Named Entity Recognition

Evelyn Gius

https://orcid.org/0000-0001-8888-8419

Abstract Dieses Kapitel führt in die automatische Erkennung von Entitäten in Texten, die so genannte *Named Entity Recognition* ein. Nach der Erläuterung von *Named Entities* werden erste Überlegungen zu ihrer Erkennung dargestellt. Es folgt ein kurzer Abriss über die Entwicklung von Named Entity-Systemen in der Sprachverarbeitung und die wichtigsten damit verbundenen Modelle. Anschließend wird die Anwendbarkeit der *Named Entity Recognition* in der Theologie beleuchtet und es werden praktische Hinweise zur Prüfung von *Named Entity*-Systemen gegeben. Das Kapitel schließt mit Hinweisen auf Tools und Ressourcen für die *Named Entity Recognition.*

Keywords *Named Entity Recognition,* Entitäten, Eigennamen, Sprachverarbeitung

1. Was ist *Named Entity Recognition*?

In der Sprachverarbeitung werden Wörter oder Ausdrücke in einem Text, die auf konkrete Entitäten in der Welt[1] verweisen, als *Named Entities* und ihre automatische Erkennung als *Named Entity Recognition* bezeichnet. Unter die *Named Entities* fallen insbesondere Ausdrücke für spezifische Personen, Orte oder Organisationen. Grundsätzlich gilt: *Named Entities* haben eine klar definierte Identität und können durch einen Namen oder einen spezifischen Begriff identifiziert werden, sie sind also benennbar. *Named Entities* umfassen neben Eigennamen auch andere Bezeichnungen. So sind etwa sowohl der Personen-Eigenname „Hildegard" als auch die spezifische Benennung „die Meisterin vom Rupertsberg" *Named Entities* des Typs Personenentität.

Die Erkennung von *Named Entities,* auch als *Named Entity Recognition* (NER) bezeichnet, ist ein wichtiger Schritt in der Textverarbeitung und -analyse. NER-Systeme identifizieren Entitäten und ordnen sie in vordefinierte Kategorien wie „Personen", „Orte" oder „Organisationen" ein. Dafür nutzen die Systeme meist maschinelle Lernverfahren, die auf linguistischen Informationen und semantischen Zusammenhängen in Texten basieren. Zudem greifen manche auf Listen bekannter Eigennamen zurück, welche in der Sprachverarbeitung *Gazetteers* genannt werden. Diese sind v. a.

1 Für viele Geisteswissenschaften ist wichtig zu ergänzen: Neben realweltlichen Entitäten können dies auch Entitäten in erzählten Welten sein.

Evelyn Gius: Named Entity Recognition, in: Christopher A. Nunn und Frederike van Oorschot (Hrsg.): Kompendium Computational Theology 1. Forschungspraktiken in den Digital Humanities, Heidelberg: heiBOOKS, 2024, S. 249–264. DOI: https://doi.org/10.11588/heibooks.1459.c21915

für Orte und andere geopolitische Entitäten hilfreich, wobei die Erkennungsergebnisse – evtl. entgegen unserer Intuition – bei der Nutzung weniger, hochfrequenten Eigennamen besser werden. Umfangreiche Listen weniger häufiger Eigennamen verschlechtern hingegen die Ergebnisse.

Wie eigentlich alle computationellen Verfahren in der Sprachverarbeitung wurden auch NER-Systeme anfangs regelbasiert entwickelt, während heutzutage maschinelle Lernverfahren bessere Ergebnisse erzielen, wobei dort anscheinend die früher in den Regeln beschriebenen Phänomene durchaus eine Rolle spielen (vgl. dazu den Abschnitt „Die Entwicklung von NER-Systemen"). Unabhängig von der eingesetzten Technik nutzen *Named Entity Recognition*-Systeme normalerweise so genannte Sequenztagging-Ansätze, in denen jedem Element einer Sequenz ein Wert zugewiesen wird. Es wird also etwa jedem Wort in einem Text die Information zugewiesen, ob es sich um eine *Named Entity* handelt und ggf., um welche Klasse. Genau genommen besteht die *Named Entity Recognition* also aus zwei Aufgaben: Aus dem Erkennen von *Named Entities* (Identifikation) und aus der Einordnung der erkannten *Named Entities* in die vordefinierten Klassen (Klassifikation).

Die tatsächlich genutzten *Named Entity*-Klassen unterscheiden sich von System zu System. Fast alle NER-Systeme erkennen die Klassen Personen, Orte und Organisation, welche typischerweise mit PER (vgl.: *person*), LOC (vgl.: *location*) und ORG (vgl.: *organization*) bezeichnet werden (Tab. 1). Die meisten Systeme haben noch eine vierte Klasse, welche häufig entweder geopolitische Entitäten (GPE, für *geo-political entity*) oder aber eine Resteklasse (MISC, für *miscellaneous*) umfasst. Neben Personen, Orten und Organisationen fallen grundsätzlich auch weitere Klassen und entsprechende Ausdrücke unter die *Named Entities.* Das sind beispielsweise Datumsangaben („17. September 1179") oder manchmal auch Mengenangaben („fünf Kilogramm"), abstrakte Begriffe („Religion") und allgemeine Klassen („Kloster").

Tabelle 1 Die drei häufigsten *Named Entity*-Klassen in NER-Systemen.

Klasse	Tag	Bsp. für Entitäten	Beispiel (*Named Entity* gefettet)
Person	PER	Menschen, Figuren, Heilige	**Abigajil** verhindert weitere Gewalt.
Ort	LOC	Städte, Berge, Länder, Gewässer	In **Südasien** gibt es zahlreiche Religionen.
Organisation	ORG	Firmen, Verbände, Institutionen	Die **römisch-katholische Kirche** ist die größte Kirche innerhalb des Christentums.

Im Kontext der Sprachverarbeitung ist die *Named Entitiy Recognition* eine gut etablierte und häufig genutzte Technik. NER ist zusammen mit anderen Grundoperationen – wie der Segmentierung des Textes in Wort- und Satzeinheiten *(Tokenization,*

Sentence Splitting) und der Auszeichnung von Worttypen und syntaktischen Einheiten (*Part-of-Speech-Tagging,* Dependenzparsing) – ein Vorverarbeitungsschritt in den meisten Sprachverarbeitungspipelines.[2] In maschinellen Lernverfahren werden *Named Entities* häufig als so genanntes *feature* genutzt, also als ein Aspekt, der in die Analyse von Texten bei den unterschiedlichsten Aufgaben mit einfließt, wobei die Systeme im Zuge des Lernverfahrens eine angemessene Berücksichtigung (die so genannte Gewichtung) des *features* errechnen.[3]

Typische Anwendungsmöglichkeiten der NER in der Sprachverarbeitung sind neben der eigentlichen Erkennung der *Named Entities* in Texten verschiedene, darauf aufbauende Verfahren, die weitere semantische Informationen in den Texten analysieren. Dazu zählen die Extraktion von Relationen der Entitäten zueinander (etwa familiäre Beziehungen zwischen Personen, räumliche Verhältnisse zwischen Orten oder Personen und Orten etc.), der Aufbau von Wissensgraphen in Datenbanken, die etwa für Suchmaschinen genutzt werden, oder die Erkennung von Ereignissen, die eine weitere semantische Analyse von Texten ermöglicht. Entsprechend vielfältig sind auch die Felder, in denen NER eingesetzt wird. Sie reichen von wissenschaftlicher Forschung über staatliche Institutionen bis hin zu Unternehmen. NER wird ebenso genutzt, um Marktanalysen zu erstellen, Kundenfeedback zu verfolgen oder geheimdienstliche Erkenntnisse zu möglichen Gefahren zu erlangen, wie um historische Texte zu analysieren und kulturelle Entwicklungen zu untersuchen.

2. Eine erste Annäherung an die Erkennung von Entitäten

Schauen wir uns einige Beispiele in Bezug auf mögliche textliche Besonderheiten an, anhand derer *Named Entities* systematisch identifiziert werden könnten.

> Beispiel 1:
> Und es begab sich in jenen Tagen, dass von **Kaiser Augustus** ein Erlass ausging, dass alle Welt geschätzt würde. Diese Schätzung war die allererste und geschah zur Zeit, als **Quirinius** Statthalter in **Syrien** war.

In diesem Beispiel werden zwei Personenentitäten („Kaiser Augustus", „Quirinus") und eine Ortsentität („Syrien") genannt. Als textliche Merkmale, die für ihre Erken-

2 Zum Aufbau von Sprachverarbeitungspipelines vgl. Biemann et al. (2022, 85 ff.). Die Einführung eignet sich darüber hinaus zur Vertiefung einiger weiterer, in diesem Beitrag genannten Verfahren der Sprachverarbeitung.

3 Zur Nutzung von *features* in maschinellen Lernverfahren vgl. Jurafsky & Martin (2023, 59; 60 ff.). Die Einführung eignet sich darüber hinaus zur Vertiefung der *Named Entity Recognition* sowie aller weiteren in diesem Beitrag genannten Verfahren der Sprachverarbeitung.

nung in Frage kommen, könnte man auf die Schreibung achten. Im Deutschen sind, wie in vielen anderen Sprachen, Eigennamen großgeschrieben.[4] Ein weiteres Merkmal für Personenentitäten könnte sein, dass Eigennamen normalerweise keinen Artikel haben, was sie von anderen Nomen unterschiedet. Beim Ausdruck „Kaiser Augustus" kommt außerdem der Titel „Kaiser" dazu, man könnte also auch eine Regel formulieren, dass Titel und darauffolgende großgeschriebene Wörter Personenentitäten bezeichnen. Auch „Syrien" ist ein Eigenname, der entsprechend an seiner Großschreibung erkannt werden kann. Hinzu kommt, dass bestimmte Präpositionen wie „in", „von" etc. auf eine Ortsentität verweisen können.

Beispiel 2:
Die **heilige Theresa von Avila** wurde im **16. Jahrhundert** in **Spanien** geboren. Ihre mystischen Erfahrungen führten zu bedeutenden Schriften.

In diesem Beispiel haben wir wieder drei Nennungen: eine Personenentität („heilige Theresa von Avila"), eine Datumsentität („16. Jahrhundert") und eine Ortsentität („Spanien"). Für die Personen- und die Ortsnennung können wir uns die bereits erwähnten Merkmale zunutze machen. Dabei ist das „heilige" ebenfalls eine Art Titel, wobei man ggf. für Heilige eigene Regeln aufstellen könnte, wie z. B. die Kombination aus dem vorangestellten Adjektiv „heilige[r]" mit „von" und einem Ortsnamen. Für das Datum könnte man eine Reihe von Formaten definieren, die typischerweise Zahlen, Interpunktionszeichen und zum Teil auch Wörter kombinieren, was sie von vielen anderen Ausdrücken unterscheidet.

Beispiel 3:
Franz von Assisi gründete den **Franziskanerorden** in **Italien.**

Die Nennungen im dritten Beispiel sind: eine Personenentität („Franz von Assisi"), eine Organisationsentität („Franziskanerorden") und eine Ortsentität („Italien"). Für „Franz von Assisi" könnte man eine Teilregel der Heiligenregel nutzen, nämlich das Schema [Eigenname] „von" [Ortsentität]. Für „Italien" gelten dieselben Regeln wie für „Syrien" und „Spanien" in den vorangegangenen Beispielen. Der „Franziskanerorden" hingegen ist eventuell daran erkennbar, dass der Ausdruck aus einem großgeschriebenen, aber nicht sehr häufigen Wort besteht, welches mit einem bestimmten Artikel eingeleitet wird. Vermutlich lässt sich auch aus der Zusammensetzung eine Regel ableiten, da „Franziskaner" ein von einem Personen-Eigennamen abgeleiteter Name ist und „Orden" eine allgemeine Organisationsbezeichnung.

4 Im Deutschen gilt allerdings die Großschreibung nicht nur für Eigennamen, sondern auch für Nomen, was dazu führt, dass auch viele andere Wörter das Merkmal aufweisen und die Erkennung von *Named Entities* weniger einfach macht als etwa im Englischen.

Die Überlegungen zu den drei Beispielen sollen zeigen, dass *Named Entities* anhand von textlichen Merkmalen von anderen Ausdrücken unterschieden werden können. Zu diesen Merkmalen gehören die Schreibung, die Nutzung von bestimmten Präpositionen oder andere Kombinationen an Wortartenfolgen, Merkmale auf Zeichenebene wie z. B. Großschreibung, bestimmte Buchstabenfolgen oder die Verwendung von für andere Wortarten untypischen Zeichen wie Ziffern oder Interpunktion, die syntaktische Struktur (wo im Satz erwartet man *Named Entities?*) oder auch typische Kontexte oder Vorkommenshäufigkeiten der *Named Entities.*

3. Die Entwicklung von NER-Systemen

Die Entstehung der *Named Entity Recognition* (NER) geht auf die Anfänge der computationellen Verarbeitung von natürlicher Sprache in den 1950er und 1960er Jahren zurück. In dieser Zeit begann die Entwicklung von Textverarbeitungssystemen, mit denen grundlegende sprachliche Informationen analysiert werden konnten. Insgesamt entspricht die Geschichte der NER der Entwicklung vieler Sprachverarbeitungs-Anwendungen, die von der regelbasierten Erkennung der Phänomene über maschinelle Lernverfahren bis zu *Deep Learning*-Ansätzen reicht.

Die ersten Ansätze der NER konzentrierten sich hauptsächlich auf die Identifizierung von Personen- und Ortsnamen. Für diese definierten sie Regeln oder Muster, die auf spezifische Eigenschaften von Eigennamen wie in den oben diskutierten Beispielen abzielten. Diese sogenannten regelbasierten Verfahren ermöglichten es, Namen in Texten anhand bestimmter Merkmale wie Großschreibung oder besonderen Zeichen zu identifizieren. Allerdings sind solche heuristischen Ansätze begrenzt und erzielten aufgrund der Vielfalt von *Named Entities* und Kontexten, in denen diese vorkommen, keine zufriedenstellenden Ergebnisse.

Die Nutzung der in den 1990er Jahren aufkommenden maschinellen Lernverfahren in der NER führten zu einer Verbesserung der Systeme.[5] Statistische Modelle und *Machine-Learning*-Algorithmen wurden verwendet, um *Named Entities* anhand von vorher manuell annotierten Trainingsdaten zu erkennen und zu klassifizieren. Dabei kamen insbesondere *Hidden Markov*-Modelle und *Maximum Entropy*-Modelle zum Einsatz, welche Kontextinformationen und statistische Wahrscheinlichkeiten bei der NER berücksichtigen können. *Hidden Markov*-Modelle (HMMs) können die Sequenz von Wörtern in einem Text analysieren und die Wahrscheinlichkeit berechnen, mit der ein Wort eine *Named Entity* ist. Sie nehmen dafür verdeckte *(hidden)* Zustände an, die am Anfang unbekannt sind (im Fall von NER: die Entitäten), sowie beobacht-

5 Für einen kurzen Überblick anhand der entwickelten Systeme von den Anfängen bis zu den aktuellen *Transformer*-basierten Ansätzen vgl. Jurafsky & Martin (2023, 183).

bare Zustände, die aus den Wörtern im Text bestehen. Das Modell wird trainiert, um die Übergangswahrscheinlichkeiten zwischen diesen Zuständen und entsprechend die Ausgabewahrscheinlichkeiten für jedes Wort zu optimieren. Die Erkennung von Entitäten basiert auf den so ermittelten wahrscheinlichsten Zustandsübergängen, die eine Verbindung zwischen den unbekannten (bzw. *hidden*) Entitätsklassen und den beobachtbaren Wörtern herstellen. Auch *Maximum Entropy*-Modelle (MaxEnt) sind probabilistische Modelle. Sie beruhen auf Prinzipien der maximalen Entropie, mit der Wahrscheinlichkeiten für eine Reihe von Klassen oder Kategorien optimiert werden. Bei der NER sagen *Maximum Entropy*-Modelle die Zugehörigkeit zu den *Named Entity*-Kategorien vorher, indem sie trainierte Gewichtungen geeigneter Textmerkmale verwenden. Diese Merkmale könnten Wörter, Kontextinformationen, Groß- oder Kleinschreibung usw. sein. Das Ziel besteht darin, die Gewichtungen der Merkmale so anzupassen, dass die Wahrscheinlichkeit für jede Entitätskategorie im Sinne einer maximalen Entropie berechnet wird.

Mit der Etablierung von *Deep Learning* wurde die Leistungsfähigkeit von NER-Systemen weiter gesteigert. Nun wurden so genannte künstliche neuronale Netze, welche Phänomene in einer sehr großen Datenmenge anhand einer Vielzahl an Schichten „lernen", für die Erkennung von *Named Entities* genutzt. Als erste *Deep Learning*-Systeme wurden *rekurrente neuronale Netzwerke* (RNNs) genutzt. Diese gibt es schon seit den 1980er Jahren, sie wurden allerdings erst in den 2000ern in der NER eingesetzt. RNNs sind neuronale Netzwerke, die speziell für die Verarbeitung sequenzieller Daten entwickelt wurden. Sie können verwendet werden, um die Sequenz von Wörtern in einem Text zu verarbeiten und für jedes Wort die Wahrscheinlichkeit zu berechnen, mit der es einer bestimmten Klasse angehört. Anders als die oben genannten Modelle (HMMs und MaxEnt), berücksichtigt das RNN dabei auch den Kontext der jeweils vorherigen Wörter. Dies ermöglicht eine bessere Erkennung von Entitäten. RNNs haben jedoch Schwierigkeiten bei der Verarbeitung langer Sequenzen. Die nächste Entwicklung, *Long Short-Term Memory*-Netzwerke (LSTM), konnte hingegen auch langfristige Abhängigkeiten in Sequenzen erfassen. In der NER ermöglichen LSTMs eine präzisere Modellierung von Zusammenhängen zwischen Wörtern und die Erkennung von Entitäten, die über längere Abschnitte hinweg variieren können. Sie sind in der Lage, sowohl lokale als auch globale Kontextinformationen effektiv zu nutzen, und haben die NER entsprechend vorangebracht. Eine weitere Verbesserung von NER-Systemen wurde in den frühen 2010er Jahren durch die Kombination von bi-direktionalen LSTM-Modellen mit *Conditional Random Fields* (CRFs) möglich. Bi-direktionale LSTMs erfassen nicht nur den Kontext der Sequenz vor einem bestimmten Wort, sondern auch den Kontext nach dem Wort, was die Qualität der Erkennung von *Named Entities* erhöht. Die Verwendung von CRFs hilft bei der Modellierung von Abhängigkeiten zwischen benachbarten Wörtern und ihrer Klassifizierung. Dies ermöglicht eine kohärentere Zuordnung von Entitätslabels. Auch die Erkennung von verschachtelten Entitäten wurde durch die Nutzung von *Deep Learning*-Methoden deutlich verbessert.

Die heutigen *State-of-the-Art*-Systeme für NER basieren auf vortrainierten *Transformer*-Modellen wie BERT oder GPT, die seit Mitte der 2010er entwickelt werden. *Transformer* sind eine Weiterentwicklung rekurrenter Netzwerke, in denen insbesondere durch die so genannten *self attention*- sowie *memory*-Schichten kontextabhängige Informationen gleichzeitig berechnet werden können. Diese Netzwerke können deshalb auf einer großen Menge von Textdaten trainiert werden und erzeugen entsprechend noch bessere Sprachmodelle. Durch sogenanntes *Transfer Learning* können diese generellen Modelle dann für spezialisierte Aufgaben wie NER angepasst werden.

Unabhängig von den genutzten Modellen hat sich ab den späteren 1990er Jahren die so genannte BIO-Annotation (Ramshaw & Marcus 1995) für NER durchgesetzt, die seitdem als Standardansatz für die Sequenzauszeichnung bei einem Spannenerkennungsproblem gilt. Der Ansatz stellt drei Labels zur Verfügung, die auch die Grenzen der *Named Entities* erfassen. Damit wird jedes Wort (bzw. *Token*) eines *Named Entity*-Ausdrucks wie folgt erfasst: Das erste Wort bekommt das Label B (für *begin*), alle evtl. folgenden Wörter werden mit I (für *inside*) ausgezeichnet und alle Wörter außerhalb der *Named Entity* mit O (für *outside*). Dabei gibt es für jede Entitätsklasse eigene B- und I-Labels, um diese abzubilden. Für den Anfang von Beispiel 2 würde eine BIO-Annotation entsprechend wie folgt aussehen:

Die	heilige	Theresa	von	Avila	wurde	im	…
O	B-PER	I-PER	I-PER	I-PER	O	O	

Abb. 1 Sequenzkodierung einer *Named Entity* des Typs Person (PER) mit BIO-Labels

4. Herausforderungen der automatischen Entitätenerkennung

Auch wenn die NER zu den grundlegenden Verfahren der Sprachverarbeitung gehört und die aktuellen NER-Systeme zudem gute Ergebnisse erzielen, so gibt es dennoch einige weiterhin bestehende Herausforderungen bei der Erkennung von *Named Entities*.

Die sprachliche Form von *Named Entities* ist sehr variantenreich. Die breite Palette an Flexionen, Derivationen und Morphemen oder syntaktischen Regeln und Wortreihenfolgen in einer Sprache erhöhen die Komplexität der Erkennung. Deshalb ist die NER in morphologisch reichen Sprachen wie etwa dem Hebräischen besonders schwierig. Hinzu kommt ein praktisches Problem: NER-Systeme basieren in der Regel auf umfangreichen Trainingsdaten, deshalb hängt die Leistungsfähigkeit der Systeme von der Verfügbarkeit ausreichend geeigneter Daten in der entsprechenden Sprache ab. Die Entwicklung universell anwendbarer NER-Systeme wird schließlich von den sprach- und kulturabhängigen Unterschieden in Grammatik, Syntax und Nomenklatur, also der Art der Benennung von Entitäten, erschwert.

Named Entities sind nicht nur oft Mehrwortphrasen – wie „Universität Tübingen“ oder „Maria Magdalena“ –, sie sind zudem teilweise verschachtelt. Für die korrekte Erkennung von Entitäten wie „Apostel Paulus“ oder „Hildegard von Bingen“ müssen nicht nur die zum Eigennamen gehörigen Wörter als Titel („Apostel“), Ort („Bingen“) etc. erkannt, sondern auch als zur Personenentität gehörig bestimmt werden. Dafür braucht es eine tiefere semantische Verarbeitung und eine bessere Modellierung von Zusammenhängen im Text.

Die Tatsache, dass *Named Entities* auch Mehrwortphrasen sein können, macht zudem die Evaluation von NER-Systemen komplexer als bei einheitlichen Segmenten. Im Gegensatz zu etwa dem *Part of Speech-Tagging*, bei welchem jedem einzelnen Wort ein Wert zugewiesen wird, oder zu Klassifikationsaufgaben, welche für ganze Texte ausgeführt werden, muss bei der NER auch die Textspanne bestimmt werden, die die jeweilige Entität umfasst. Nachdem in der NER typischerweise Wörter die Trainingseinheit, die Ausgabeeinheit aber die Entitäten – und damit potenziell: Mehrwortausdrücke – sind, gibt es hier eine Nichtübereinstimmung. Entsprechend werden im oben skizzierten BIO-Annotationssystem nur teilweise erkannte Mehrwort-Entitäten mehrfach falsch gewertet, weil die Annotationen aufgrund der fehlenden Wörter falsch sind (die B-Annotation kommt ein oder mehrere Wörter zu spät bzw. die O-Annotation zu früh, mit entsprechenden Konsequenzen für die I-Annotationen). Dies ist insofern problematisch, als die Nichterkennung derselben Entität gleich gewertet und damit als gleich gut bzw. schlecht gewertet würde. Dieses Problem kann man allerdings durch eine entsprechende Fehlergewichtung in der Evaluation abschwächen.

Ein anderes Evaluationsproblem ist ein grundsätzliches Problem datenarmer Sprachen wie etwa vormoderner Sprachen: Für diese gibt es häufig keine weiteren annotierten Daten, die man als so genannte *Benchmarks* nutzen kann, um zu prüfen, ob das evaluierte NER-System auch bei unbekannten Texten ähnlich gute Ergebnisse erzielt oder ob ein sogenanntes *Overfitting* auf die Trainingsdaten vorliegt, wodurch nur diese gut erkannt werden.

Schließlich gibt es eine Reihe von Herausforderungen, die eher Desiderate sind. So erkennt ein NER-System die für eine Fragestellung relevanten Entitäten im Text nur, wenn diese explizit und mit klar definierten Namen oder Ausdrücken benannt werden. Es eignet sich deshalb u. a. nicht für die Erkennung von Pronomen, generischen Ausdrücken, unspezifischen Begriffen und indirekten Referenzen, die sich auf *Named Entities* beziehen. Hinzu kommen Schwierigkeiten bei der Erkennung von Entitäten wie abstrakten Konzepten und bei Entitäten, wenn diese durch generell nicht sehr häufig auftretende Fachbegriffe oder lokale Namen benannt werden.

Während die letztgenannten Schwierigkeiten von NER-Systemen durchaus angegangen werden, ist die Erkennung von Pronomen etc. nicht Teil der NER. Dies hat u. a. pragmatische Gründe, denn dafür müssten zusätzliche Herausforderungen der damit eng verbundenen Aufgabe der Koreferenzauflösung ebenfalls gelöst werden. Bei der Koreferenzauflösung geht es darum, zu bestimmen, wann Pronomen, demonstrative Ausdrücke oder andere referenzielle Elemente im Text auf bereits zuvor

erwähnte Entitäten verweisen. Dies erfordert ein noch weitergehendes Verständnis des Textkontexts und der semantischen Beziehungen. Hinzu kommen – gerade im Bereich der Theologie – durchaus relevante Identitätsfragen, da alle auf dieselbe Entität verweisenden Ausdrücke identifiziert werden müssen. Dies ist auch in weniger offensichtlichen Fällen als der Dreifaltigkeit ein oft kniffliges Problem, weil die Identität von vielen Entitäten z. B. bei zeitlichen oder anderen Veränderungen schwer zu fassen ist. So kann eine Denkschule über Jahrzehnte hinweg als dieselbe aufgefasst oder in bestimmte Abschnitte – und entsprechend mehrere Organisationsentitäten – unterteilt werden, eine Familie eine einzelne Entität sein oder das Hinzukommen und Wegfallen von Familienmitgliedern jeweils als neue Familien aufgefasst werden oder die Lebensphasen einer Person mit stark unterschiedlichen Anschauungen und Handlungen können auch als getrennte Personenentitäten aufgefasst werden.[6]

5. NER in der Theologie?

Da im Prinzip alle Techniken aus der automatischen Sprachverarbeitung in auf die Analyse von Texten ausgerichteten Wissenschaften anwendbar sind, kann auch die NER in theologischen Kontexten genutzt werden.[7] Grundsätzlich ist die Anwendung in all jenen Bereichen möglich und sinnvoll, in denen Entitäten wie Personen, Orte, Daten oder Konzepte bzw. deren Verhältnis zueinander für ein Forschungsinteresse relevant sind. Potenzielle Einsatzfelder reichen von der Identifikation spezifischer Phänomene in einzelnen Texten bis zur Analyse von großen Textmengen bzw. Korpora. Die NER eignet sich neben der Identifikation der entsprechenden *Named Entities* auch für die Analyse von Fragen der Verteilung, des Verhältnisses zueinander und zeitliche Entwicklungen. Entwicklungen können zudem im Vergleich zwischen verschiedenen Textgruppen betrachtet werden oder für eine Gruppierung von Texten anhand von *Named Entities.* Eine *Named Entitiy*-basierte Analyse kann die Frage nach den häufigsten Erwähnungen von Akteur*innen oder Orten in religiösen Texten oder auch den quantitativen Vergleich der jeweiligen Anteile der Erwähnungen zwischen verschiedenen Texten oder Textgruppen zum Ziel haben. Auch Fragen der ersten Nennung und anschließenden Entwicklung der Nennungshäufigkeit von Personen, Orten oder auch Konzepten in einem Korpus aus diachronen Texten, also Texten, die eine größere Zeitspanne abdecken, können analysiert werden. Mit einer NER kann

6 Für eine tiefergehende, allgemeine Betrachtung lohnt sich ein Blick in die *Stanford Encyclopedia of Philosophy;* zur Identitätsproblematik vgl. Noonan & Curtis (2022) zur Problematik fiktionaler Entitäten vgl. Kroon & Voltolini (2023).

7 Für eine Übersicht von Sprachverarbeitungsmethoden in den Geisteswissenschaften vgl. Piotrowski (2012) und Sporleder (2010) sowie die Methoden-Einführungen im forTEXT-Portal unter https://fortext.net/routinen/methoden, zuletzt aufgerufen am 17.06.2024.

man stilistische Analysen – etwa in der Homiletik – durchführen oder Texte in einem Korpus identifizieren, die eine bestimmte, über *Named Entities* erkennbare Thematik betreffen.

Diese und weitere Anwendungen können mit der NER zu durchaus interessanten Erkenntnissen führen. In der Theologie ist die *Named Entity Recognition* trotzdem bislang wenig verbreitet und hat auch in für digitale Ansätze einschlägigen Publikationen bisher keine erkennbare Bedeutung.[8] Die Gründe dafür sind vermutlich vielfältig. Zunächst ist die Anwendung von Sprachverarbeitungs-Techniken in den Geisteswissenschaften jenseits der Computer- und Korpuslinguistik generell ein noch recht junger Forschungszweig. Außerdem gibt es in der Theologie genauso wie in anderen, eher exemplarisch oder hermeneutisch arbeitenden Geisteswissenschaften vermutlich eine gewisse Zögerlichkeit gegenüber dem Einsatz von computationellen Mitteln. Schließlich ist die so genannte Operationalisierung einer Fragestellung, also die Übersetzung der Frage in durch *Named Entities* messbare Qualitäten, keine triviale Aufgabe, die zudem methodologisch eher konträr zu den etablierten Praktiken der theologischen Textanalyse steht. Betrachtet man allerdings die neueren Entwicklungen im Feld der *Digital Theology,* so kann man davon ausgehen, dass auch im Feld der computationellen Theologie in den kommenden Jahren einige Fortschritte gemacht und auch NER-Verfahren genutzt werden. Doch auch wenn evtl. Vorbehalte ausgeräumt und die nötigen Kompetenzen für die Umsetzung einer NER vorhanden sind, gibt es Einschränkungen der Qualität der Analysen, die berücksichtigt werden müssen. Der mögliche Umgang mit diesen wird nachfolgend skizziert.

6. Hinweise zur Anwendung von NER-Systemen

Wie die meisten Sprachverarbeitungsverfahren werden auch NER-Systeme typischerweise für das Englische und anhand von Nachrichtenartikeln oder im Internet auffindbaren Texten entwickelt. Deshalb stehen für Sprachen jenseits des Englischen

8 So wird NER zum Beispiel in Heyden & Schröder (2020) oder Sutinen & Cooper (2021) und den noch recht neuen Publikationsreihen *Introductions to Digital Humanities – Religion* (herausgegeben von Claire Clivaz, Frederik Elwert, Kristian Petersen, Ortal-Paz Saar und Jeri Wieringa) und *Digital Biblical Studies* (herausgegeben von Claire Clivaz und Ken M. Penner) kaum einmal erwähnt und kein einziges Mal angewendet. Auch Suchen in Katalogen blieben praktisch ergebnislos: Eine Recherche nach NER in der Religionswissenschaftlichen Bibliographie des Fachinformationsdienst (FID) Religionswissenschaft unter https://www.relbib.de ergibt nur einen Treffer (Blouin 2021), welcher zwar potenziell relevant, aber nicht einschlägig ist. Unter den Ergebnissen der Suche im Katalog der Universität Frankfurt nach „Named Entity Recognition“ in den Geisteswissenschaften ist kein theologischer Titel. Auch wenn hierbei bestimmt einzelne Publikationen durch das Raster gefallen sind, so deutet das doch mindestens auf eine bislang geringe Relevanz von NER in der Theologie.

oder für andere Textsorten als Nachrichten- und Internettexte meist weniger gute Systeme zur Verfügung. Zudem sind die Ergebnisse abhängig von der *Named Entity*-Klasse oft sehr unterschiedlich. Während die klassischen Kategorien für Personen, Orte und z. T. auch Organisationen zumeist recht gut erkannt werden und in den besten Systemen Erkennungsraten von deutlich über 90 % erreichen, ist die Erkennungsqualität für andere Kategorien häufig erheblich geringer. Nichtsdestotrotz kann die NER auch in Fällen eingesetzt werden, in denen die Systeme nicht optimal arbeiten, wenn man dies entsprechend vorbereitet und umsetzt.

In jedem Fall sollte man vor der Anwendung eines NER-Systems beurteilen, inwiefern die Qualität der Erkennung ausreichend ist, um auf den Ergebnissen aufbauende belastbare Aussagen zu machen. In der Sprachverarbeitung gelten Ergebnisse mit einem so genannten F1-Wert von 0,8 und mehr als sehr gut, Ergebnisse von 0,95, die für das Englische – und vereinzelt auch für andere Sprachen wie u. a. das Deutsche – in der NER mittlerweile erreicht werden, als (nahezu) perfekt. Der F1-Wert setzt sich dabei aus den Werten für die Maße für *Recall* und *Precision* zusammen. Ein F1-Wert von 0,8 bedeutet entsprechend, dass der durchschnittliche Anteil der im Text vorhandenen Phänomene, die gefunden werden *(Recall),* und der korrekt identifizierten Stellen unter den gefundenen Stellen *(Precision)* bei 80 % liegt. Da der Wert ein Durchschnittswert aus den beiden Werten ist und diese wiederum für mehrere Unterkategorien (Personen, Orte, Organisationen etc.) berechnet werden, sagt der F1-Wert allerdings noch nichts über die Qualität der Erkennung für spezifische Aspekte aus. Für die tatsächliche Qualität der Anwendung ist der F1-Wert also – wie übrigens jedes Evaluationsmaß – nur ein Richtwert. Er sagt meist nichts über die Eignung des Systems für das konkrete Forschungsinteresse aus. Es gilt deshalb zu prüfen, inwiefern ein System geeignete Ergebnisse für die Forschungsfrage und das genutzte Textkorpus liefert. Eine solche Qualitätsprüfung ist umso wichtiger, wenn man weitere, auf die NER aufbauende Schritte automatisiert umsetzt, wie etwa die bereits erwähnte Erkennung von Relationen von Entitäten oder auch die so genannte Koreferenzauflösung, bei der alle Entitäten-Benennungen und weitere mögliche Referenzen – wie etwa Pronomen – auf ein und dieselbe Entität erkannt werden.

Wenn ein NER-System für Texte eingesetzt wird, die sich von den bei der Entwicklung des Systems genutzten und evaluierten Texten unterscheiden, sollte deshalb vorher eine spezifische Überprüfung der Erkennungsqualität erfolgen. Im Idealfall sollte das System anhand annotierter Testdaten aus dem genutzten Korpus evaluiert werden, also ein für den konkreten Forschungsbedarf aussagekräftiger F1-Wert erstellt werden. Auf jeden Fall sollte aber zumindest eine Stichprobenprüfung der ausgegebenen Ergebnisse und einzelner Textteile in Bezug auf die dort gefundenen und nicht gefundenen Phänomene erfolgen. Durch die Stichprobenprüfung kann eingeschätzt werden, ob ein System die gesuchten Phänomene korrekt erkennt und inwiefern es ggf. falsche Phänomene miteinschließt. Außerdem können mögliche systematische Fehler erkannt werden, etwa, ob ein Ortsname meistens falsch als Personennamen, bestimmte Mehrwortausdrücke nicht oder nur teilweise oder einzelne

Bezeichnungen tendenziell nicht erkannt werden. Solche Fehler können die Textanalyse je nach Art des Fehlers stark verfälschen. Will man zum Beispiel die Relevanz von bestimmten Konzepten in Texten vergleichen, sollte sichergestellt werden, dass nicht eines der Konzepte deutlich schlechter erkannt wird als die anderen und entsprechend nur deswegen vermeintlich weniger häufig in den Texten zu finden ist.

Zeigt sich, dass die Qualität des Systems nicht zufriedenstellend ist und man es nicht ohne weiteres für die automatische Analyse nutzen kann, gibt es trotzdem zwei Möglichkeiten, es einzusetzen. Beide beinhalten eine weitere manuelle Prüfung der Ergebnisse und sichern so eine auf diesen aufbauende Analyse ab. Erstens kann jedes System als heuristisches System eingesetzt werden und so auf ggf. interessante Aspekte der untersuchten Texte hinweisen. Auch wenn man die Ergebnisse der NER nicht quantitativ auswertet – was man bei nicht ausreichend guten Ergebnissen ohnehin nicht sollte – kann man ihre Ergebnisse als Hinweis auf potenziell interessante Texte oder Textstellen nutzen. Vielleicht findet man mit der NER einen Text, der vorher noch nicht in einem bestimmten Kontext als relevant erkannt wurde, oder man stößt auf Bezeichnungen, die bisher nicht betrachtet wurden, obwohl sie zu einer bestimmten Zeit oder in bestimmten Texten durchaus prominent und relevant waren. Eventuell zeigen sich in den Ergebnissen auch Verbindungen durch gemeinsames Auftreten von Entitäten, die man bisher nicht in den Blick genommen hat.

Für eine automatische Analyse nicht ausreichend gute NER-Systeme lassen sich zweitens auch als Vorverarbeitungsschritt nutzen, der Daten für eine anschließende manuelle Weiterbearbeitung liefert. Insbesondere wenn ein System einen akzeptablen *Recall* hat, also einen guten Anteil der gesuchten Phänomene auch findet, kann die Qualität der Daten durch eine manuelle Bearbeitung deutlich gesteigert werden. Dafür werden aus den Ergebnissen jene aussortiert, die falsch sind. Die verbleibenden Daten können dann in weiteren – ggf. ebenfalls manuell unterstützten – Analyseschritten oder für eine quantitative Auswertung genutzt werden. Dies ist etwa für die Koreferenzauflösung in längeren Texten ein gangbarer Weg, weil die Überprüfung und Korrektur der sogenannten Koreferenzketten einen vergleichsweisen kleinen Aufwand bedeutet. Die manuelle Bearbeitung besteht im Wesentlichen darin, in den Koreferenzketten, die alle Erwähnungen einer Entität in einem Text enthalten, die falschen Erwähnungen von Entitäten zu korrigieren und eventuell aufgrund von Erkennungsfehlern getrennte Ketten wieder zusammenzufügen. Je nach Erkenntnisgewinn, welchen die so aufbereiteten Daten ermöglichen, sind derartige manuelle Prüfungen eine Möglichkeit, die man in Betracht ziehen sollte.

7. Tools und Ressourcen

Es gibt eine schier unübersichtliche Menge von NER-Systemen, die in den letzten Jahrzehnten entwickelt wurden. Bei der Auswahl der Systeme und Plattformen sollte man beachten, dass regelbasierte NER-Methoden oft für einfachere Fälle geeignet sind, während komplexere Szenarien maschinelles Lernen erfordern könnten. Außerdem sollten idealerweise mehrere Systeme auf denselben Daten getestet werden, um das am besten geeignete System zu identifizieren. Zurzeit werden vor allem drei *open source*-Systeme in Anwendungen häufig genutzt: das *Natural Language Toolkit* (NLTK)[9], *spaCy*[10] und der *Stanford Named Entity Recognizer*[11]. Alle drei erzielen gute Ergebnisse für verschiedene natürliche Sprachen, werden regelmäßig aktualisiert und ihre Python- bzw. Java-basierte Modelle sind vergleichsweise einfach anzuwenden. Eine Suche nach weiteren, sprachspezifischen NER-Systemen kann aber lohnenswert sein.[12] Insbesondere für (noch) nicht versierte Nutzer*innen sind außerdem Plattformen interessant, mit denen man selbst eine Verarbeitungspipeline zusammenstellen kann. So ist etwa die deutsche Plattform *WebLicht* für Angehörige vieler wissenschaftlicher Einrichtungen frei zugänglich und stellt sowohl für die Vorverarbeitung als auch für die NER selbst verschiedene Systeme zur Verfügung, die auf einer grafischen Oberfläche kombiniert und auf zur Verfügung gestellte oder eigene Texte in vielen Sprachen angewendet werden können.[13]

Wenn man selbst NER-Systeme weiterentwickeln möchte, sollte man geeignete Daten auswählen. Es gibt eine Reihe von annotierten Korpora, die man je nach Anwendungsfeld nachnutzen kann, wie etwa das englischsprachige Korpus für literarische Texten von Bamman et al. (2019) oder die (u. a.) deutschsprachigen Zeitungstextkorpora von Tjong Kim Sang & De Meulder (2003) und Benikova et al. (2014). Eine weitere Annotation bereits vorverarbeiteter Daten für die NER ist insbesondere für ressourcenärmere Sprachen möglicherweise sinnvoll (z. B. für Latein beim bereits mit Informationen zu Lemmatisierung und *Part-Of-Speech*-Tagging angereicherten *EvaLatin* Korpus von Sprugnoli et al. 2020).

9 Vgl. https://www.nltk.org. Für die Anwendung von NER vgl. https://www.nltk.org/book/ch07.html. Alle in diesem Abschnitt genannten Adressen wurden zuletzt am 17.06.2024 geprüft.

10 Vgl. https://spacy.io/models. Für die Anwendung von NER vgl. https://spacy.io/universe/project/video-spacys-ner-model-alt.

11 Vgl. https://stanfordnlp.github.io/CoreNLP/ner.html. Für die Anwendung der Pipeline vgl. https://stanfordnlp.github.io/CoreNLP/pipeline.html.

12 So gibt es gute Ansätze u. a. zu Latein (z. B. Erdmann et al. 2016), Altgriechisch (z. B. Yousef et al. 2022), Hebräisch (z. B. Bareket & Tsarfaty 2021) und zu vormodernen bzw. klassischen Sprachen (z. B. Johnson et al. 2021 und Burns 2019).

13 Vgl. https://weblicht.sfs.uni-tuebingen.de/weblicht. Für die Beschreibung der zur Verfügung stehenden NER-Modelle vgl. https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Tools_in_Detail#Named_Entity_Recognition.

Für die Erstellung von *Gazetteers* können oft bestehende Verzeichnisse nachgenutzt werden. Grundsätzlich bieten sich dafür große, idealerweise frei unter entsprechenden Lizenzen wie der Creative-Commons-Lizenz verfügbare Verzeichnisse an. So können entsprechende *Wikipedia*-Kategorien genutzt werden (wie etwa Mann, Frau, Figur, Heiliger für Personennamen oder entsprechende Kategorien für Ortsnamen etc.), um an Entitätennamen zu kommen.[14] Eine weitere Quelle ist die Gemeinsame Normdatei (GND), welche Normdaten aus Katalogdaten zu Personen und anderen Bereichen in einer Reihe von Metadaten- und Datendiensten zur Verfügung stellt.[15] Außerdem lohnt es sich, nach spezifischen Daten zu suchen. Für historische Texte kann etwa das Personenlexikon zur Führungsschicht des Römischen Reiches in der Frühen und Hohen Kaiserzeit der Berlin-Brandenburgischen Akademie der Wissenschaften[16] oder das Lexikon griechischer Personennamen der Universität Oxford interessant sein.[17] Auch Institutionen wie die EU oder einzelne Staaten stellen zahlreiche für die NER relevante Daten zur Verfügung. So bietet die EU neben einer Reihe von anderen Informationen auch ein großes Namensverzeichnis[18] und das U. S. Geological Survey (USGS) diverse Daten zu Orten und andere geologischen Informationen an.[19] Viele Verzeichnisse lassen sich außerdem leicht mit einer entsprechenden Internetsuche finden.

Als Einstieg in die NER empfiehlt sich allerdings eher die auch für Anfänger*innen geeignete deutschsprachige Übung von Schumacher (2019) zur Anpassung des *Stanford Named Entity Recognizer* für einen literarischen Text oder die englischsprachige Übung von Grunewald et al. (2022), die niedrigschwellig in eine *Python*-Analyse von Orten in Daten zu Kriegsgefangenen einführt und auch das Einbinden eines *Gazetteers* erläutert.

Literaturverzeichnis

Bamman, D., Popat, S., & Shen, Sh. (2019). An Annotated Dataset of Literary Entities. In *Proceedings of the 2019 Conference of the North* (S. 2138–2144). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1220 [zuletzt aufgerufen am 17. 06. 2024].

14 Diese und andere Informationen können strukturiert über Wikidata abgefragt werden, vgl. https://www.wikidata.org.

15 Vgl. etwa den GND-Dienst *Entity Facts* unter https://www.dnb.de/DE/Professionell/Metadatendienste/Datenbezug/Entity-Facts/entityFacts_node.html.

16 Vgl. *Prosopographia Imperii Romani saec. I. II. III,* verfügbar unter https://pir.bbaw.de.

17 Vgl. https://www.lgpn.ox.ac.uk.

18 Vgl. https://data.jrc.ec.europa.eu für eine Übersicht https://data.jrc.ec.europa.eu/dataset/jrc-emm-jrc-names für das Namensverzeichnis.

19 Vgl. https://www.usgs.gov/products/data/all-data.

Bareket, D., & Tsarfaty, R. (2021). Neural Modeling for Named Entities and Morphology (NEMO2), *Transactions of the Association for Computational Linguistics,* 9, 909–928. https://doi.org/10.1162/tacl_a_00404 [zuletzt aufgerufen am 17.06.2024].

Benikova, D., Biemann, Ch., & Reznicek, M. (2014). NoSta-D Named Entity Annotation for German. Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (S. 2524–2531). Reykjavik: European Language Resources Association. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf [zuletzt aufgerufen am 17.06.2024].

Biemann, Ch., Heyer, G., & Quasthoff, U. (2022). *Wissensrohstoff Text. Eine Einführung in das Text Mining,* 2. Wesentlich überarbeitete Auflage. Lehrbuch. Wiesbaden [Heidelberg]: Springer Vieweg. https://doi.org/10.1007/978-3-658-35969-0 [zuletzt aufgerufen am 17.06.2024].

Blouin, B., Magistry, P., & Van Den Bosch, N. (2021). Creating Biographical Networks from Chinese and English Wikipedia, *Journal of Historical Network Research,* 5(1), 303–317. https://doi.org/10.25517/JHNR.V5I1.120 [zuletzt aufgerufen am 17.06.2024].

Burns, P. J. (2019). Building a Text Analysis Pipeline for Classical Languages. In M. Berti (Hrsg.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution* (S. 159–176). Berlin/Boston: De Gruyter Saur [= *Age of Access? Grundfragen der Informationsgesellschaft,* 10]. https://doi.org/10.1515/9783110599572-010 [zuletzt aufgerufen am 17.06.2024].

Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification in Historical Documents. A Survey, *ACM Computing Surveys,* 56(2), 1–47. https://doi.org/10.1145/3604931 [zuletzt aufgerufen am 17.06.2024].

Grunewald, S., & Janco, A. (2022). Finding Places in Text with the World Historical Gazetteer, *Programming Historian,* 11, o. S. https://doi.org/10.46430/phen0096 [zuletzt aufgerufen am 17.06.2024].

Heyden, K., & Schröder, B. (Hrsg.) (2020), *Theologie Im Digitalen Raum,* Gütersloh: Gütersloher Verlagshaus [= *Verkündigung und Forschung* 65(2)].

Johnson, K. P., Burns, P. J., Stewart, J., Cook, T., Besnier, C., & Mattingly, W. J. B. (2021). The Classical Language Toolkit. An NLP Framework for Pre-Modern Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. System Demonstrations* (S. 20–29). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-demo.3 [zuletzt aufgerufen am 17.06.2024].

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3rd Edition [Draft]. URL: https://web.stanford.edu/~jurafsky/slp3 [zuletzt aufgerufen am 17.06.2024].

Kroon, F., & Voltolini, A. (2023). Fictional Entities. In E. N. Zalta & U. Nodelman (Hrsg.), *The Stanford Encyclopedia of Philosophy.* Stanford University: Metaphysics Research Lab. URL: https://plato.stanford.edu/archives/fall2023/entries/fictional-entities [zuletzt aufgerufen am 17.06.2024].

Noonan, H., & Curtis, B. (2022). Identity. In E. N. Zalta & U. Nodelman (Hrsg.), *The Stanford Encyclopedia of Philosophy.* Stanford University: Metaphysics Research Lab. URL: https://plato.stanford.edu/archives/fall2022/entries/identity [zuletzt aufgerufen am 17.06.2024].

Piotrowski, M. (2012). NLP Tools for Historical Languages. In Ders. (Hrsg.), *Natural Language Processing for Historical Texts* (S. 85–100). Cham: Springer International Publishing [= *Synthesis Lectures on Human Language Technologies*]. https://doi.org/10.1007/978-3-031-02146-6_7 [zuletzt aufgerufen am 17.06.2024].

Ramshaw, L., & Marcus, M. (1995). Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora.* URL: https://aclanthology.org/W95-0107 [zuletzt aufgerufen am 17.06.2024].

Schumacher, M. (2019). Named Entity Recognition mit dem Stanford Named Entity Recognizer, *forTEXT. Literatur digital erforschen*, 1–53. URL: https://fortext.net/routinen/lerneinheiten/named-entity-recognition-mit-dem-stanford-named-entity-recognizer [zuletzt aufgerufen am 17.06.2024].

Sporleder, C. (2010). Natural Language Processing for Cultural Heritage Domains, *Language and Linguistics Compass*, 4(9), 750–768. https://doi.org/10.1111/j.1749-818X.2010.00230.x [zuletzt aufgerufen am 17.06.2024].

Sprugnoli, R., Passarotti, M., Cecchini, F. M., & Pellegrini, M. (2020). Overview of the EvaLatin 2020 Evaluation Campaign. In *Proceedings of LT4HALA 2020. 1st Workshop on Language Technologies for Historical and Ancient Languages* (S. 105–110). Marseille: European Language Resources Association (ELRA). URL: https://aclanthology.org/2020.lt4hala-1.16 [zuletzt aufgerufen am 17.06.2024].

Sutinen, E., & Cooper, A.-P. (2021). *Digital Theology. A Computer Science Perspective.* Bingley: Emerald Publishing Limited. https://doi.org/10.1108/9781839825347 [zuletzt aufgerufen am 17.06.2024].

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task. Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, 4, 142–147. Edmonton: Association for Computational Linguistics. https://doi.org/10.3115/1119176.1119195 [zuletzt aufgerufen am 17.06.2024].

Yousef, T., Palladino, Ch., & Jänicke, S. (2023). Transformer-Based Named Entity Recognition for Ancient Greek. In W. Scholger, G. Vogeler, T. Tasovac, A. Baillot & P. Helling (Hrsg.), *Digital Humanities 2023. Collaboration as Opportunity (DH2023)* (S. 1–3). Graz: Zenodo. https://doi.org/10.5281/zenodo.8107629 [zuletzt aufgerufen am 17.06.2024].