setzt jedoch selektive Informatikkenntnisse voraus (Jannidis, 2017, S. 95).

Die Anzahl an Ausarbeitungen, in denen automatisierte Verfahren eingesetzt werden, wächst zwar, dennoch gehört die Anwendung solcher Verfahren in der Kommunikationswissenschaft noch nicht zum Ausbildungsstandard (Strippel et al., 2018, S. 18) und es mangelt an Studien, die dezidiert für einzelne Verfahren die Qualitätsanforderungen oder Validität benennen (Niekler, 2018, S. 179). Mit zunehmender Verbreitung und Bedeutung solcher Verfahren steigt allerdings auch die Notwendigkeit der methodologischen Diskussion über ihre Anwendung und Validierung (Strippel et al., 2018, S. 18). Der vorliegende Aufsatz setzt an dieser Stelle an und prüft ein Verfahren, mit dem Personen-, Orts- und Organisationsnamen aus journalistischen Texten automatisiert extrahiert werden können, die sogenannte Named Entity Recognition (NER). Ziel ist es, die Güte der automatisiert erhaltenen Ergebnisse von verschiedenen NER-Paketen zu bestimmen und dadurch das Verfahren zu validieren, um dessen Einsatz für künftige Analysen deutschsprachiger Textdaten weiter zu etablieren.[1]

Die automatisierte Identifikation von Eigennamen ist für die Kommunikationswissenschaft vor allem vor dem Hintergrund relevant, dass die Untersuchung der

1 Ähnliche Arbeiten gibt es bereits für Verfahren wie Sentiment Analysis (Boukes et al., 2020; van Atteveldt et al., 2021) oder Topic Modeling (Maier et al., 2018).

Präsenz von Personen, Unternehmen oder Organisationen in (medialer) Kommunikation ein wichtiges Anwendungsgebiet von Inhaltsanalysen ist (Schneider, 2014, S. 41). Viele Untersuchungen befassen sich mit der Akteursvielfalt medialer Diskurse sowie den Fragen, wann und wie häufig spezifische Akteure[2] erwähnt werden (bspw. Boberg et al., 2020, S. 12; Burggraaff & Trilling, 2020, S. 121; Eisenegger et al., 2020, S. 10; Niekler, 2018, S. 159). Dabei wird nicht nur analysiert, wie oft bestimmte Akteure in der Berichterstattung vertreten sind, sondern auch, wie sich dies zwischen verschiedenen Medien unterscheidet oder im Zeitverlauf entwickelt. Dadurch können Aussagen über die Sichtbarkeit und Relevanz bestimmter Personen, Unternehmen oder Institutionen getroffen, ebenso wie Veränderungen in der Akteurskonstellation innerhalb der Berichterstattung erkannt werden (Strippel et al., 2018, S. 17). Auch wird das Spektrum der genannten Akteure mitsamt ihren Äußerungen teilweise als Indikator für eine gehaltvolle Berichterstattung gewertet (Schweiger, 2017, S. 32), wenn für einen gelungenen Meinungsbildungsprozess der Gesellschaft eine Berichterstattung mit einer Vielzahl an Standpunkten und entgegengesetzten Sichtweisen als wertvoll angesehen wird.

2. Automatisierung inhaltsanalytischer Untersuchungen

Traditionell ist eine manuelle Inhaltsanalyse die Methode der Wahl für die Erhebung von Akteuren aus Medientexten. Seit ungefähr zehn Jahren halten jedoch bedingt durch verschiedene Entwicklungen automatisierte Verfahren zunehmend Einzug in die Kommunikationswissenschaft. Der entscheidende Faktor kann mit dem Stichwort ‚Big Data' benannt werden: die Verfügbarkeit einer gewaltigen Menge von digitalen Spurendaten menschlichen Verhaltens und insbesondere auch von Kommunikation (Lewis et al., 2013; Parks, 2014). Diese Datenbestände bieten neue oder erweiterte Möglichkeiten, menschliche Kommunikation zu erforschen und besser zu verstehen, indem beispielsweise neue Fragestellungen eröffnet oder bestehende aus neuen Perspektiven betrachtet werden können (van Atteveldt & Peng, 2018, S. 82–84). Sie machen es aber auch erforderlich, dass sich die Sozialwissenschaften neuen Ansätzen und Methoden zuwenden. Um auf Veränderungen der sozialen (und medialen) Umwelt reagieren zu können, die unter Konzepten wie ‚Datafizierung' diskutiert werden (Hepp, 2016, S. 229–230), sind Methoden aus Gebieten wie der angewandten Informatik notwendig, die besser an die neuen Datenformen und -mengen angepasst sind. Praktisch kommt hinzu, dass komplizierte algorithmische Verfahren inzwischen auch für Nicht-Expert:innen deutlich leichter zugänglich (durch Open-Source-Implementierungen in verschiedenen Programmiersprachen) und anwendbar sind (auch auf ‚normalen' Endgeräten oder durch Cloud Computing) (van Atteveldt & Peng, 2018, S. 81).

Die Folge dieser Entwicklungen ist eine starke Zunahme der Anwendung automatisierter Verfahren in der Kommunikationswissenschaft, die sich unter

2 Nach längerem Abwägen haben sich die Autor:innen dazu entschieden, den Begriff ‚Akteure' nicht zu gendern, da die Bezeichnung hier nicht nur individuelle Personen umfasst, sondern sich auch auf Institutionen bzw. korporative und kollektive Akteure bezieht.

anderem in einer Reihe von Special Issues einschlägiger Fachzeitschriften zeigt (Domahidi et al., 2019; Karlsson & Sjøvaag, 2016; van Atteveldt & Peng, 2018) oder sogar in der Gründung neuer Zeitschriften (van Atteveldt et al., 2019). Die Vorteile automatisierter Verfahren insbesondere für Inhaltsanalysen liegen auf der Hand: Codierentscheidungen, die sonst aufwändig durch menschliche Codierer:innen getroffen werden mussten, können von Computerprogrammen übernommen werden, wodurch insbesondere große Textmengen deutlich schneller analysiert werden können. Neben der erhöhten Effizienz und Kapazität weist eine automatisierte Inhaltsanalyse 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, S. 200). Grenzen haben automatisierte Methoden unter anderem in der Erhebung abstrakter Konstrukte, bei denen menschliche Codierer:innen weiterhin überlegen sind, oder auch durch den teilweise sehr großen Aufwand bei der Konzeption, Implementierung und Validierung der Verfahren. Diese Vor- und Nachteile führen dazu, dass automatisierte Verfahren in erster Linie als Ergänzung klassischer Methoden verwendet werden (bspw. Grimmer & Stewart, 2013, S. 270).

2.1 Unterschiedliche Typen automatisierter Inhaltsanalysen

Bei der computergestützten Durchführung von Inhaltsanalysen steht inzwischen ein ganzes ‚toolkit' von Verfahren zur Auswahl (Boumans & Trilling, 2016), die sich für verschiedene Anwendungen eignen. Gängig ist in der Literatur die Unterscheidung in drei Gruppen von Verfahren: diktionär- oder regelbasierte; trainierte und deskriptive bzw. explorative Verfahren (Boumans & Trilling, 2016; Günther & Quandt, 2016).

Unter diktionärbasierten Verfahren wird die automatisierte Informationsextraktion mittels Schlagwörtern oder Wortlisten verstanden (Scharkow, 2012, S. 60). Bei diesem deduktiven Vorgehen handelt es sich um einen einfachen Vergleich von Zeichen oder Suchbegriffen, die im Vorfeld als maschinenlesbares Wörterbuch definiert werden (Wettstein, 2014, S. 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 (Züll & Mohler, 2001, S. 4). Da alle Kategorien samt Ausprägungen und Codieranweisungen vor der Analyse definiert werden müssen, ist der Aufwand im Vorfeld besonders hoch und die Einsatzmöglichkeiten und die Anwendungstiefe sind recht begrenzt (Stoll et al., 2020, S. 113).

Bei trainierten Verfahren werden Machine-Learning-Algorithmen eingesetzt, die anhand von speziell angefertigten Trainingsdokumenten mit richtigen Klassifikationen eigenständig die Codier-Zuordnungen und -Regeln erlernen. Dieser Prozess ist nicht rein induktiv, da auch hier im Vorfeld manueller Aufwand nötig ist, um einen Trainingsdatensatz mit richtigen Zuordnungen und maschinenlesbaren Annotationen zu erstellen. In diesen annotierten Beispieltexten werden von dem Algorithmus statistische Zusammenhänge und Strukturen erkannt, aus denen ein Vorhersagemodell erstellt wird, welches dann auf neue Texte angewandt werden kann (Kelm et al., 2020). Mit jedem zusätzlichen Beispiel in dem Trainingsdaten-

satz kann der Algorithmus dazulernen und seine Leistung optimieren (Augenstein et al., 2017, S. 69).

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 (van der Meer, 2016, S. 959). Unüberwachte Verfahren erfordern den geringsten Aufwand im Vorfeld der Analyse, da keine manuellen Regelspezifikationen für ihren Einsatz notwendig sind (Rössler, 2017, S. 196). Im Gegensatz zu den wörterbuchbasierten und überwacht trainierten Ansätzen werden hierbei Muster und Wortcluster in einem Textdatensatz mittels unbeaufsichtigtem maschinellem Lernen identifiziert. Statt nach vordefinierten Kategorien zu suchen, werden durch den Algorithmus eigene Zuordnungen vorgenommen. Ein in der Kommunikationswissenschaft häufig verwendetes unüberwachtes Verfahren ist die Themenklassifizierung mittels Latent Dirichlet Allocation (LDA; Blei et al., 2003; für einen Überblick zur Anwendung in der Kommunikationswissenschaft Maier et al., 2018).[3]

2.2 Named Entity Recognition als Verfahren zur Identifikation von Eigennamen

Unser Interesse gilt einem speziellen Verfahren aus dem Bereich automatisierter Inhaltsanalysen, der sogenannten Named Entity Recognition (NER). NER gehört zu den Verfahren der maschinellen Verarbeitung menschlicher Sprache (Natural Language Processing – NLP) und wird zur Extraktion von Informationen aus unstrukturierten Texten angewandt, mit dem Ziel benannte ‚reale Objekte' zu identifizieren, die aus Eigennamen bestehen (Marrero et al., 2013, S. 482). Dabei kann es sich beispielsweise um die Erkennung der Namen von Personen, Orten, Unternehmen oder Institutionen handeln (Schneider, 2014, S. 41). Beim Einsatz eines NER-Verfahrens werden die Eigennamen in einem Text identifiziert und einer bestimmten NE-Klasse zugeordnet (Eftimov et al., 2017, S. 3). Die vier gängigen Kennzeichnungen umfassen die Klassen PER, ORG, LOC und MISC für die Kategorien Person, Organisation, Ort und Sonstiges (Faruqi & Padó, 2010, S. 130).

Theoretisch kann NER als regelbasiertes, überwachtes oder unüberwachtes Verfahren angewandt werden. Regelbasiert könnte man beispielsweise versuchen, sich bestimmte Eigenschaften von Texten zur Erkennung von Akteuren zunutze zu machen. Über sogenannte Regular Expressions könnte man nach je zwei nacheinander stehenden großgeschriebenen Begriffen in einem Text suchen, um so in zahlreichen Sprachen alle vorkommenden Vor- und Nachnamen oder Titel und

3 LDA ist eine Form eines sogenannten Topic Models, dabei handelt es sich um „algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents." (Blei, 2012, S. 78) LDA leitet die in einem Textkorpus vorkommende Themenstruktur induktiv aus dem (gemeinsamen) Vorkommen von Worten in den einzelnen Texten ab. Neben LDA gibt es inzwischen eine Vielzahl weiterer Varianten von Topic Models, die unter anderem unterschiedliche statistische Modelle verwenden, bei der Ermittlung der Themenstruktur Metadaten der einzelnen Dokumente einbeziehen oder modellieren, dass sich Themen über die Zeit verändern (Blei, 2012, S. 82–84).

Zunamen zu ermitteln (ein Beispiel für eine solche Anwendung findet sich in Eisenegger et al., 2020). Unüberwacht kann ein explorativer Algorithmus basierend auf wenigen Beispielnamen nach Begriffen suchen, die in ähnlichen strukturellen Kontexten vorkommen. Dieser Lernprozess wird dann erneut auf die neu gefundenen Beispiele angewendet, um neue relevante Zusammenhänge zu entdecken (Nadeau & Sekine, 2007, S. 5).

Praktisch wird Named Entity Recognition jedoch üblicherweise als überwachtes Verfahren umgesetzt. Basierend auf Textkorpora, in denen menschliche Codierer:innen verschiedene Arten von Eigennamen identifiziert und klassifiziert haben, können mit Machine-Learning-Verfahren Modelle trainiert werden, die dann in unbekannten Texten Eigennamen erkennen. Die Erstellung eines Trainingskorpus ist relativ aufwendig, weswegen sich für die meisten Analysen bereits vortrainierte Verfahren anbieten. Es gibt inzwischen eine Vielzahl von frei zugänglichen Code-Bibliotheken mit vortrainierten Modellen, die mit wenig Aufwand für eigene Analysen genutzt werden können.

Diese Modelle unterscheiden sich zum einen in den Textdaten, mit denen die Verfahren trainiert werden, und zum anderen in den Machine-Learning-Algorithmen, die für das Training verwendet werden. Die Zusammensetzung der für das Training verwendeten Texte wirkt sich erheblich auf die Leistung der damit trainierten Algorithmen aus (Augenstein et al., 2017, S. 76). Beispielsweise kann dadurch, dass in einem Textkorpus Personen-, Organisationen- und Ortsnamen in unterschiedlich hohen Anteilen vorkommen, auch die Klassifikationsleistung in den jeweiligen NE-Klassen uneinheitlich ausfallen (Maynard et al., 2016, S. 35). Bei den verwendeten Algorithmen sind inzwischen Deep-Learning-Verfahren der Standard, die auf künstlichen neuronalen Netzen basieren. Insbesondere Convolutional und Recurrent Neural Networks sowie sogenannte Transformer Architekturen gelten heutzutage in der Informatik als state-of-the-art Werkzeug für die Textverarbeitung und sind aktuell das Kernforschungsgebiet für Lösungsansätze der Computerlinguistik (Jurafsky & Martin, 2021, S. 176; Stoll et al., 2020, S. 130).

3. Methode

Ziel dieses Aufsatzes ist die Validierung von vortrainierten NER-Verfahren, um zu beurteilen, ob sie sich eignen, um präzise und vollständig Akteure aus journalistischen Texten zu extrahieren. Dafür werden drei verschiedene NER-Codepackages ausgewählt, getestet und ihre ausgegebenen Ergebnisse mit denen einer manuellen Inhaltsanalyse verglichen.

Bei manuellen Inhaltsanalysen werden in den Sozialwissenschaften Reliabilitätstests zur Qualitätssicherung und Sicherstellung der Güte der durchgeführten Messungen eingesetzt (Dumm & Niekler, 2014, S. 21). Beim Einsatz von automatisierten Verfahren zur Extraktion von Textinhalten erfolgt diese Gütebeurteilung typischerweise auf andere Art und Weise, denn auf technischer Ebene ist ein automatisiertes Verfahren vollständig zuverlässig und reproduzierbar, während auf inhaltlicher Ebene geprüft werden muss, ob die erhaltenen Ergebnisse gültig sind (Scharkow, 2013, S. 290).

In den Computational Social Sciences fehlen gegenwärtig universelle Richtwerte zur Bewertung von automatisierten Verfahren (Niemann-Lenz et al., 2019, S. 3892), doch in der Forschungsliteratur wird empfohlen, die automatisiert erhaltenen Ergebnisse mit manuell durchgeführten Erhebungen zu vergleichen (Burggraaff & Trilling, 2020, S. 125; Grimmer & Stewart, 2013, S. 271; Schwotzer, 2014, S. 55). Als Gütekriterien werden dabei die Leistungskennzahlen *Precision* und *Recall* genutzt, welche im Fachbereich der angewandten Informatik für die Leistungsevaluation von NLP-Aufgaben verwendet werden (Derczynski, 2016, S. 262; Dumm & Niekler, 2014, S. 21). Die Precision gibt dabei die ‚Exaktheit' oder ‚Verlässlichkeit' des Verfahrens an, während der Recall die ‚Vollständigkeit' der Ergebnisse bewertet (Ketschik et al., 2020, S. 204; Rössler, 2007, S. 92). Precision und Recall stehen dabei in einer Wechselbeziehung: Bei der Optimierung der Precision eines algorithmischen Verfahrens, geht dies auf Kosten des Recalls und vice versa. Zusätzlich ist beim Vergleich von Machine-Learning-Verfahren ein drittes Gütemaß gängig, der sogenannte F-Score. Dieser ist das harmonische Mittel von Precision und Recall, welches in einem Wertebereich zwischen 0 und 1 liegt (je höher der Wert, desto höher die erzielte Leistung des Verfahrens).

3.1 Auswahl der NER-Codepackages

Voraussetzung für die Durchführung jeglicher NLP-Aufgaben ist die Einrichtung eines lauffähigen Programms, welches am Anfang die Textdaten einliest, alle anschließenden Textverarbeitungsschritte umfasst und am Ende die Ergebnisse der Textanalyse beispielsweise in Form einer Liste oder als Datei ausgibt. Der für die hier durchgeführte NER-Analyse verwendete Programmcode wurde in Form eines Jupyter Notebooks aufgesetzt und wird anderen Forschenden als Open Source zur Verfügung gestellt.[4]

In den meisten Fällen werden dafür vorgefertigte Softwarepakete verwendet, in denen bereits verschiedene Funktionalitäten implementiert wurden (Lane et al., 2019, S. 4). Für das Gebiet der natürlichen Sprachverarbeitung gilt Python – auch in der Kommunikationswissenschaft – als führende Anwendungssprache und verspricht, leicht erlernbar und gleichzeitig leistungsfähig einsetzbar zu sein (u. a. Burggraaff & Trilling, 2020, S. 118; Lewis et al., 2013, S. 48; Stoll et al., 2020, S. 131). Für diese Analyse wurden die NER-Packages[5] der Python-Bibliotheken spaCy (Honnibal et al., 2020), Stanza (Qi et al., 2020) und FLAIR (Akbik et al., 2019) ausgewählt, weil diese in verschiedenen aktuellen Publikationen zu den überlegenen Verfahren bei der Erkennung von Eigennamen gezählt werden (bspw. Lane et al., 2019, S. 353; Shelar et al., 2020, S. 324; Qi et al., 2020, S. 6). Auf den Plattformen, auf denen die Codepackages der jeweiligen Bibliotheken veröffentlicht werden, sind aktuelle Angaben zu ihren höchsterzielten Werten bei der Ana-

4 Jupyter Notebook hier verfügbar: https://gitlab.com/wisskomm-in-digitalen-medien/validierung-ner-public

5 spaCy: https://spacy.io/usage/linguistic-features#named-entities-101
Stanza: https://stanfordnlp.github.io/stanza/
FLAIR: https://github.com/flairNLP/flair

lyse deutschsprachiger Texte vermerkt. Das NER-Codepackage von Stanza wird mit einem F-Score von 0,85 ausgewiesen (Qi et al., 2020, S. 6), spaCy soll F-Scores von 0,86 erreichen (Honnibal et al., 2020, o. S.) und FLAIR wirbt mit einem Wert von 0,88 (Akbik et al., 2018, S. 1645).

Die drei Bibliotheken sind frei verfügbar und können ohne tiefergehende Informatik- oder Machine-Learning-Kenntnisse verwendet werden.[6] Die gewählten NER-Codepackages setzen verschiedene Ausführungen neuronaler Netze ein: FLAIR und Stanza verwenden sogenannte contextual string embeddings zur Repräsentation der Texte und zum Labeling ein sogenanntes Bi-LSTM, ein bidirectional Recurrent Neural Network mit Long Short-Term Memory (Akbik et al., 2018, S. 1638–1642; Qi et al., 2020, S. 3). SpaCy setzt zum Labeling eine andere Variante eines neuronalen Netzes ein, ein Convolutional Neural Network (CNN) und Wortvektoren, die auf Bloom Embeddings basieren (Honnibal o.J.).[7]

Das ausgewählte Modell von FLAIR wurde mit dem Textkorpus ‚CoNLL03' - bestehend aus Artikeln der Frankfurter Rundschau von 1992 (Tjong Kim Sang & Meulder, 2003, S. 143) - trainiert (Akbik et al., 2018, S. 1645), das in Stanza gewählte Modell wiederum basiert auf dem Textkorpus ‚GermEval14', der Wikipedia-Artikel und Online-Zeitungsnachrichten als Trainingsdaten nutzt (Qi et al., 2020, S. 5). Auch der Algorithmus des NER-Modells von spaCy wurde mit einem Wikipedia-Datensatz trainiert und zusätzlich mit dem TIGER Corpus, der sich aus Artikeln der Frankfurter Rundschau von 1995-1997 zusammensetzt (Dipper & Kübler, 2017, S. 601).

3.2 Auswahl des Textkorpus und Vergleichsdatensatz der Analyse

Die Grundlage für die Evaluation der drei NER-Codepackages stellen 887 Online- sowie digitalisierte Print-Beiträge des SPIEGEL, der Deutschen Presseagentur (dpa) und der WELT aus dem Zeitraum von Januar bis Juni 2020 dar. Der Datensatz wurde im Rahmen einer manuellen Inhaltsanalyse der Corona-Berichterstattung deutscher Medientitel zusammengestellt (Leidecker-Sandmann et al., 2021). In dieser Inhaltsanalyse wurde unter anderem ermittelt, welche individuellen, institutionellen oder generischen Akteure in der Berichterstattung zu Wort kommen. Im Rahmen der Analyse wurde ein Reliabilitätstest mit 20 Artikeln bzw. 73 Aussagen durchgeführt, an dem die drei Codierer:innen teilnahmen, die die Analyse

6 Neben den oben genannten drei ausgewählten Codepackages können für NER-Analysen auch andere Python-Bibliotheken verwendet werden wie zum Beispiel Stanford CoreNLP (https://stanfordnlp.github.io/CoreNLP), AllenNLP (https://allennlp.org), NERsuite (https://nersuite.nlplab.org), NLTK (https://www.nltk.org), LingPipe (http://www.alias-i.com/lingpipe) oder Polyglot (https://polyglot.readthedocs.io/en/stable/NamedEntityRecognition.html). Auch in der Programmiersprache R sind verschiedene NER-Packages verfügbar wie z. B.: spacyR (https://spacy.io/universe/project/spacyr), coreNLP (https://www.rdocumentation.org/packages/coreNLP) oder openNLP (https://cran.r-project.org/package=openNLP). Darüber hinaus existieren für NER auch Stand-alone-Tools wie TXTWerk (https://www.txtwerk.de), Dandelion (https://dandelion.eu) oder WebLicht (https://weblicht.sfs.uni-tuebingen.de) (siehe auch Gaus, 2018).

7 Auf die Details und Unterschiede der verschiedenen Arten neuronaler Netze einzugehen ist im Rahmen dieser Arbeit nicht möglich. Für eine tiefergehende Auseinandersetzung mit den zugrundeliegenden Techniken verweisen wir auf einführende Literatur aus der Informatik oder verwandten Disziplinen (bspw. Di Franco & Santurro, 2021; LeCun et al., 2015; Schmidhuber, 2015).

durchführten. Zur Ermittlung der Reliabilität bei der Erkennung von Akteuren wurde für alle identifizierten Eigennamen verglichen, ob diese von allen Codierer:innen gleich identifiziert wurden. Dabei wurde sehr übereinstimmend codiert (Übereinstimmung nach Holsti 0,94; Krippendorffs Alpha 0,87), was dafür spricht, dass sich die händisch erhobenen Daten gut als Vergleichsmaßstab anbieten (weitere Details zum Reliabilitätstest finden sich in Leidecker-Sandmann et al., 2021).

Das Codebuch der manuellen Inhaltsanalyse schrieb vor, dass nur solche Akteure zu codieren sind, deren Aussagen in der Berichterstattung direkt oder indirekt zitiert werden. Es wurden daher manuell nicht alle Akteure codiert, die in der Berichterstattung vorkamen.[8] Ein komplett deckungsgleicher Vergleich der manuell und automatisiert erhobenen Akteure ist somit nicht gegeben, da der NER-Algorithmus darauf trainiert ist, alle Eigennamen in den Texten zu identifizieren, unabhängig davon, ob die Personen oder Institutionen direkt oder indirekt Aussagen tätigen. Das hat zur Folge, dass mittels der NER-Verfahren mehr Eigennamen (korrekt) extrahiert werden, als bei der Inhaltsanalyse mit menschlichen Codierer:innen. Hinzu kommt, dass in der händischen Codierung Personen oder Organisationen, die mehrmals im selben Artikel zu Wort kamen, nur einmal gezählt wurden. Diese Komprimierung auf Artikelebene musste beim Einsatz eines automatisierten Verfahrens als separater Schritt nachträglich erfolgen. Dennoch ist von Relevanz, ob die händisch erfassten Personen und Organisationen grundsätzlich auch automatisiert identifiziert werden können und mit welchem Paket dies am besten gelingt. Der Datensatz der manuellen Codierung ist somit der Referenzwert, mit dem die Vollständigkeit der automatisiert extrahierten Eigennamen berechnet wird.

3.3 Qualitätskriterien zur Messung der Güte

Um die Leistungskennzahlen zu berechnen und damit die automatisiert extrahierten Ergebnisse zu beurteilen, werden meist manuell erhobene Vergleichsdaten als Bewertungsmaßstab für ‚richtige' oder ‚falsche' Ergebnisse genutzt (Schwotzer, 2014, S. 55). Bei der Auswertung werden dann entsprechend die automatisiert extrahierten Ergebnisse, die manuell nicht erhoben wurden, als False Positives betitelt. Da die für diese Studie verfügbaren, manuell selektierten Akteure nicht alle vorhandenen Personen- und Organisationsnamen des Datensatzes abdecken, musste mit einer anderen, weiteren Definition der False Positives gearbeitet werden. Es wurden alle extrahierten Eigennamen manuell geprüft und inhaltlich beurteilt, ob es sich bei dem ausgegebenen Ergebnis tatsächlich um den Eigennamen einer Person oder Institution handelt. Dazu wurden die Richtlinien für die automatisierte Erkennung von deutschsprachigen Eigennamen hinzugezogen (Beni-

8 Codieranweisungen des Codebuchs im Wortlaut: „Akteure sind Individuen oder Organisationen, die in einem Artikel direkt oder indirekt zu Wort kommen. […] Akteure, die nicht direkt oder indirekt zitiert werden, werden auch nicht codiert. Wenn es etwa heißt, Obama hätte Hilfstruppen geschickt oder Anan hätte jemanden ernannt, werden die Akteure nicht codiert. Akteure werden nur dann codiert, wenn ihre Erwähnung im Zusammenhang steht mit einer Wortmeldung dieser Akteure."

kova et al., 2014). Die Precision ergibt sich als Anteil der richtig erkannten Eigenamen an allen identifizierten Eigennamen (bspw.: Zahl der richtig erkannten Akteure/Zahl aller erkannten Akteure).

Zusätzlich wurden bei der Berechnung der Precision die Begriffskombinationen, die fälschlich als Personen oder Organisationen eingestuft wurden (z. B. ‚April-Miete' oder ‚Quarantäne'), händisch genauer betrachtet, um nach Mustern in den Fehlern zu suchen. Diese Muster lassen Rückschlüsse darauf zu, wie die einzelnen Verfahren bei der Erkennung der Eigennamen vorgehen und wo dadurch bedingt ihre Schwächen liegen (siehe 4.2).

Für eine allumfassende Beurteilung der drei Verfahren reicht es nicht aus, zu untersuchen, ob es sich bei den ermittelten Wörtern tatsächlich 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 die Vollständigkeit (Recall) der Ergebnisse zu evaluieren, wurde daher 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. Der Recall ergibt sich demnach aus der Zahl der richtig erkannten Eigennamen geteilt durch alle in den Texten manuell codierten Eigennamen. Hierzu wurden auf Artikelebene alle automatisiert extrahierten Akteure mit den manuell codierten Akteuren abgeglichen und bei Nichtübereinstimmung ebenfalls tiefergehend untersucht, ob bestimmte Gründe dafür erkennbar waren (siehe 4.3).

4. Ergebnisse

Für die Validierung der getesteten NER-Verfahren wird neben der Ermittlung der Güte der automatisiert erhaltenen Eigennamen, auch die Verarbeitungszeit und der generelle Umfang der extrahierten Ergebnisse der drei angewandten Codepackages gegenübergestellt. Beides spielt eine Rolle für die praktische Anwendung der verschiedenen Pakete als Ersatz oder Ergänzung manueller Analysen.

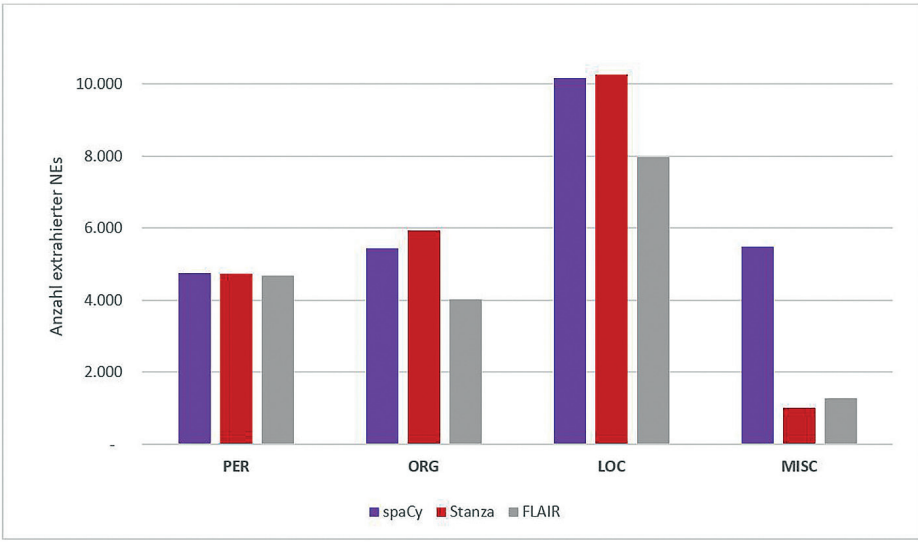
4.1 Verarbeitungsgeschwindigkeit und Zahl identifizierter Eigennamen

Der Prozess des Identifizierens und Klassifizierens der Eigennamen in dem Datensatz mit n = 887 Corona-Nachrichtenartikeln dauert je nach genutztem NER-Codepackage unterschiedlich lange. Alle drei NER-Analysen wurden an demselben Endgerät mit dem zentralen Prozessor (CPU) durchgeführt, um eine Vergleichbarkeit der Verarbeitungsgeschwindigkeit zu gewährleisten.[9] Während spaCy für einen Durchlauf des gesamten Datensatzes nur drei Minuten in Anspruch nimmt, benötigt Stanza 106 Minuten. Verglichen dazu braucht FLAIR mit

9 Um die Performanz der getesteten Pakete besser vergleichen zu können werden hier weitere Spezifikationen des genutzten Computersystems aufgeführt: AMD Ryzen 3 Prozessor (3200U Radeon Vega Mobile Gfx, 2600 MHz, 2 Kernen und 4 logischen Prozessoren) mit 8,00 GB RAM und Windows 10 mit 64-Bit Betriebssystem.

mehr als 2,5 Stunden am längsten, um den Datensatz zu verarbeiten. Auffällig ist hierbei, dass die benötigte Verarbeitungszeit konträr zum Umfang der ausgegebenen Ergebnisse ist. Das NER-Verfahren von FLAIR mit der längsten Verarbeitungsdauer liefert die geringste Anzahl an Ergebnissen (ca. 18.000 Named Entities), während spaCy mit der kürzesten Dauer die größte Menge an Eigennamen markiert (ca. 25.000 NEs). Stanza liegt mit knapp 22.000 Eigennamen im Mittelfeld. Dabei handelt es sich um die ungeprüften und unbereinigten Ergebnisse. Abbildung 1 zeigt diese absolute Anzahl der identifizierten Eigennamen pro NE-Klasse. 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.

Abbildung. 1. Unbereinigte absolute Anzahl der NEs pro Klasse im Vergleich



Auf den ersten Blick kann nicht beurteilt werden, was das Identifizieren vieler oder weniger Eigennamen über die Leistung des jeweiligen NER-Verfahrens aussagt. Die Ausgabe von vielen Ergebnissen kann viele False Positives beinhalten, während wenige identifizierte NEs auf eine schlechtere Erkennungsleistung oder aber auf ein präziseres Verfahren hinweisen können.

Bei der Betrachtung der Akteure und Organisationen, die am häufigsten durch die drei Verfahren identifiziert wurden, zeigen sich große Übereinstimmungen. Das gilt insbesondere für die NE-Klasse ‚PER', sowohl bei den konkreten Eigennamen als auch bei der Häufigkeit ihres Vorkommens (siehe Tabelle 1).

Tabelle 1. Die zehn am häufigsten extrahierten Personennamen nach Bibliothek

PER					
spaCy		Stanza		FLAIR	
Donald Trump	57	Donald Trump	57	Donald Trump	57
Angela Merkel	56	Angela Merkel	56	Angela Merkel	56
Jens Spahn	44	Jens Spahn	46	Jens Spahn	45
Markus Söder	27	Markus Söder	27	Markus Söder	27
Olaf Scholz	26	Olaf Scholz	26	Olaf Scholz	26
Christian Drosten	23	Christian Drosten	23	Christian Drosten	23
Peter Altmaier	17	Ursula von der Leyen	19	Peter Altmaier	17
Ursula von der Leyen	15	Peter Altmaier	17	Ursula von der Leyen	17
Winfried Kretschmann	14	Winfried Kretschmann	14	Winfried Kretschmann	14
Xi Jinping	14	Xi Jinping	14	Xi Jinping	14

Anmerkung. Hier wurden alle extrahierten Namensvarianten, Abkürzungen und Deklinationen des jeweiligen Personennamens zusammengefasst und pro Artikel einmal gezählt.

Bei der Klasse ‚ORG' sind im Vergleich mehr Abweichungen zwischen den Ergebnissen der einzelnen NER-Codepackages sichtbar. Die erkannten Eigennamen pro Artikel unterscheiden sich stärker, sodass die Rangfolge der vorkommenden Akteure nicht identisch ist. Auffällig ist, dass die Ergebnisse von FLAIR und spaCy sich bei der Anzahl der erkannten Organisationen stärker ähneln, während Stanza oft mehr Nennungen erkennt (siehe Tabelle 2).

Tabelle 2. Die zehn am häufigsten extrahierten Organisationsnamen nach Bibliothek

ORG					
spaCy		Stanza		FLAIR	
CDU	123	dpa	159	CDU	126
dpa	98	CDU	144	dpa	95
SPD	68	EU	102	SPD	68
WHO	63	SPD	86	EU	61
EU	61	RKI	62	WHO	57
RKI	52	WHO	59	RKI	53
Die Grünen	46	Die Grünen	56	Die Grünen	47
CSU	36	CSU	46	CSU	37
EU-Kommission	35	Twitter	39	Bundestag	23
Bundestag	31	Bundestag	34	Lufthansa	22

Anmerkung. Hier wurden alle extrahierten Namensvarianten, Abkürzungen und Deklinationen des jeweiligen Organisationsnamens zusammengefasst und pro Artikel einmal gezählt.

4.2 Precision der Verfahren

Entscheidend für die Precision ist die Anzahl an Begriffen und Begriffskombinationen, die fälschlich als Eigennamen klassifiziert wurden, die sogenannten False Positives. Je höher ihr Anteil an den erkannten Eigennamen, desto geringer die Precision. Zur Berechnung der Precision wurden alle durch die drei Verfahren ausgegebenen Eigennamen händisch geprüft und all jene Begriffe als False Positives markiert und gewertet, bei denen es sich unverkennbar nicht um Eigennamen handelt. Allgemein lässt sich feststellen, dass sich die Identifikationsleistung zwischen den verschiedenen NE-Klassen zum Teil relativ deutlich unterscheidet (dazu auch Jiang et al., 2016, S. 25; Shelar et al., 2020, S. 327).

Tabelle 3. Precision-Werte der drei Verfahren 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**

Beim Vergleich der drei Verfahren weist spaCy in allen NE-Klassen die meisten fälschlich als Eigennamen identifizierten Begriffe und somit im Durchschnitt den geringsten Precision-Wert auf (siehe Tabelle 3). Gerade bei Personen und Organisationen schneidet es deutlich schlechter ab. Stanza liegt mit einem Precision-Wert von insgesamt 0,83 im Mittelfeld. Auffällig ist allerdings die hohe Fehlerrate bei der Identifikation von Orten (Precision 0,75), die im Vergleich zu den beiden anderen NE-Klassen deutlich abfällt. Dies liegt hauptsächlich darin begründet, dass das Verfahren neben Namen von Orten, Gebäuden und Plätzen auch fälschlicherweise eine Unmenge von Nationalitäten und Regionalbezüge (‚amerikanische' oder ‚baden-württembergische') extrahiert. In dieser Auswertung wurden diese alleinstehenden Begriffe als False Positives markiert, eventuell wurden hier jedoch beim Training des Algorithmus andere Codieranweisungen verwendet als bei den anderen beiden Codepackages.

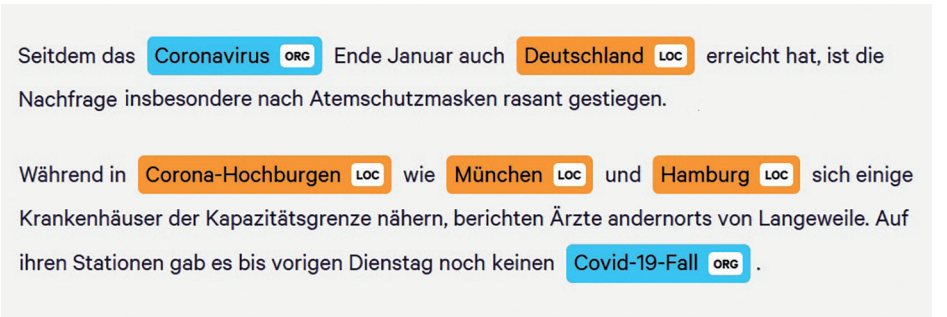
FLAIR erreicht in allen drei NE-Klassen und damit auch insgesamt die höchste Precision. Gerade in der Klasse ‚LOC' ordnet FLAIR erstaunlich wenig Begriffe falsch zu (nur 1,8 % der erkannten Begriffe oder Begriffskombinationen bezeichnen keine Ortsnamen).

Bei der Prüfung der identifizierten Eigennamen aller drei Verfahren sind verschiedene Fehlermuster aufgefallen, von denen die wichtigsten im Folgenden kurz vorgestellt werden. Dies soll einerseits die Transparenz und Nachvollziehbarkeit der Evaluation der Ergebnisse gewährleisten und kann andererseits einen Einblick in die unterschiedliche Funktionsweise der drei Verfahren bieten.

4.2.1 Fehlerquelle 1: Neologismen

Große Probleme hatten alle drei Verfahren mit Neologismen und Fremdwörtern, wie beispielsweise ‚Super-Spreader' und ‚Brexit' oder Wörtern wie ‚Hygge' und ‚Homeoffice'. Gerade in der Coronaberichterstattung findet sich eine Vielzahl von Begriffen oder Begriffsneubildungen, die zuvor überhaupt nicht oder kaum im allgemeinen Sprachgebrauch Verwendung fanden. Es liegt die Vermutung nahe, dass dies für die NER-Algorithmen unbekannte Begriffe sind und sie diese in manchen Fällen daher fälschlicherweise als Eigennamen einordnen. Wenn sich die zu analysierenden Texte inhaltlich stark von den Trainingsdaten unterscheiden, steigt die Wahrscheinlichkeit für solche Fehlleistungen der Verfahren (Maynard, 2016, S. 27). Eine Rolle könnten auch Personalisierungen spielen, durch die beispielsweise das Coronavirus in der Satzstruktur wie ein Akteur erscheint. Ein Beispiel hierfür wird in Abbildung 2 gezeigt. Die hier von spaCy automatisiert identifizierten Eigennamen sind farblich hervorgehoben. Sichtbar wird, dass Wörter mit Coronabezug irrtümlich als Eigennamen deklariert werden.

Abbildung 2. Beispielsätze aus dem Datensatz, in denen Corona-Begriffe als Eigennamen identifiziert werden



Knapp zehn Prozent der von spaCy extrahierten Wörter stellen solche False Positives mit Corona-Begrifflichkeiten dar (Sars, Coronavirus, Covid-Test, Corona-Infektion), bei FLAIR liegt der Anteil bei etwa sechs Prozent und bei Stanza bei leicht über drei Prozent.

Bei der gesamten Betrachtung der über 5.000 ermittelten False Positives innerhalb der Klassen ‚PER' sowie ‚ORG' und ‚LOC' ist spaCy für einen Großteil dieser falschidentifizierten Wörter verantwortlich (spaCy: 50 %; Stanza: 41 %; FLAIR: 9 %). Das liegt unter anderem daran, dass das NER-Verfahren von spaCy vermehrt Satzanfänge mit Artikeln irrtümlich als Eigennamen einstuft 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 beispielsweise das Nomen im Ausdruck ‚Ernst der Lage' handelt.

4.2.2 Fehlerquelle 2: Chunking

Ein weiterer Aspekt, der eine deutliche Unterscheidung zwischen den Leistungen der drei verschiedenen NER-Codepackages 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 (Chunking), also auch der Kombination mehrerer Begriffe, die zu einem Namen gehören, wie z. B. der Vor- und Nachname einer Person. Bei der Sichtung der erhaltenen Ergebnisse fällt auf, dass dieses sogenannte Chunking bei Doppelnamen von Personen und bei Namenszusätzen in Form von Berufsbezeichnungen recht unterschiedlich umgesetzt wird. Tabelle 4 zeigt exemplarisch, wie unterschiedlich die Grenzen von gewissen Personennamen identifiziert werden.

Tabelle 4. Beispiel für unterschiedliches Chunking der Bibliotheken

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

Das NER-Verfahren von spaCy extrahiert häufig zusätzliche Informationen mitsamt den Namen von Personen, wie ‚*Charité-Professor* Henning Rüden' oder ‚*Gesundheitssenator* Mario Czaja' (in beiden Fällen Hervorhebung durch die Autor:innen). Eine solche Zusatzinformation kann zwar in Einzelfällen eine hilfreiche, ergänzende Auskunft darstellen, entspricht jedoch nicht der eigentlichen Aufgabe bei der Erkennung von Eigennamen und stellt daher eine Fehlklassifikation dar. Das NER-Verfahren von Stanza hat dagegen bei zusammengesetzten Eigennamen mit Bindestrichen oft Schwierigkeiten, den gesamten Namen als Eigennamen zu erkennen und gibt in zahlreichen Fällen zwei separate Eigennamen aus (siehe Tabelle 4). 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. Insbesondere bei den Namen der Bundesländer fällt auf, dass es einen gravierenden Unterschied macht, wenn beispielsweise fälschlicherweise die Eigennamen ‚Baden' und ‚Sachsen' anstatt der gesamten Eigennamen ‚Baden-Württemberg' und ‚Sachsen-Anhalt' ausgegeben werden.

4.3 Recall der Verfahren

Eine allgemein gute Precision der NER-Verfahren sagt noch nichts darüber aus, ob relevante Ergebnisse fehlen, sondern nur wie fehlerarm die vorhandenen Ergebnisse sind. Um die Leistung vollständig bewerten zu können, müssen die automatisiert erkannten Eigennamen der drei Code-Bibliotheken mit den manuell erhobenen Daten abgeglichen werden. Hierfür werden alle Namen der individuellen und institutionellen Akteure, die manuell selektiert wurden, mit den erhaltenen Eigennamen der NER-Verfahren verglichen.

4.3.1 Individuelle Akteure

Für die Berechnung des Recalls der NER-Verfahren gelten die manuell codierten Akteure als zu erzielende Identifikationsleistung. In dem untersuchten Datensatz handelt es sich bei den individuellen Akteuren um 973 verschiedene Personennamen. 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. Doch obwohl die NER-Verfahren bis zu 55 Prozent zusätzliche Personennamen extrahieren, stimmen die meistgenannten Akteure in den Ergebnissen beider Erhebungsmethoden stark überein. Die Auflistung in Tabelle 5 zeigt, dass mit Ausnahme eines Namens die ermittelten Personen des Datensatzes bei der manuellen und der automatisierten Inhaltsanalyse identisch sind.

Tabelle 5. Häufigkeiten manuell und automatisiert erhobener Akteure

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

Für die konkrete Berechnung des Recalls wurde händisch je Artikel abgeglichen, wie viele der in diesem Artikel manuell codierten Akteure auch von den NER-Verfahren identifiziert wurden. Bei der Berechnung können zwei Herangehensweisen gewählt werden. Wenn bei dem artikelweisen Abgleich der identifizierten Akteure nur exakt übereinstimmende Personennamen gezählt werden, wird bei der Auswertung von exact matching gesprochen (Jiang et al., 2016, S. 23). Entsprechen sich die Namen nicht genau, stimmen aber weitestgehend überein, ist von loose matching die Rede (ebd.). Bei der Auswertung solcher partiellen Treffer werden ergänzend die Eigennamen mitgezählt, bei denen erkennbar ist, dass es sich um manuell selektierte Akteure handelt, doch ein Namensteil fehlt oder der Name mitsamt einer zusätzlichen Berufsbezeichnung extrahiert wurde (s. Tabelle 3).

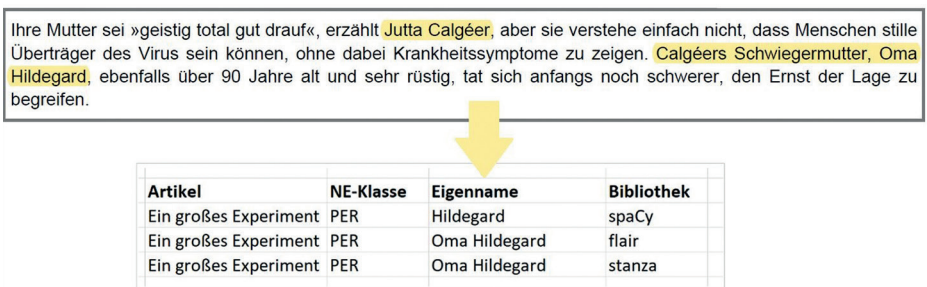
Tabelle 6 zeigt, wie viele der 973 individuellen Akteure aus der manuellen Codierung von den jeweiligen NER-Verfahren identifiziert wurden. Beim exact matching weist das NER-Verfahren von FLAIR die besten Werte auf. Die größeren Unterschiede der Werte zwischen exact und loose matching bei spaCy und Stanza kommen unter anderem durch die Fehler beim Chunking von Eigennamen zustande, die oben beschrieben wurden.

Tabelle 6. Recall-Werte bei der Identifikation von individuellen Akteuren

manuell: 973	spaCy	Stanza	FLAIR
Exact Matching	**925**	**924**	**950**
Recall	0,92	0,95	0,98
Loose Matching	**956**	**955**	**955**
Recall	0,98	0,98	0,98

Werden die partiellen Treffer im loose matching mitgezählt, unterscheiden sich die Recall-Werte der Verfahren fast nicht mehr voneinander. Alle drei NER-Verfahren finden 98 Prozent der manuell erhobenen individuellen Akteure. In Tabelle 6 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 Namen 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 entnommen, kommt dort jedoch nicht in dieser Form vor (siehe Abb. 3).

Abbildung 3. Textbeispiel mit zugehörigen extrahierten Eigennamen



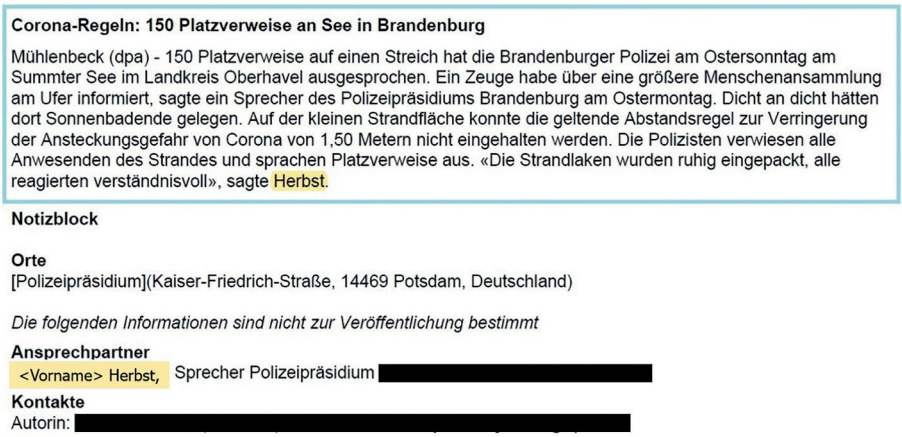
Quelle. SPIEGEL-Artikel im Datensatz und Screenshot der NER-Ergebnisse.

Die NER-Verfahren können nur die im Text vorhandenen Informationen ermitteln und geben daher korrekterweise die in Abbildung 3 aufgeführten Ergebnisse aus. Nur bei manueller Prüfung auf Artikelebene kann nachträglich ermittelt werden, dass damit die gleiche Akteurin bezeichnet wird.

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 vorab definierten Analysebereich (Textbody) des NER-Ver-

fahrens befinden. Abbildung 4 liefert hierzu ein Beispiel aus dem Datensatz. Zu sehen ist, dass im Artikel der gesamte Personenname nicht im Artikeltext vorkommt.

Abbildung 4. Eigenname befindet sich nicht in dem maschinell zu verarbeitendem Textbereich



Corona-Regeln: 150 Platzverweise an See in Brandenburg

Mühlenbeck (dpa) - 150 Platzverweise auf einen Streich hat die Brandenburger Polizei am Ostersonntag am Summter See im Landkreis Oberhavel ausgesprochen. Ein Zeuge habe über eine größere Menschenansammlung am Ufer informiert, sagte ein Sprecher des Polizeipräsidiums Brandenburg am Ostermontag. Dicht an dicht hätten dort Sonnenbadende gelegen. Auf der kleinen Strandfläche konnte die geltende Abstandsregel zur Verringerung der Ansteckungsgefahr von Corona von 1,50 Metern nicht eingehalten werden. Die Polizisten verwiesen alle Anwesenden des Strandes und sprachen Platzverweise aus. «Die Strandlaken wurden ruhig eingepackt, alle reagierten verständnisvoll», sagte Herbst.

Notizblock

Orte
[Polizeipräsidium](Kaiser-Friedrich-Straße, 14469 Potsdam, Deutschland)

Die folgenden Informationen sind nicht zur Veröffentlichung bestimmt

Ansprechpartner
<Vorname> Herbst, Sprecher Polizeipräsidium [redacted]

Kontakte
Autorin: [redacted]

Quelle. dpa-Artikel im Datensatz.

Es handelt sich dabei um keine Schwäche in der Texterkennung des genutzten Codepackages, sondern um ein Datenaufbereitungsproblem. Grundsätzlich ist der Arbeitsaufwand hinter der Datenaufbereitung im Vorfeld und der Datenbereinigung im Nachgang nicht zu unterschätzen. Diese Anwendungsschritte stellten den größten Zeitaufwand in dem hier durchgeführten Erhebungsprozess dar. Bei den übrigen individuellen Akteuren, die weder von spaCy, Stanza noch FLAIR erkannt werden, handelt es sich um Personen, die manuell nicht mit Vor- und Nachnamen, sondern nur mit ihren Berufsbezeichnungen (z. B. französischer Innenminister, Berliner Senatssprecherin) codiert wurden. Die automatisierte Extraktion solcher Akteure ist grundsätzlich schwierig, da die NER-Verfahren nicht dafür konzipiert wurden, Berufsbezeichnungen als Eigennamen zu identifizieren.

4.3.2 Institutionelle Akteure

Bei den manuell codierten institutionellen Akteuren handelt es sich um 862 verschiedene Organisationen, Institutionen und Behörden. Vereinzelt wurden auch Staaten und deutsche Bundesländer codiert, weshalb für den Vergleich auch die extrahierten Eigennamen der NE-Klasse ‚LOC' einbezogen werden. Im Vergleich zu den Personennamen erkennen die NER-Verfahren hier einen geringeren Anteil an Eigennamen. Für den Corona-Datensatz bildet Tabelle 7 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 Ei-

gennamengruppe spaCy die meisten Treffer liefert, während FLAIR die schlechteste Leistung erzielt.

Tabelle 7. Recall-Werte bei der Identifikation institutioneller Akteure

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

Bei dieser Auswertung 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 bei der manuellen Codierung. Dem NER-Verfahren von spaCy kommt hier 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.

4.4 F-Score und Einordnung der Genauigkeitsmaße

Um die Güte der getesteten Verfahren zu beurteilen und für den Einsatz in zukünftigen inhaltsanalytischen Untersuchungen zu validieren, werden die insgesamt erzielten Leistungskennzahlen im Überblick betrachtet. Gegenwärtig liegen für deutschsprachige Textkorpora kaum Referenzwerte vor, die als Maßstab genutzt werden können, um zu bestimmen, 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 (Qi et al., 2020, S. 6; Yadav & Bethard, 2019, S. 5). Die hier ermittelten F-Scores der NER-Verfahren von spaCy und Stanza liegen in diesem Wertebereich, während das Verfahren von FLAIR ihn übertrifft. Tabelle 8 bildet diese Gütemaße für alle untersuchten NE-Klassen zusammengefasst ab.

Tabelle 8. Darstellung der übergreifenden F-Scores pro Bibliothek

		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

Tabelle 8 zeigt außerdem deutlich, dass spaCy das Verfahren mit den höchsten Recall-Werten ist, während FLAIR die beste Precision bei der Identifikation von Eigennamen aufweist. Die hier erhaltenen finalen F-Scores decken sich außerdem gut mit den höchsterzielten Werten der jeweiligen Bibliotheken, die auf deren Publikationsplattformen genannt werden (siehe 3.3).

5. Diskussion

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

Im Vorfeld war ungewiss, wie gut sich vorab trainierte ML-Verfahren auf unbekannte Datensätze anwenden lassen, da unter anderem der zugrundeliegende Trainingskorpus des gewählten Sprachmodells die grundsätzliche Leistung der NER-Verfahren beeinflussen kann. Doch die errechneten Kennzahlen der Untersuchung zeigen, dass die getesteten NER-Verfahren in der Lage sind, in unbekannten Textdaten mit hoher Verlässlichkeit Eigennamen zu identifizieren.

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 sollen, da sich die Identifikationsleistung je nach Gruppe stark unterscheidet. Bei der Identifikation individueller Akteure weisen alle getesteten Verfahren hohe Trefferquoten auf. Im Vergleich zu den erzielten Werten bei der Identifikation von Personen lässt die Erkennungsleistung von Institutions- und Organisationsnamen jedoch noch Raum für Verbesserungen.

Die unterschiedlichen Fehler, die bei der Extraktion von Eigennamen erkannt wurden, beeinträchtigen insbesondere die Arbeit mit den ausgegebenen Analyseergebnissen von Stanza und spaCy. Daher kann hier keine uneingeschränkte Empfehlung für den Einsatz gegeben werden. Von den getesteten Verfahren eignet sich für eine präzise und vollständige Erkennung von *Personennamen* das NER-Codepackage von FLAIR am besten. Es extrahiert hier 99 Prozent der relevanten Akteure und liefert wenig irrelevante und kaum unvollständige Ergebnisse (Precision = 0,94 und Recall = 0,99).

Wenn daher die Identifikation von Personen im Fokus einer Analyse liegt, kann dieses Codepackage zweifelsfrei zur Extraktion der individuellen Akteure empfohlen werden. Wenn hingegen die Identifikation von institutionellen Akteuren

Schwerpunkt der 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 - aufgrund des erhöhten Recalls des Verfahrens - voraussichtlich vollständigere Ergebnisse bei der Extraktion von Organisationsnamen erhalten werden.

5.2 Replikation manueller Codierungen durch die NER-Verfahren

Der durchgeführte Vergleich zeigt, dass mit einer Trefferquote von 83-88 Prozent automatisiert dieselben Akteure im Datensatz ermittelt werden können, die auch manuell erhoben wurden – bei einer grundsätzlich schnelleren Bearbeitung und Ausgabe. Das automatisierte Verfahren weist somit gemessen an einer manuellen Codierung eine vergleichbare Unsicherheit bei der Akteursidentifizierung auf. Manuelle Inhaltsanalysen können allerdings je nach Komplexität des verwendeten Codebuchs und Größe der untersuchten Stichprobe Wochen bis Monate in Anspruch nehmen. Der Einsatz der automatisierten Verfahren kann dagegen je nach vorhandenen Vorkenntnissen beim Aufsetzen eines lauffähigen Programmiercodes und dem Bereinigen der daraus erhaltenen Daten in wenigen Tagen abgeschlossen sein. Zudem lassen sich nach dem ersten Einrichtungs- und Auswertungsaufwand gleichartige Analysen in kürzester Zeit durchführen. Bei einer Erweiterung der Stichprobe, um beispielsweise 200 (oder 2000) weitere maschinenlesbare Nachrichtenartikel, würde der Unterschied in der Verarbeitungsdauer bei den automatisierten NER-Verfahren nur wenige Minuten oder Stunden betragen, während menschliche Codierer:innen 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 (Rössler, 2017, S. 141) und Akteure nur dann codiert, wenn sie direkt oder indirekt in dem Nachrichtenartikel zu Wort kommen. Mit einem NER-Verfahren allein ist die Extraktion nach solchen inhaltlichen Kriterien nicht möglich. Die hier automatisiert erhaltenen Eigennamen lassen damit keine Aussage darüber zu, in welcher Rolle die Akteure in einem Text vorkommen. Zudem ist es wichtig, zu unterscheiden zwischen Namen und den Objekten (oder Entitäten), die sie bezeichnen. Die Named Entity Verfahren, die in dieser Studie verwendet wurden, sind nur in der Lage, Eigennamen zu bestimmen, nicht aber, welche konkrete Personen oder Objekte mit diesen Namen bezeichnet werden. Dadurch sind sie auch nicht in der Lage, sogenannte Koreferenzen aufzulösen, also Eigennamen zu gruppieren, die sich auf das gleiche Objekt beziehen. Für solche Informationen ist der Einsatz zusätzlicher, spezialisierter Verfahren notwendig, bspw. Named Entity Linking, um Eigennamen zu konkreten Personen oder Organisationen zuzuordnen (bspw. Al-Moslmi, 2020; Rao, 2013) oder Entity Relation Extraction, um den semantischen Zusammenhang zwischen den Eigennamen zu extrahieren (Bach & Badaskar 2007, S. 13; Jurafsky & Martin, 2021, S. 340).

Nichtsdestotrotz können mit den Ergebnissen der automatisierten NER-Analyse verschiedene Auswertungen durchgeführt werden: So kann zum Beispiel das Vorkommen bestimmter Akteure im Zeitverlauf abgebildet sowie eine Analyse

der Kookkurrenz von Akteuren in den Nachrichtenartikeln durchgeführt werden. Wenn die ausgewählte Stichprobe repräsentativ für bestimmte Medientitel ist, ließen die Ergebnisse der NER-Analyse außerdem Aussagen über die Akteursvielfalt in den jeweiligen Medientitel zu.

Außerdem eignen sich die Ergebnisse automatisierter Analysen durchaus auch als Vergleichsmaßstab für händisch identifizierte Akteure, um Rechtschreib- oder Codierungsfehler zu entdecken. Im hier verwendeten Vergleichsdatensatz kamen bei den manuell codierten individuellen Akteuren in etwa fünf Prozent der Fälle solche Fehler vor. 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 eher irrelevant sind.

Doch es wurden auch Namen gefunden, die uneinheitlich codiert wurden, und unentdeckt eine Zusammenfassung und Auswertung der Ergebnisse beeinträchtigen könnten. Tippfehler in anderen Kategorien können darüber hinaus in der Zuteilung einer inkorrekten Merkmalsausprägung resultieren (‚1' = ‚individueller Akteur' statt ‚2' = ‚institutioneller Akteur') und bleiben eventuell unentdeckt. Mit Hilfe eines Vergleichs von automatisiert extrahierten Namen können solche Unstimmigkeiten schnell erkannt und die Ergebnisse korrigiert werden.

5.3 Limitationen

Die Limitationen dieser Studie beziehen sich im Wesentlichen auf die Stichprobe sowie die eingeschränkte Auswahl der NER-Packages. Der Datensatz befasst sich mit der Medienberichterstattung zu Corona in der ersten Jahreshälfte 2020. Dieser kleine und relativ spezielle Datensatz beinhaltet nur wenige Medientitel, was auch mit der Verfügbarkeit des digitalen Textkorpus zusammenhängt. Repräsentative Aussagen werden daher in dieser Studie nicht getroffen. Einschränkend muss weiterhin bemerkt werden, dass ein Vergleich der hier getesteten Codepackages mit Paketen aus weiteren Programmiersprachen sinnvoll wäre.

6. Fazit

Ziel war die Anwendung und Validierung eines automatisierten Verfahrens zur Identifikation von Akteuren in journalistischen Texten. Hierfür wurden drei unterschiedliche NER-Softwarepakete gewählt, mit denen Eigennamen aus deutschsprachigen Nachrichtenartikeln extrahiert wurden. Die Ergebnisse wurden mit Ergebnissen aus einer manuellen Inhaltsanalyse verglichen. So konnten die jeweiligen Stärken und Schwächen der drei getesteten Verfahren ermittelt werden. Trotz großer Unterschiede in der Verarbeitungszeit, der Menge an extrahierten Eigennamen und der Fehlerraten bei ihrer Identifikation erwiesen sich alle drei Verfahren als gut geeignet: Die automatisiert identifizierten Eigennamen deckten 99 Prozent der individuellen Akteure ab, die händisch erhoben wurden. Von den manuell selektierten institutionellen Akteuren wurden dagegen nur zwischen 68 Prozent und 80 Prozent von den NER-Verfahren erkannt. Die berechneten Leistungskennzahlen zur Bewertung der Verfahren liegen in den Wertebereichen, die

auch in anderen Publikationen beim Test von NER-Verfahren und bei manuellen Inhaltsanalysen erzielt werden.

Da gegenwärtig kaum Orientierungsmaßstäbe und standardisierte Vorgehensweisen zur Anwendung automatisierter Verfahren in der Fachliteratur vorliegen, leistet diese Studie einen Beitrag zur Klärung, welche NER-Packages für die Analyse von Akteuren in deutschsprachigen journalistischen Texten zur Anwendung kommen können und zur Frage, ob die Qualität der NER-Packages hoch genug ist – was klar bejaht werden kann.

Literaturverzeichnis

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In W. Ammar, A. Louis, & N. Mostafazadeh (Hrsg.), *Annual Conference of the North American Chapter of the Association for Computational Linguistics* (S. 54–59). https://doi.org/10.18653/v1/N19-4010

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In E. Bender, L. Derczynski, & P. Isabelle (Hrsg.), *Proceedings of the 27th International Conference on Computational Linguistics* (S. 1638–1649). Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1139/

Al-Moslmi, T., Gallofré Ocaña, M., Opdahl, A. L., & Veres, C. (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access, 8*, 32862–32881. https://doi.org/10.1109/ACCESS.2020.2973928

Augenstein, I., Derczynski, L., & Bontcheva, K. (2017). Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language 44*, 61–83. https://doi.org/10.1016/j.csl.2017.01.012

Bach, N., & Badaskar, S. (2007). *A review of relation extraction.* Carnegie Mellon University. https://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf

Benikova, D., Biemann, C., & Reznicek, M. (2014). NoSta-D named entity annotation for German: Guidelines and dataset. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Hrsg.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (S. 2524–2531). http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3(Jan)*, 993–1022.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). *Pandemic populism: Facebook pages of alternative news media and the corona crisis - a computational content analysis* [Preprint]. arXiv. https://arxiv.org/abs/2004.02566v3

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures, 14*(2), 83–104. https://doi.org/10.1080/19312458.2019.1671966

Boumans, J., & Trilling, D. (2016). Taking stock of the toolkit. *Digital Journalism, 4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Burggraaff, C., & Trilling, D. (2020). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism, 21*, 112–129. https://doi.org/10.1177/1464884917716699

Derczynski, L. (2016). Complementarity, F-score, and NLP evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Hrsg.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (S. 261–266). European Language Resources Association. https://www.aclweb.org/anthology/L16-1040

Di Franco, G., & Santurro, M. (2021). Machine learning, artificial neural networks and social research. *Quality & Quantity*, *55*(3), 1007–1025. https://doi.org/10.1007/s11135-020-01037-y

Dipper, S., & Kübler, S. (2017). German treebanks: TIGER and TüBa-D/Z. In N. Ide & J. Pustejovsky (Hrsg.), *Handbook of linguistic annotation* (S. 595–639). Springer. https://doi.org/10.1007/978-94-024-0881-2_22

Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). 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". *International Journal of Communication, 13*, 3876–3884. https://ijoc.org/index.php/ijoc/article/view/10533

Dumm, S., & Niekler, A. (2014). Methoden und Gütekriterien. Computergestützte Diskurs- und Inhaltsanalysen zwischen Sozialwissenschaft und Automatischer Sprachverarbeitung [Methods and quality criteria. Computer-aided discourse and content analyses between social science and automatic language processing]. *Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus*, *Discussion Paper 4*. http://www.epol-projekt.de/discussion-paper/discussion-paperdiscussion-paper-4/

Eftimov, T., Korousić Seljak, B., & Korosec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE, 12(6)*, Article e0179488. https://doi.org/10.1371/journal.pone.0179488

Eisenegger, M., Oehmer, F., Udris, L., & Vogler, D. (2020). *Qualität der Medien Studie 1/2020: Die Qualität der Medienberichterstattung zur Corona-Pandemie* [The quality of media 1/2020: The quality of media coverage of the Corona pandemic]. Forschungszentrum Öffentlichkeit und Gesellschaft (fög). http://www.foeg.uzh.ch/dam/jcr:ad278037-fa75-4eea-a674-7e5ae5ad9c78/Studie_01_2020.pdf

Faruqi, M., & Padó, S. (2010). Training and evaluating a German named entity recognizer with semantic generalization. In M. Pinkal, I. Rehbein, S. Schulte im Walde, & A. Storrer (Hrsg.), *Semantic approaches in natural language processing: Proceedings of the Conference on Natural Language Processing 2010* (S. 129–134). Universaar.

Gaus, A. (2018). 8 lessons learned about NER. Medium, dpa-newslab, 06.02.2018, https://medium.com/dpa-newslab/8-lessons-learned-about-ner-f40b263490db

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21* (3), 267–297. https://doi.org/10.1093/pan/mps028

Günther, E., & Quandt, T. (2016). Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism, 4*(1), 75–88. https://doi.org/10.1080/21670811.2015.1093270

Hepp, A. (2016). Kommunikations- und Medienwissenschaft in datengetriebenen Zeiten [Communciation and media science in datadriven times]. *Publizistik 61*, 225–246. https://doi.org/10.1007/s11616-016-0263-y

Honnibal, M. (o.J.). spaCy's NER model [Video]. Abgerufen von https://spacy.io/universe/project/video-spacys-ner-model.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python.* Zenodo. https://doi.org/10.5281/zenodo.1212303

Jannidis, F. (2017). Grundbegriffe des Programmierens [Basic concepts of programming]. In F. Jannidis, H. Kohle, & M. Rehbein (Hrsg.), *Digital Humanities: Eine Einführung* (S. 68–95). J. B. Metzler.

Jiang, R., Banchs, R. E., & Li, H. (2016). Evaluating and combining name entity recognition systems. In X. Duan, R. E. Banchs, M. Zhang, H. Li, & A. Kumaran (Hrsg.), *Proceedings of the Sixth Named Entity Workshop* (S. 21–27). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2703

Jurafsky, D., & Martin, J. H. (2021). *Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition* (3rd edition draft). Pearson. https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf

Karlsson, M., & Sjøvaag, H. (2016). Introduction: Research methods in an age of digital journalism. *Digital Journalism, 4*(1), 1–7. https://doi.org/10.1080/21670811.2015.1096595

Kelm, O., Gerl, K., & Meißner, F. (2020). Machine learning. In I. Borucki, K. Kleinen-von Königslöw, S. Marschall, & T. Zerback (Hrsg.), *Handbuch Politische Kommunikation* (S.1–9). Springer. https://doi.org/10.1007/978-3-658-26242-6_55-1

Ketschik, N., Overbeck, M., Murr, S., Pichler, A., & Blessing, A. (2020). Interdisziplinäre Annotation von Entitätenreferenzen. Von fachspezifischen Fragestellungen zur einheitlichen methodischen Umsetzung [Interdisciplinary annotation of entity references. From subject-specific issues to uniform methodological implementation]. In N. Reiter, A. Pichler, & J. Kuhn (Hrsg.), *Reflektierte algorithmische Textanalyse* [Reflective algorithmic text analysis] (S. 203–236). De Gruyter. https://doi.org/10.1515/9783110693973-010

Lane, H., Howard, C., & Hapke, H. (2019). *Natural language processing in action. Understanding, analyzing, and generating text with Python.* Manning, Pearson Education.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leidecker-Sandmann, M., Attar, P., & Lehmkuhl, M. (2021). *Selected by expertise? Scientific experts in German news coverage on Covid-19 compared to other pandemics* [Preprint]. SocArXiv. https://osf.io/preprints/socarxiv/cr7dj/

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of Big Data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media 57*(1), 34–52. https://doi.org/10.1080/08838151.2012.761702

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces, 35*(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 23*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Maynard, D., Bontcheva, K., & Augenstein, I. (2016). Natural language processing for the semantic web. *Synthesis lectures on the semantic web: Theory and technology, 6*, 1–194. https://doi.org/10.2200/S00741ED1V01Y201611WBE015

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes 30*(1), 3–26. https://doi.org/10.1075/li.30.1.03na

Niekler, A. (2018). *Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen* [Automated procedures for topic analysis of news-oriented text sources]. Herbert von Halem Verlag.

Niemann-Lenz, J., Bruns, S., Hefner, D., Knop-Hülß, K., Possler, D., Reich, S., Reinecke, L., Scheper, J., & Klimmt, C. (2019). Crafting a strategic roadmap for computational methods in communication science: Learnings from the CCS 2018 Conference in Hanover – Commentary. *International Journal of Communication, 13*, 3885–3893. https://ijoc.org/index.php/ijoc/article/view/10534

Parks, M. R. (2014). Big Data in communication research: Its contents and discontents. *Journal of Communication*, *64*(2), 355–360. https://doi.org/10.1111/jcom.12090

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python natural language processing toolkit for many human languages* [Preprint]. arXiv. https://arxiv.org/abs/2003.07082

Rao, D., McNamee, P., & Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In T. Poibeau, H. Saggion, J. Piskorski, & R. Yangarber (Hrsg.) *Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing, 93–116*. Springer. https://doi.org/10.1007/978-3-642-28569-1_5

Rössler, M. (2007). *Korpus-adaptive Eigennamenerkennung* [Corpus-adaptive proper name recognition] [Dissertation, Universität Duisburg-Essen]. Winterthur. https://duepublico2.uni-due.de/receive/duepublico_mods_00014746

Rössler, P. (2017). *Inhaltsanalyse* (3. Auflage) [Content analysis (3rd ed.)]. UVK.

Scharkow, M. (2012). *Automatische Inhaltsanalyse und maschinelles Lernen* [Automatic content analysis and machine learning] [Dissertation, Universität der Künste Berlin]. epubli. https://opus4.kobv.de/opus4-udk/frontdoor/deliver/index/docId/28/file/dissertation_scharkow_final_udk.pdf

Scharkow, M. (2013). Automatische Inhaltsanalyse [Automatic content analysis]. In W. Möhring & D. Schlütz (Hrsg.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* [Handbook of standardized survey procedures in communication science] (S. 289–306). Springer. https://doi.org/10.1007/978-3-531-18776-1_16

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schneider, G. (2014). Automated media content analysis from the perspective of computational linguistics. In K. Sommer, J. Matthes, M. Wettstein, & W. Wirth (Hrsg.), *Automatisierung in der Inhaltsanalyse* [Automation in content analysis.] (S. 40–54). Herbert von Halem.

Schweiger, W. (2017). *Der (des)informierte Bürger im Netz: Wie soziale Medien die Meinungsbildung verändern* [The (dis)informed citizen on the internet: How social media are changing the way opinions are formed]. Springer.

Schwotzer, B. (2014). Automatische Selektion von Beiträgen für themenspezifische Inhaltsanalysen mittels Schlagwortlisten [Automatic selection of articles for topic-specific content analyses using keyword lists]. In K. Sommer, J. Matthes, M. Wettstein, & W. Wirth (Hrsg.), *Automatisierung in der Inhaltsanalyse* [Automation in content analysis] (S. 55–72). Herbert von Halem.

Shelar, H., Kaur, G., Heda, N., & Agrawal, P. (2020). Named entity recognition approaches and their comparison for custom NER model, *Science & Technology Libraries 39*(3), 324–337. https://doi.org/10.1080/0194262X.2020.1759479

Sommer, K., Matthes, J., Wettstein, M., & Wirth, W. (2014). Ein Vorwort – Automatisierung in der Inhaltsanalyse. Methoden und Forschungslogik der Kommunikationswissenschaft [Preface – Automation in content analysis]. In K. Sommer, J. Matthes, M. Wettstein, & W. Wirth (Hrsg.), *Automatisierung in der Inhaltsanalyse* (S. 9–16). Herbert von Halem.

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research*, *2*, 109–134. https://doi.org/10.5117/CCR2020.1.005. KATH

Strippel, C., Bock, A., Katzenbach, C., Mahrt, M., Merten, L., Nuernbergk, C., Pentzold, C., Puschmann, C., & Waldherr, A. (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) [The future of communication science is already here, it's just not evenly distributed. A collective response to contributions in „Forum" (Publizistik, Issue 3 and 4, 2016)]. *Publizistik, 63*, 11–27. https://doi.org/10.1007/s11616-017-0398-5

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In W. Daelemans & M. Osborne (Hrsg.*), Proceedings of CoNLL-2003* (S. 142–145). Morgan Kaufman Publishers.

van Atteveldt, W., Margolin, D., Shen, C., Trilling, D., & Weber, R. (2019). A roadmap for computational communication research. *Computational Communication Research*, *1*(1). https://computationalcommunication.org/ccr/article/view/37

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures, 15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

van der Meer, T. G. L. A. (2016). Automated content analysis and crisis communication research. *Public Relations Review, 42*(5), 952–961. https://doi.org/10.1016/j.pubrev.2016.09.001

Wettstein, M. (2014). "Best of both worlds": Die halbautomatische Inhaltsanalyse [The half-automatic content analysis]. In K. Sommer, J. Matthes, M. Wettstein, & W. Wirth (Hrsg.), *Automatisierung in der Inhaltsanalyse* (S. 16–39). Herbert von Halem.

Yadav, V., & Bethard, S. (2019). *A survey on recent advances in named entity recognition from deep learning models* [Preprint]. arXiv. https://arxiv.org/abs/1910.11470

Züll, C., & Mohler, P. (2001). Computerunterstützte Inhaltsanalyse: Codierung und Analyse von Antworten auf offene Fragen [Computer-aided content analysis: coding and analysis of replies to outstanding questions]. *GESIS-How-to, 8. Zentrum für Umfragen, Methoden und Analysen (ZUMA)*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-201405

SCM
Studies in Communication and Media

EXTENDED ABSTRACT

Validation of NER methods for the automated identification of actors in German journalistic texts

Cecilia Buz, Nikolai Promies, Sarah Kohler & Markus Lehmkuhl

Cecilia Buz (M. A.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunikation, Englerstraße 2, D-76131 Karlsruhe. Contact: cecilia.buz(at)partner.kit.edu.

Nikolai Promies (M. A.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunikation, Englerstraße 2, D-76131 Karlsruhe. Contact: nikolai.promies(at)kit.edu. ORCID: https://orcid.org/0000-0002-4804-4155

Sarah Kohler (Dr.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunikation, Englerstraße 2, D-76131 Karlsruhe. Contact: sarah.kohler(at)kit.edu. ORCID: https://orcid.org/0000-0002-3548-010X

Markus Lehmkuhl (Prof. Dr.), Karlsruher Institut für Technologie, Institut für Technikzukünfte (ITZ), Department für Wissenschaftskommunıkation, Englerstraße 2, D-76131 Karlsruhe. Contact: markus.lehmkuhl(at)kit.edu. ORCID: https://orcid.org/0000-0001-8295-6548

EXTENDED ABSTRACT

Validation of NER methods for the automated identification of actors in German journalistic texts

Cecilia Buz, Nikolai Promies, Sarah Kohler & Markus Lehmkuhl

1. Introduction

In recent years, the interdisciplinary field of computational communication science has emerged within the social sciences as a mixed discipline between applied computational science and communication science (Domahidi et al., 2019, p. 3877). Among other things, it focuses on the use of automated procedures to analyze the content of large text collections via algorithms to e.g. identify patterns in them and to visualize these data structures (Grimmer & Stewart, 2013, p. 267). As such methods become more widespread and important, there is an increasing need for a methodological discussion of their application and validation (Niekler, 2018, p. 179; Strippel et al., 2018, p. 18. This is where this paper wants to contribute: It aims to validate a method referred to as Named Entity Recognition (NER). NER is able to automate a sub-step in the content analysis of text data by automatically extracting names of persons, places, and organizations from texts. For the validation, three different NER code packages were used to extract individual and institutional actors from a set of journalistic texts dealing with the coronavirus pandemic. The results were then compared to the results of a manual content analysis of the same texts.

2. Named Entity Recognition

Named Entity Recognition is a method from the field of Natural Language Processing (NLP) and is used to extract information from unstructured texts with the goal of identifying named 'real objects' that consist of proper names (Marrero et al., 2013, p. 482). When using a NER procedure, the proper names in a text are identified and assigned to a specific class (Eftimov et al., 2017, p. 3). The four most common classes include the labels PER, ORG, LOC, and MISC for the categories person, organization, place, and miscellaneous (Faruqi & Padó, 2010, p. 130).

Based on text corpora in which human coders have identified different types of proper names, machine learning methods can be used to train NER models that then recognize proper names in unknown texts. The creation of a training corpus for this supervised learning process is relatively time-consuming, which is why most analyses use pre-trained models. There is a large number of easily accessible

code libraries in many programming languages with different pre-trained models, which can be adapted and used with little effort. The available models differ in the text data they were trained with and in the machine learning algorithms used for their training. As these models have been trained and optimized with specific data sets, it is uncertain whether these NER packages can provide correct and accurate results when analyzing unknown journalistic news texts.

3. Method

We selected the NER functions of the Python libraries spaCy (Honnibal et al., 2020), Stanza (Qi et al., 2020), and FLAIR (Akbik et al., 2019) for our analysis, because they are considered the best-performing methods for named entity recognition in various recent publications (e.g. Shelar et al., 2020, p. 324; Qi et al., 2020, p. 6). All three software packages are freely available and can be used without in-depth knowledge of programming or machine learning.

The basis for the evaluation of the three NER packages are 887 news articles published by the German media outlets SPIEGEL and WELT and by the news agency Deutsche Presse-Agentur (dpa) from January to June 2020. This dataset was compiled as part of a manual content analysis of COVID-19 reporting in Germany (Leidecker-Sandmann et al., 2021). This manual content analysis determined, among other things, which individual, institutional, or generic actors appear in the coverage. We used the actors and institutions identified in the manual content analysis as the reference with which we compared the automatically extracted proper names. On the one hand, it was tested whether the same actors are identified, and on the other hand, whether the determined frequency of actors is similar across all articles between the manual method and the automated method.

In the field of applied computer science, the parameters *precision* and *recall* are often used as quality criteria to evaluate the performance of NLP tasks (Dumm & Niekler, 2014, p. 21). Here, precision indicates the 'exactness' or 'reliability' of the procedure, while recall evaluates the 'completeness' of the results (Rössler, 2007, p. 92). In addition to precision and recall, a third common quality measure for comparing machine learning methods is the so-called F-score. This is the harmonic mean of the precision and recall values. We calculated the precision, recall, and F-score for all three compared methods for the classes 'PER', 'ORG', and 'LOC' in order to be able to generally assess how reliable the automatically generated results are and to determine the 'most accurate' method.

4. Results

To determine the precision, the number of terms that were incorrectly classified as proper names (false positives) were counted. The higher their share in all of the recognized proper names, the lower the precision.

When comparing the three methods, spaCy shows the highest error rates in all classes and thus the lowest precision value on average (0.81). Especially for persons and organizations, it performs worse than the other packages. Stanza is in the midfield with an average precision value of 0.83 and values of 0.91 for per-

sons and organizations, while FLAIR achieves the highest precision in all NE classes and thus overall (0.93).

For calculation of the recall of the NER procedures, the manually coded actors are considered as the gold standard to be reached. All names of the individual and institutional actors that were manually identified are compared with the obtained proper names of the NER procedures. Two approaches can be used to determine what percentage of the manually coded actors were also identified by the NER procedures. If only exactly matching names are counted when comparing the identified actors, the evaluation is referred to as *exact matching* (Jiang et al., 2016, p. 23). If the names do not correspond exactly, but match closely, this is referred to as loose matching (ibid.). In the evaluation of such partial matches, we also counted named entities, whose names were only partially extracted or that included additional terms that were not part of the actual name.

In the dataset studied, the individual actors consist of 973 different person names. Although the NER methods extract up to 55 percent additional individual names, the most frequently named actors strongly agree in the results of both survey methods. Using exact matching, all three packages found at least 92 % (spaCy) and up to 98 % (FLAIR) of the manually identified actors. The manually coded institutional actors comprise 862 different organizations, institutions, and authorities. Occasionally, states and German federal states were also coded, which is why the extracted proper names of the NE class 'LOC' are also included for comparison. Compared to the personal names ('PER'), the NER methods recognize only a smaller proportion of those proper names. When the twenty most frequently named institutional actors in the manual survey of the coronavirus dataset are compared with the twenty most frequently extracted proper names of the classes 'ORG' and 'LOC', only half of the names in each listing match. Overall, it can be seen that for all entity classes, spaCy provides the most hits (recall of 88 % with exact matching) while FLAIR performs the worst (recall of 83 % with exact matching).

5. Discussion

Currently, there are hardly any reference values available for German-language text corpora that can be used as a benchmark to determine which F-score values to aim for or to meet for a successful validation. The listed values of other publications range between 0.64 and 0.86 for German texts (Qi et al., 2020, p. 6; Yadav & Bethard, 2019, p. 5). The F-scores of the NER procedures of spaCy and Stanza determined here fall within this range of values, while the procedure of FLAIR exceeds it (see Table 1).

Table 1. Overview of the performance metrics per NER package

		Overall performance (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**

The comparison carried out shows that with a recall rate of 83–88 percent, the same actors that were collected manually can be identified automatically in the data set– in a much shorter span of time. Manual content analyses can take weeks to months, depending on the complexity of the codebook used and the size of the sample studied. The use of automated methods, on the other hand, can be completed in a matter of days, depending on previous knowledge of setting up an executable programming code and cleaning the data obtained from it.

In a manual survey, however, it is much more flexible and easier to specify which actors are of interest in the journalistic texts. In the manual content analysis discussed here, the function of an actor was taken into account (Rössler, 2017, p. 141) and actors were only coded if they were quoted directly or indirectly in the news article. Extracting such content-related criteria is only possible in combination with additional text analysis methods. Moreover, the automatically obtained proper names do not yet allow any statement about the role in which the actors appear in a text, e.g. about how prominently they are represented.

It is difficult to evaluate whether the recall of the three packages is sufficient for the identification of all relevant actors in journalistic texts. For this assessment, the type of actors to be extracted from the texts has to be specified, since the identification performance differs strongly depending on the entity class. For the identification of individual actors (NER class 'PER'), all tested methods show high hit rates. However, compared to the achieved values for the identification of individuals, the recognition performance of institution and organization names (NER class 'ORG') still leaves room for improvement. Of the methods tested, the NER code package from FLAIR is best suited for the accurate and complete recognition of personal names. It extracts 99 percent of the relevant actors and provides few irrelevant and hardly any incomplete results (precision = 0.94 and recall = 0.99). Therefore, if the focus of an analysis is the identification of persons, this code package can be recommended for the extraction of individual actors.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In W. Ammar, A. Louis, & N. Mostafazadeh (Eds.), *Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 54–59). https://doi.org/10.18653/v1/N19-4010

Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). 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". *International Journal of Communication 13*, 3876–3884. https://ijoc.org/index.php/ijoc/article/view/10533

Dumm, S., & Niekler, A. (2014). Methoden und Gütekriterien. Computergestützte Diskurs- und Inhaltsanalysen zwischen Sozialwissenschaft und Automatischer Sprachverarbeitung [Methods and quality criteria. Computer-aided discourse and content analyses between social science and automatic language processing]. *Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus. Discussion Paper 4*. http://www.epol-projekt.de/discussion-paper/discussion-paperdiscussion-paper-4/

Eftimov, T., Korousić Seljak, B., & Korosec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE 12(6)*, Article e0179488. https://doi.org/10.1371/journal.pone.0179488

Faruqi, M., & Padó, S. (2010). Training and evaluating a German named entity recognizer with semantic generalization. In M. Pinkal, I. Rehbein, S. Schulte im Walde, & A. Storrer (Eds.), *Semantic approaches in natural language processing: Proceedings of the Conference on Natural Language Processing 2010* (pp. 129–134). Universaar.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303

Jiang, R., Banchs, R. E., & Li, H. (2016). Evaluating and combining name entity recognition systems. In X. Duan, R. E. Banchs, M. Zhang, H. Li, & A. Kumaran (Eds.), *Proceedings of the Sixth Named Entity Workshop* (pp. 21–27). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2703

Leidecker-Sandmann, M., Attar, P., & Lehmkuhl, M. (2021). *Selected by expertise? Scientific experts in German news coverage on Covid-19 compared to other pandemics*. SocArXiv. https://osf.io/preprints/socarxiv/cr7dj/

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces 35*(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

Niekler, A. (2018). *Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen* [Automated procedures for topic analysis of news-oriented text sources]. Herbert von Halem Verlag.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python natural language processing toolkit for many human languages* (Preprint). arXiv. https://arxiv.org/abs/2003.07082

Rössler, M. (2007). *Korpus-adaptive Eigennamenerkennung* [Corpus-adaptive proper name recognition] (Dissertation, Universität Duisburg-Essen). Winterthur. https://duepublico2.uni-due.de/receive/duepublico_mods_00014746

Rössler, P. (2017). *Inhaltsanalyse* [Content analysis] (3rd ed.). UVK.

Shelar, H., Kaur, G., Heda, N., & Agrawal, P. (2020). Named entity recognition approaches and their comparison for custom NER model, *Science & Technology Libraries 39*(3), 324–337. https://doi.org/10.1080/0194262X.2020.1759479

Strippel, C., Bock, A., Katzenbach, C., Mahrt, M., Merten, L., Nuernbergk, C., Pentzold, C., Puschmann, C., & Waldherr, A. (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) [The future of communication science is already here, it's just not evenly distributed. A collective response to contributions in 'Forum' (Publizistik, Issue 3 and 4, 2016)]. *Publizistik 63*, 11–27. https://doi.org/10.1007/s11616-017-0398-5

Yadav, V. & Bethard, S. (2019). *A survey on recent advances in named entity recognition from deep learning models* (Preprint). arXiv. https://arxiv.org/abs/1910.11470