Methodenbeitrag: Named Entity Recognition (NER)

Mareike Schumacher [1]

1. Universität Regensburg

forTEXT

Thema:	Named Entity Recognition	DOI:	10.48694/fortext.3765
Jahrgang:	1	Ausgabe:	9
Erscheinungsdatum:	30-10-2024	Erstveröffentlichung:	2018-05-17 auf fortext.net
Lizenz:	CC BY SA		open ∂ access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

1. Definition

Named Entity Recognition (NER) ist ein Verfahren, mit dem klar benennbare Elemente (z.B. Namen von Personen oder Orten) in einem Text automatisch markiert (vgl. Text Mining) werden können. Named Entity Recognition wurde im Rahmen der computerlinguistischen Methode des *Natural Language Processing* (NLP) entwickelt, bei der es darum geht, natürlichsprachliche Gesetzmäßigkeiten maschinenlesbar aufzubereiten.

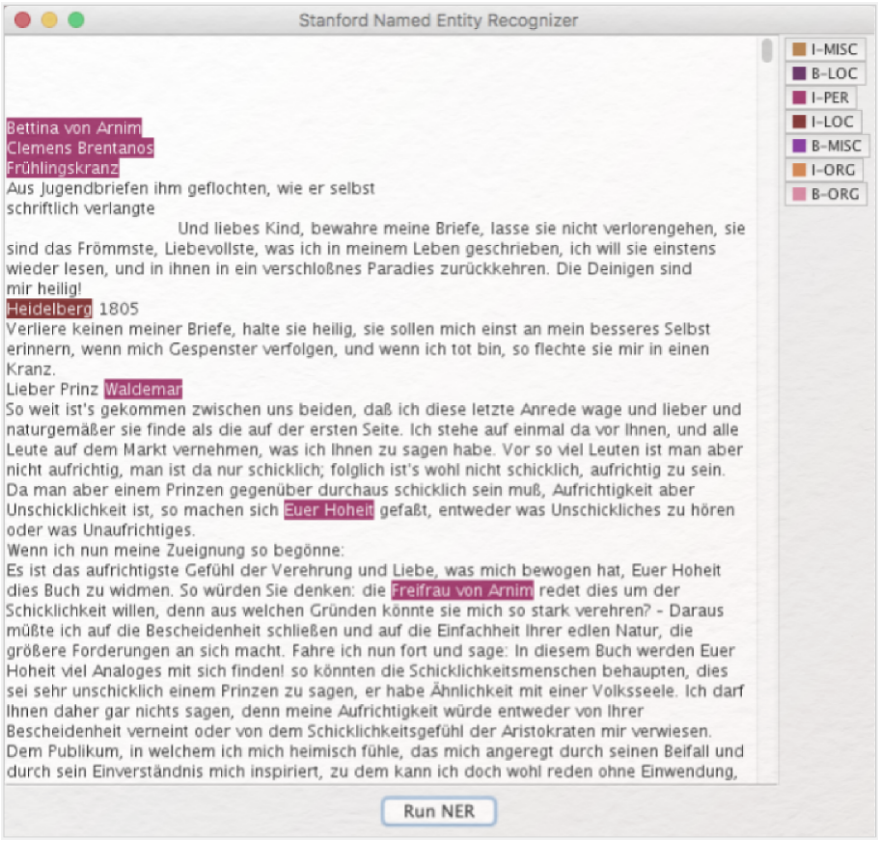

Abb. 1: Named Entity Recognition mit dem Stanford Named Entity Recognizer

2. Anwendungsbeispiel

Wollen Sie beispielsweise untersuchen, wie häufig Frauenfiguren in norddeutschen Kriminalromanen des 20. Jahrhunderts vorkommen, so können Sie Ihre Untersuchung mit *Named Entity Recognition* (vgl. Distant Reading) beginnen. In einer Textsammlung von 100 Kriminalromanen (vgl. Korpus) werden alle Figuren mit einem *NER*-Tool automatisch markiert (vgl. Annotation) und dann manuell nach weiblichen und männlichen Figuren differenziert. Es wird sichtbar, dass Frauenfiguren in Ihrem gesamten Textkorpus nur 15% aller Figurennennungen ausmachen.

Mit diesem Befund gehen Sie weiter und teilen Ihr Korpus in zeitliche Einheiten auf. Nun fällt Ihnen auf, dass Frauenfiguren zu Beginn des 20. Jahrhunderts nur 5% der Namensnennungen in norddeutschen Kriminalromanen ausmachen, während es zum Ende des 20. Jahrhunderts schon 40% sind. Sie spalten nun Ihr Korpus nach Autoren und Autorinnen auf und stellen fest, dass Autoren Frauenfiguren viel häufiger erwähnen als Autorinnen. Schließlich untersuchen Sie auch noch, ob es sich bei den benannten Figuren um Ermittler*innen, Verdächtige, Opfer, Hacker*innen usw. handelt.

3. Literaturwissenschaftliche Tradition

Mit *Named Entity Recognition* werden meistens drei Parameter gleichzeitig erfasst:

1. Personen,
2. Orte und
3. Organisationen.

In der Literaturwissenschaft wurden vor allem Personen (im Sinne von Figuren) und Orte bereits auf unterschiedliche Weise analysiert und sind bis heute wichtige Aspekte literaturwissenschaftlicher Forschung. Im Forschungsgebiet des „Organizational Storytelling", welches der Literaturwissenschaft und insbesondere der Erzähltheorie nahe steht, werden auch Organisationen in Narrativen untersucht. Im Folgenden werden die drei typischen Aspekte der *Named Entity Recognition* und ihre literaturwissenschaftliche Bedeutung einzeln betrachtet.

Personen: Schon seit Aristoteles werden Figuren als Elemente von Erzählungen verstanden und analysiert (Eder 2013, 40f.). Dabei kann der Fokus auf der Funktion einzelner Figuren (vgl. ebd., 41), dem Gesamtgefüge von Figurenkomplexen (vgl. ebd., 43) oder der psychologischen Ausgestaltung einzelner oder mehrerer Figuren liegen (vgl. ebd., 47ff.). Eine Besonderheit des literaturwissenschaftlichen Verständnisses von Figuren gegenüber einer linguistischen Definition von Person als *named entity* ist, dass neben der Nennung von Eigennamen auch allgemeines Weltwissen, Typisierungen und kulturelle Codes dazu beitragen, eine Figur auszugestalten (Jannidis 2012, 2). Für die *Named Entity Recognition* ist dabei von besonderer Bedeutung, dass neben Eigennamen auch andere Referenzen auf eine Figur verweisen können (vgl. ebd., 3). Trotzdem bleibt aber stets ein Eigenname zentral für die Erkennung von Figuren (Jannidis 2004, 109). Dies entspricht einem (post)strukturalistischen Verständnis von Figur (vgl. ebd., 104).
Auch in der literarischen Onomastik (Lamping 1983, 9), ist die Typologisierung und Bedeutungsanalyse von Namen in literarischen Texten vorherrschend (Stiegler 1994, 14). Hier wurden sowohl einzelne Werke (Krappmann 2012) als auch Gruppen literarischer Werke (Trauner 2012) und Korpora (Brütting 2013) auf Eigennamen untersucht. Darüber hinaus gab es bereits in den 1980er Jahren den Versuch, computergestützte Verfahren in dieses Forschungsfeld zu integrieren (Dalen-Oskam 2016, 345). Diese frühe Verknüpfung von Digital Humanities mit einem spezifischen literaturwissenschaftlichen Forschungsgebiet gibt einen Hinweis darauf, wie naheliegend eine solche Verbindung ist.

Raum: Die erzähltheoretische Forschung zur Thematik des Raumes folgt zwei Traditionslinien. Entweder wird der Raumbegriff primär im eigentlichen Wortsinne definiert oder hauptsächlich metaphorisch gebraucht (Ryan 2012, 2). Die buchstäbliche Verwendung des Raumbegriffs ähnelt dem linguistischen Verständnis von Orten, das der *Named Entity Recognition* zu Grunde liegt. Im Rahmen dieser Forschungstradition wurde Raum sowohl als Phänomen der Textoberfläche (Fludernik 2008, 51ff.) als auch der Tiefendimension betrachtet, welchem unterschiedliche Funktionen zuzuordnen sind. Dazu gehören die räumliche Rahmung von Ereignissen, die Ausgestaltung des Settings und die Gestaltung ganzer narrativer Welten (Ryan 2012, 6–10).
Für die Literaturgeografie oder -kartografie sind literarische Orte und Räume ebenfalls zentral (Piatti 2008). In diesem Forschungsfeld werden geografische Karten genutzt, um fiktive Räume zu visualisieren und zu analysieren. Orte und Räume werden hier als lokalisierbare Einheiten bzw. als räumliche Objekte (Reuschel, Piatti und Hurni 2012, 149) verstanden – eine Auffassung, die ebenfalls eine Ähnlichkeit mit dem linguistischen Ortsbegriff in der *Named Entity Recognition* aufweist.

Organisationen: Die dritte Kategorie, die in der *Named Entity Recognition* häufig mit berücksichtigt wird, ist weit weniger eindeutig als die beiden zuvor betrachteten. Sogar die Benennung variiert je nach Tool und manchmal sogar auch nach Modellen, die mit einem Tool verwendet werden. Es ist also denkbar, dass unterschiedliche literaturwissenschaftlich bedeutsame Motive zu dieser Kategorie gerechnet werden. Da z.B. Schulen und/oder Universitäten als Institutionen der Bildung als klar benennbare Einheiten interpretiert werden können, kann die Betrachtung dieser „Organisationen" analog zu Verfahren der *NER* gedacht werden (Röser 1975; Mix 1995; Mikota 2014; Pauldrach 2016).
Im erzähltheoretisch geprägten Forschungszweig zum „Corporate Storytelling" stehen Unternehmen und ihre narrative Außendarstellung sowie zugehörige Gegennarrative im Fokus des Interesses (Hansen, Narlyk und Wolff Lundholt 2013). In diesem Feld werden unter anderem handlungspraktische Methoden narrativer Kommunikation entwickelt (Thier 2004). Obwohl der Forschungsgegenstand hier fast deckungsgleich mit dem ist, was in der *Named Entity Recognition* erkannt wird, gibt es allerdings bisher nur wenige methodische Überschneidungen.

Indem literaturwissenschaftliche Konzepte so aufbereitet wurden, dass sie maschinenlesbar sind (Jannidis u. a. 2015), konnte die Methode bereits in Ansätzen auf literarische Texte angewandt werden (vgl. Domäneadaption). Da aber noch nicht alle Parameter in einem solchen Verfahren für die Literaturwissenschaft aufbereitet wurden, kann an dieser Stelle noch Grundlagenforschung betrieben werden.

4. Diskussion

NER-Tools erreichen bei der automatischen Extraktion von Informationseinheiten (*named entities*) aus Sachtexten eine hohe Verlässlichkeit, weshalb *NER* in der Linguistik bereits als gelöstes Problem betrachtet wurde (Cunningham 2005, 668). Die erzielten Erfolgsquoten bei der Erkennung von Personen, Orten und Organisationen variieren in unterschiedlichen Sprachen allerdings stark. Vorhandene *NER*-Tools wurden zumeist mit großen Korpora bestehend aus journalistischen Texten trainiert (das Training von *NER* Tools wird in Abschnitt 5 genauer beschrieben). Es gibt mehrere sprachenspezifische *NER*-Modelle, die mit diesen Tools verwendet werden können. Diese Modelle haben durchschnittliche Erfolgsquoten (*F-Scores*) von 68% in deutschen Texten und 89% in englischen Texten (Faruqui und Padó 2010, 1). Wie Jannidis u. a. (2015) am Beispiel von Figuren zeigten, erkennt das Modell für die deutsche Sprache von Faruqui und Padó in literarischen Texten nur knapp 20% aller Vorkommnisse von Figuren im Text („Recall-Quote" genannt) im Vergleich zu 88% erkannten Personen in journalistischen Texten. Ein Grund hierfür kann darin gesehen werden, dass in literarischen Texten Figuren seltener mit ihren Eigennamen und häufiger mit indirekten Verweisen oder Umschreibungen aufgerufen werden. Eine Nutzung von *NER*-Tools ohne vorhergehende Domänenadaption ist also problematisch.
Der Einsatz von *NER* bedeutet eine enorme Zeitersparnis gegenüber der manuellen Annotation von *named entities* in Texten. Damit ermöglicht *NER* vor allem die Betrachtung großer Textmengen, d. h. ganzer Romane und Korpora. *NER* sollte aber stets in dem Bewusstsein der eingeschränkten Zuverlässigkeit der Methode eingesetzt werden. Sogar wenn durch die Domäneadaption für Figuren in (deutschen) literarischen Texten eine Trefferquote von rund 85% erreicht werden konnte (Jannidis u. a. 2015), bedeutet dies, dass rund 15% der relevanten *named entities* nicht gefunden wurden. Vor allem bei Analysen von Einzeltexten kann diese Fehlerquote durchaus so gewichtig sein, dass die qualitative Interpretation, die anhand des Datenmaterials geleistet werden soll, fehlgeleitet wird.
Der Einsatz von *NER* in der Literaturwissenschaft eignet sich besonders für *Distant Reading*-Ansätze. Während im *Close Reading*-Verfahren zumeist einzelne Texte oder eine kleine Anzahl von Texten exemplarisch analysiert werden, werden beim *Distant Reading* größere Korpora in den Fokus gerückt. Moretti beschreibt, wie die Literaturwissenschaft mittels *Close Reading* bisher vielleicht 1% der Weltliteratur tatsächlich in Betracht gezogen hat. Die restlichen 99%, die er als „the great unread" bezeichnet, kommen in der Literaturwissenschaft bisher eher selten vor (Moretti 2013, 63–70). Digitale Methoden machen es möglich, sich diesem „great unread" zuzuwenden. Statt wie beim *Close Reading* herausragende Einzeltexte zu studieren, wird Literatur beim *Distant Reading* eher relational betrachtet. Auf einer solchen Basis wird auch der Einsatz digitaler *NER*-Tools bedeutsam. Denn mit der zunehmenden Distanz von einzelnen Texten und dem Fokus auf Relationen zwischen bestimmten quantitativen Aspekten in unterschiedlichen Texten relativiert sich auch die oben hervorgehobene Fehlerquote. In solchen Analysen ist es nicht mehr entscheidend, an welcher Stelle welche Figur in welchem Kontext genannt wird. Stattdessen werden eher Muster fokussiert, die in Gruppen von Texten oder sogar in einem großen Korpus auftreten.
Abschließend bleibt noch festzuhalten, dass auch in *Distant-Reading*-Verfahren ein bewusster Umgang mit der Fehleranfälligkeit von *NER* gepflegt wird. Statt davon auszugehen, dass sich Fehlerquoten durch die Vergrößerung von Korpora nivellieren, werden diese eher in Kauf genommen, um überhaupt Analysen großer Textmengen durchführen zu können. Es werden hier erste Annäherungen an das „great unread" gewagt, um dadurch haltbare Aussagen über Nationalliteraturen, Weltliteratur oder Literatur als solche treffen zu können.

5. Technische Grundlagen

Im ersten Schritt der *Named Entity Recognition* geht es zunächst darum, dem Computer zu vermitteln, wie die Worte erkannt werden können, die als *named entity* gekennzeichnet werden sollen. Dazu werden eine Reihe von Merkmalen definiert, die vom Tool statistisch ausgewertet werden (technisch: Festlegen der *Natural Language Processing (NLP)*-Feature) und so eine möglichst präzise Erkennung möglich machen sollen. Zum Beispiel können Wortlisten berücksichtigt werden, die alle Figurennamen, Orte und Organisationen verzeichnen, die vorkommen könnten. Zusätzlich können Wortarten nicht nur der *named entity* selbst sondern auch der vorangehenden und nachgestellten Worte einbezogen werden. Weitere *features* können sein:

- häufig vorher genannte Wörter (sowie z.B. bei Orten das Wort „in"),
- Darstellungsformat (bei Daten so etwas wie Zahl Monat Jahr),
- Groß- und Kleinschreibung,
- Position im Satz
- usw.

Mit Hilfe dieser im Tool vordefinierten *features* kann ein als *Machine Learning* bezeichneter Prozess durchgeführt werden. Der Lernprozess des *NER*-Tools (auch „Training" genannt) besteht darin, dass diese *features* mit einem manuell annotierten Text abgeglichen werden; dem sogenannten Trainingskorpus. Ergebnis dieses Abgleichs ist das *NER*-Modell. Die Anzahl der *features*, die ein Tool berücksichtigt, kann variieren. Nur die Kombination verschiedener *features* führt zu guten Ergebnissen bei der automatischen Erkennung, da *named entities* in unterschiedlichen Zusammenhängen unterschiedliche Bedeutungen tragen können. So wäre z.B. „Paris" mit vorangestelltem „in" wahrscheinlich eine Referenz auf die Stadt, mit nachgestelltem Verb wahrscheinlich eine Person. Ausnahmen können durch den Abgleich mit dem manuell annotierten Text ebenfalls mit einbezogen werden. Wenn im Trainingskorpus z.B. in Sätzen wie „Paris hat schöne Museen" Paris als Ort ausgezeichnet wurde, obwohl ein Verb dahinter steht, so kann das Tool erkennen, dass das Wort „hat" direkt hinter einer *named entity* darauf hindeuten kann, dass es sich um einen Ort handelt. Letztendlich errechnet das Tool anhand der *feature*-Kombinationen und der Trainingsdaten, welche Bedeutung in welchem Zusammenhang wahrscheinlicher ist.
NER kann z.B. mit Hilfe des Stanford Named Entity Recognizers (Schumacher 2024b) oder der virtuellen Arbeitsumgebung WebLicht (Schumacher 2024a) über eine grafische Nutzeroberfläche (vgl. GUI) durchgeführt werden, die auch mit geringen technischen Kenntnissen bedienbar sind. In diesen Tools werden *NER*-Modelle genutzt, die zumeist mittels Korpora aus Sachtexten trainiert wurden. Um das Tool mit Beispieldaten einer anderen Domäne wie z.B. der Literatur neu zu trainieren oder die vordefinierten *features* zu ergänzen, ist es nötig, über die Commandline des Computers eine Reihe von Befehlen in der Programmiersprache des jeweiligen Tools einzugeben. Die erforderlichen Befehle finden sich allerdings oft in der Dokumentation der Tools, sind dort erklärt und können per copy/paste eingefügt werden. Profunde Kenntnisse einer oder mehrerer Programmiersprache(n) sind also auch hierfür nicht nötig. Es ist aber durchaus hilfreich, die Sprache des Codes zumindest lesen und verstehen zu können.

Externe und weiterführende Links

- Stanford CoreNLP: https://web.archive.org/save/https://stanfordnlp.github.io/CoreNLP/ (Letzter Zugriff: 10.10.2024)
- WebLicht: https://web.archive.org/save/https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page (Letzter Zugriff: 10.10.2024)

Bibliographie

Brütting, Richard. 2013. *Namen und ihre Geheimnisse in Erzählwerken der Moderne*. Hamburg: Baar.

Chlopczyk, Jaques. 2017. *Beyond Storytelling. Narrative Ansätze und die Arbeit mit Geschichten in Organisationen*. Berlin: Springer Gabler.

Cunningham, H. 2005. Information extraction, automatic. In: *Encyclopedia of Language and Linguistics*, hg. von Keith Brown, 665–677. Oxford: Elsevier.

Dalen-Oskam, Karina. 2016. Corpus-based Approaches to Names in Literature. In: *The Oxford Handbook of Names and Naming*, hg. von Carole Hoigh. Oxford: Oxford University Press.

Eder, Jens. 2013. *Die Figur im Film. Grundlagen der Figurenanalyse*. Marburg: Schüren.

Faruqui, Manaal und Sebastian Padó. 2010. Training and Evaluating a German Named Entity Recognizer with Semantic Generalization. In: *Proceedings of Konvens 2010*. Saarbrücken. https://nlpado.de/~sebastian/pub/papers/konvens10_faruqui.pdf (zugegriffen: 4. Mai 2018).

Fludernik, Monika. 2008. *Erzähltheorie. Eine Einführung*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Hansen, Per Krogh, Brigitte Narlyk und Marianne Wolff Lundholt. 2013. Corporate Storytelling. In: *the living handbook of narratology*, hg. von Peter Hühn, Jan Christoph Meister, John Pier, und Wolf Schmid. Hamburg: Hamburg University. http://www.lhn.uni-hamburg.de/article/corporate-storytelling (zugegriffen: 8. Mai 2018).

Jannidis, Fotis. 2004. *Figur und Person. Beitrag zu einer historischen Narratologie*. Berlin: de Gruyter.

———. 2012. Character. In: *the living handbook of narratology*, hg. von Peter Hühn, Jan Christoph Meister, John Pier, und Wolf Schmid. Hamburg: Hamburg University. http://www.lhn.uni-hamburg.de/article/character (zugegriffen: 8. Mai 2018).

Jannidis, Fotis, Markus Krug, Frank Puppe, Isabella Reger, Martin Toepfer und Lukas Weimer. 2015. Automatische Erkennung von Figuren in deutschsprachigen Romanen. In: *DHd 2015. Von Daten zu Erkenntnissen. Book of Abstracts*. http://gams.uni-graz.at/o:dhd2015.abstracts-vortraege (zugegriffen: 4. Mai 2018).

Krappmann, Tamara. 2012. *Die Namen in Uwe Johnsons „Jahrestagen"*. Göttingen: V & R Unipress.

Lamping, Dieter. 1983. *Der Name in der Erzählung. Zur Poetik des Personennamens*. Bonn: Bouvier.

Mikota, Jana. 2014. Lehrer als Täter - Schüler als Opfer, oder doch umgekehrt? Schule in der Gegenwartsliteratur. *Der Deutschunterricht* 1: 70–78.

Mix, York-Gothart. 1995. *Die Schulen der Nation. Bildungskritik in der Literatur der Moderne*. Stuttgart: Metzler.

Moretti, Franco. 2013. *Distant Reading*. London, New York: Verso.

Pauldrach, Matthias. 2016. „Nicht für das Leben, sondern für die Schule lernen wir!“ Eine Reflexion der Beziehung von Schule und Leben anhand von Wes Andersons Spielfilm Rushmore. *Schule in Literatur und Film. Zeitschrift für den Deutschunterricht in Wissenschaft und Schule* 1: 20–29.

Piatti, Barbara. 2008. *Die Geografie der Literatur. Schauplätze, Handlungsräume, Raumphantasien*. Göttingen: Wallstein.

Reuschel, Ann Kathrin, Barbara Piatti und Lorenz Hurni. 2012. Modelling Uncertain Geodata for the Literary Atlas of Europe. In: *Understanding Different Geographies*, hg. von Karel Kritz, William Cartwright, und Michaela Kinberger, 135–157. Berlin, Heidelberg: Springer.

Röser, Dietmar. 1975. *Das Bild der höheren Schule in der neueren deutschen Literatur*. Köln: Dissertationsdruck Hansen.

Ryan, Marie-Laure. 2012. Space. In: *the living handbook of narratology*, hg. von Peter Hühn, Jan Christoph Meister, John Pier, und Wolf Schmid. Hamburg: Hamburg University. http://www.lhn.uni-hamburg.de/node/55.html (zugegriffen: 8. Mai 2018).

Schumacher, Mareike. 2024a. Toolbeitrag: Abbyy FineReader. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 3. Textdigitalisierung und Edition (12. Juni). doi: 10.48694/fortext.3742, https://fortext.net/tools/tools/abbyy-finereader.

———. 2024b. Toolbeitrag: Stanford Named Entity Recognizer. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3767, https://fortext.net/tools/tools/stanford-named-entity-recognizer.

Stiegler, Bernd. 1994. *Die Aufgabe des Namens. Untersuchung zur Funktion der Eigennamen in der Literatur des zwanzigsten Jahrhunderts*. München: Fink.

Thier, Karin. 2004. *Die Entdeckung des Narrativen für Organisationen. Entwicklung einer effizienten Story-Telling-Methode*. Hamburg: Kovač.

Trauner, Karl. 2012. *Die Namenwelt in den Kinder- und Hausmärchen der Brüder Grimm*. Szentendre: Tillinger.

Viehhauser-Mery, Gabriel und Florian Barth. 2017. Towards a Digital Narratology of Space. In: *Digital Humanities 2017. Conference Abstracts*. Montréal, Canada. https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf (zugegriffen: 3. Mai 2018).

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

Domäneadaption Domäneadaption beschreibt die Anpassung einer in einem Fachgebiet entwickelten digitalen Methode an ein anderes Fachgebiet.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z. B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

NLP *Natural Language Processing* (NLP), maschinelle Sprachverarbeitung zu Deutsch, ist ein Teilgebiet der Linguistik, der Informatik und der künstlichen Intelligenz, welches sich damit beschäftigt, wie Computer so programmiert werden, dass sie große Mengen an natürlichsprachlichen Daten verarbeiten und analysieren können.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

POS PoS steht für *Part of Speech*, oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von

Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

Wordcloud Eine *Wordcloud*, oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.