Ressourcenbeitrag: DWDS: Digitales Wörterbuch der Deutschen Sprache

Jan Horstmann [1]

1. Universität Münster

Thema:	Projektkonzeption	DOI:	10.48694/fortext.3804
Jahrgang:	1	Ausgabe:	12
Erscheinungsdatum:	30-11-2024	Erstveröffentlichung:	2019-07-01 auf fortext.net
Lizenz:			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. Kurzbeschreibung

Das DWDS ist ein digitales Lexikon, das Ihnen die freie Suche nach Begriffen der deutschen Sprache und ihrer historischen und gegenwärtigen Bedeutung ermöglicht. Sie können bestimmen, in welchen der großen Textsammlungen (z. B. DWDS-Kernkorpora (vgl. Korpus) des 19., 20. oder 21. Jahrhunderts, Zeitungs-, Blog-, Webkorpora etc.) und welcher Textsorte (Belletristik, Wissenschaft, Gebrauchsliteratur oder Zeitungen) gesucht werden soll.

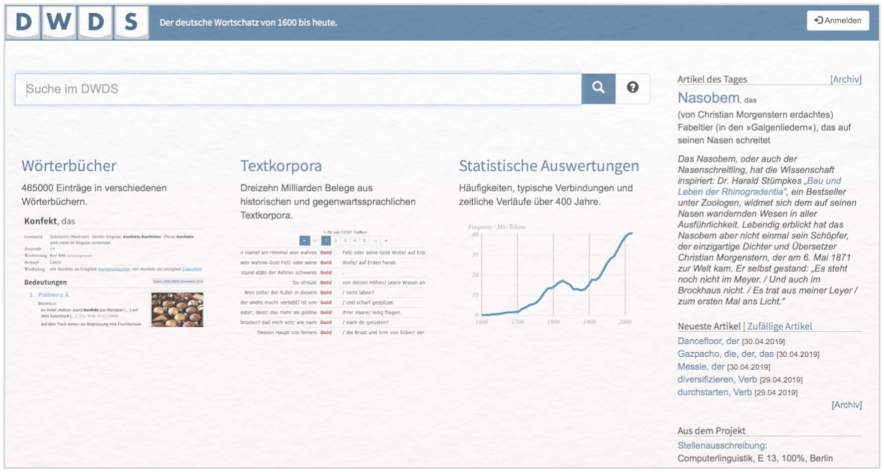

Abb. 1.: Startseite des DWDS

Steckbrief

- https://www.dwds.de
- Projekt der Berlin-Brandenburgischen Akademie der Wissenschaften zur Erstellung eines digitalen allgemein zugänglichen Wörterbuchsystems mit derzeit 13.521.774.869 Tokens (vgl. Type/Token)
- Referenzkorpora zum 19. (Deutsches Textarchiv (DTA) siehe Horstmann und Kern (2024)), 20. und 21. Jahrhundert: Das Kernkorpus zum 20. Jahrhundert (über 121 Millionen Tokens) ist über das gesamte Jahrhundert gestreut und nach Textsorten ausgewogen: Belletristik (28,42 %), Zeitung (27,36 %), wissenschaftliche Fachtexte (23,15 %) und Gebrauchstexte (21,05 %); das Kernkorpus zum 21. Jahrhundert (derzeit gut 15 Millionen Tokens) wird laufend erweitert, ist noch nicht ausgewogen, jedoch ebenfalls zeitlich und nach Textsorten differenziert
- verknüpfte lexikalische Informationstypen: Artikel des Wörterbuchs der deutschen Gegenwartssprache (WDG) inkl. automatisch generierter Informationen zu Synonymen, Hyponymen und Hyperonymen, Textbeispiele aus den DWDS-Kernkorpora und statistische Kookkurrenz-Informationen
- mehr als 10.000 registrierte Benutzer*innen (einige Korpora benötigen zur Recherche eine kostenfreie Registrierung)
- Wörterbücher: Wörterbuch der deutschen Gegenwartssprache (WDG), DWDS-Wörterbuch, Etymologisches Wörterbuch des Deutschen, Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm (DWB), Das Große Wörterbuch der deutschen Sprache in 10 Bänden (Duden 1999), OpenThesaurus

- Zeitungskorpora: BILD (1996-2018), Berliner Zeitung (1945-2005), Frankfurter Rundschau (1997-2000), neues deutschland (1946-1990), NZZ (1970-2018), SPIEGEL (1947-2014), Der Standard (2000-2016), Süddeutsche Zeitung (1992-2017), Tagesspiegel (1996-2005), taz (1986-1999), Welt (1997-2018), Die ZEIT (1946-2018)
- Spezialkorpora: Blogs, Webkorpus (Auswahl von Webseiten auf Deutsch), Dortmunder Chat-Korpus, Filmuntertitel, Polytechnisches Journal, DDR (1100 Texte von 1949-1990), Gesprochene Sprache (Transkripte von Reden, Parlamentsprotokollen, Interviws des 20. Jhs.), Text+Berg (Jahrbuch Schweizer-Alpenclub), Berliner Wendekorpus (77 Interviews mit Ost- und Westberliner*innen)

2. Anwendungsbeispiel

Sie vergleichen drei literarische Werke aus dem 19., 20. und 21. Jahrhundert in gendertheoretischer Perspektive und begegnen dabei unterschiedlichen Verwendungen des Begriffs „Geschlecht“. Eine Recherche im DWDS bietet Ihnen die diversen Bedeutungen des Begriffes, seine Etymologie, Verknüpfungen mit einem Thesaurus, ein Wortprofil mit einer interaktiven Wordcloud, automatisch generierte Beispiele aus den DWDS-Korpora wie „Aber jeder von uns besitzt alle nötigen Gene für beide Geschlechter“ aus der *Süddeutschen Zeitung* am 07. November 2003 (https://www.dwds.de/wb/Geschlecht, Zugriff: 07. Mai 2019), Angaben über die Worthäufigkeit, eine Wortverlaufskurve (die ihren Höhepunkt um 1800 hat), Zugriffsmöglichkeiten auf die älteren Wörterbücher DWB und WDG sowie Angaben über Trefferquoten in den einzelnen Korpora des DWDS (sodass Sie bei literaturwissenschaftlichem Interesse auch noch in die Referenzkorpora zu den einzelnen Jahrhunderten schauen können).

3. Diskussion

3.1 Kann ich das Digitale Wörterbuch der Deutschen Sprache für wissenschaftliche Arbeiten nutzen?

Ja. Das DWDS ist bibliographisch referenzierbar und bei der Textauswahl und Aufbereitung wurde und wird auf inhaltliche und qualitative Streuung geachtet, sodass der deutsche Wortschatz von 1600 bis in die Gegenwart repräsentativ dargestellt wird. Zur Recherche von Volltexten bietet sich das DWDS jedoch nicht an. Stattdessen ermöglicht es dezidiert, Wörter in ihren Gebrauchskontexten zu erforschen. Volltexte finden Sie für das Referenzkorpus des 19. Jahrhunderts auf der Webseite des Deutschen Textarchivs (DTA). Für das 20. und 21. Jahrhundert können Volltexte aufgrund des Urheberrechts i. d. R. noch nicht angeboten werden, das DWDS stellt in dieser Hinsicht keine Ausnahme dar. Die Metadaten der hinterlegten Dokumente sind auf sehr hohem Niveau (die Redaktion achtet auf Vollständigkeit und Einheitlichkeit) und die mit dem DWDS gefundenen Belege können unter Beachtung der Nutzungsbedingungen frei weiterverwendet werden. Zudem bietet das DWDS eine Zitationshilfe an.

3.2 Wie benutzerfreundlich ist die Arbeit mit dem Digitalen Wörterbuch der Deutschen Sprache?

Das DWDS kann in den meisten Bereichen intuitiv bedient werden und die Webseite ist übersichtlich gestaltet (vgl. GUI). Etliche Korpora können ohne vorherige Registrierung kostenfrei durchsucht werden und vor allem die Visualisierung (vgl. Text Mining) von bis zu vier Begriffen als Verlaufskurven (siehe Abb. 2) stellt ein hilfreiches Tool zur Herstellung von Übersichten dar.

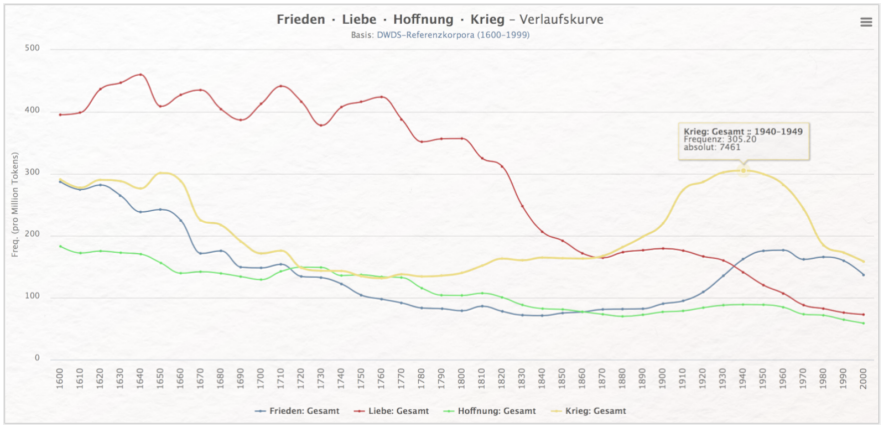

Abb. 2: Verlaufskurven der Begriffe „Frieden“, „Liebe“, „Hoffnung“ und „Krieg“ in den DWDS-Referenzkorpora

Diese in unterschiedlichen Formaten exportierbaren Verlaufskurven bieten nicht nur einen synoptischen Überblick, sondern können auch interaktiv exploriert werden. Die Diagramme lassen sich zudem mit einem Klick dahingehend ausdifferenzieren, dass die einzelnen Textsorten zu den ausgewählten Begriffen getrennt voneinander visualisiert werden (siehe Abb. 3).

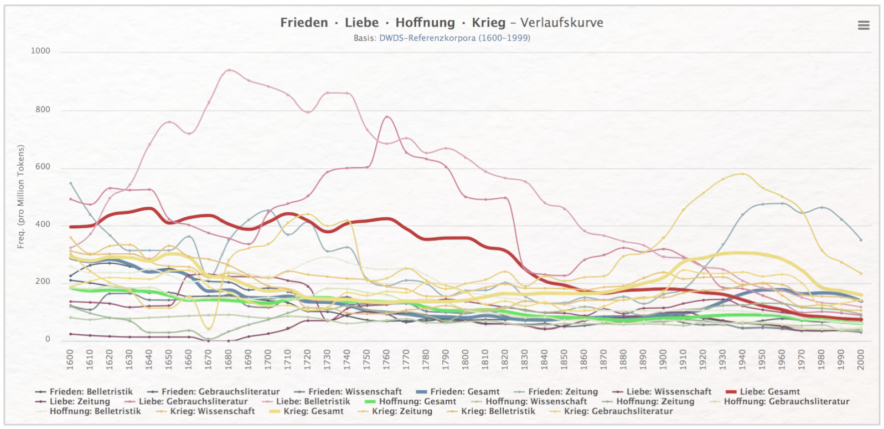

Abb. 3: Verlaufskurven der Begriffe „Frieden“, „Liebe“, „Hoffnung“ und „Krieg“ in den DWDS-Referenzkorpora, differenziert nach Textsorten

Insbesondere der große Funktionsumfang der Suchoptionen (vgl. Query) bedarf jedoch einer genaueren Einarbeitung; eine kompakt gestaltete Überblicksseite ermöglicht Ihnen hierbei den Einstieg in die Grammatik der Suchabfragen. Einige Korpora des DWDS können aufgrund von Nutzungsvereinbarungen mit den Rechtegebern lediglich mit vorheriger Registrierung – dann jedoch ebenfalls kostenfrei – genutzt werden.
Das mit beinahe allen Korpora im DWDS verknüpfte, von der Forschungsinfrastruktur CLARIN-D entwickelte Analysetool DiaCollo zur diachronen Kollokationsanalyse (vgl. Kollokation) ermittelt typische Wortverbindungen nach deren zeitlichem Auftreten. Der ausgewählte Begriff wird auf Grundlage des jeweils bestimmten Korpus zusammen mit anderen in seinem Umfeld häufig vorkommenden Begriffen bspw. als animierte Wordcloud oder animierte Bubble-Visualisierung dargestellt (siehe Abb. 4). Wie das etwa mit dem Begriff „Liebe“ im Verlauf des 20. Jahrhunderts aussieht, können Sie hier verfolgen.

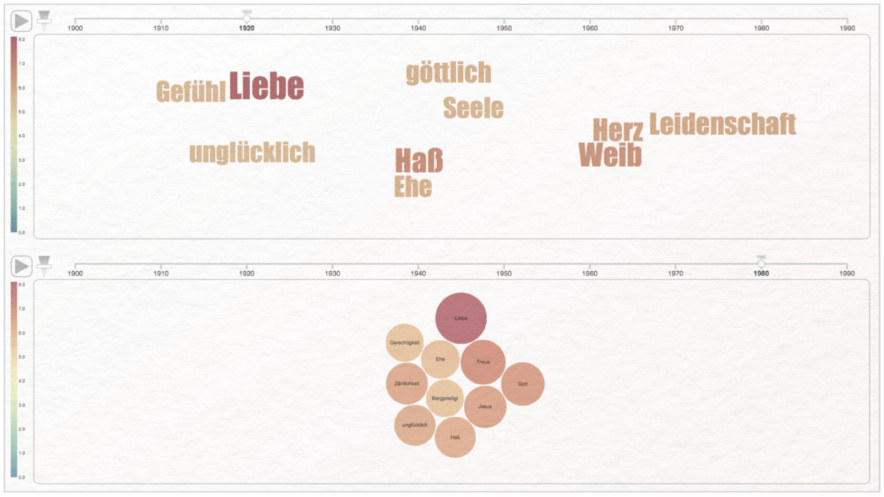

Abb. 4: Kollokationsanalyse des Begriffs „Liebe“ im DWDS-Kernkorpus des 20. Jahrhunderts als Wordcloud und Bubble-Visualisierung

4. Wie funktioniert die Textsuche im Digitalen Wörterbuch der Deutschen Sprache?

Die Begriffssuche im DWDS funktioniert denkbar einfach: Bereits auf der Startseite haben Sie ein großes Sucheingabefeld, in das Sie Ihren Begriff eintippen können. Bereits während Sie tippen, werden Ihnen aus den Korpora des DWDS automatisch Vervollständigungen angeboten, wie Sie das auch von der Arbeit mit größeren onlinebasierten Suchmaschinen kennen. Sie können Ihren Begriff nun entweder vollständig eingeben (Groß- oder Kleinschreibung spielt hierbei keine Rolle) und dann die Suche starten (per Klick auf das Lupensymbol oder die Enter-Taste), oder Sie wählen einen der vorgeschlagenen Begriffe per Mausklick aus. Anschließend gelangen Sie zur Übersichtsseite des jeweiligen Begriffes mit allen oben im Abschnitt Anwendungsbereich beschriebenen Kategorien.
Auf dieser Ergebnisseite sehen Sie in der rechten Spalte außerdem die sog. Korpustreffer. Klicken Sie hier auf das von Ihnen präferierte Korpus, gelangen Sie zu den einzelnen Vorkommnissen des gesuchten Begriffes im ausgewählten Korpus. Dort finden Sie außerdem eine differenzierte Suchmaske, um einzelne Korpora, Textsorten und genauer definierte Zeitabschnitte zu durchsuchen. Die Korpussuche bietet Ihnen zudem die Möglichkeit, Suchergebnisse in unterschiedlichen Ansichten darzustellen.
Wie im Deutschen Textarchiv (DTA) ist es im gesamten DWDS möglich, die Suchabfragesprache der korpuslinguistischen Suchmaschine DDC zu verwenden, mithilfe derer komplexe Suchanfragen bspw. nach Wortgruppen, Phrasen, Lemmata, Satzanfängen etc. vorgenommen werden können.

Externe und weiterführende Links

- DiaCollo Manual: https://web.archive.org/web/20240927130528/https://clarin-d.de/de/kollokationsanalyse-in-diachroner-perspektive (Letzter Zugriff: 06.11.24)
- DTA (Deutsches Textarchiv): https://web.archive.org/web/20241106155219/https://www.deutschestextarchiv.de/ (Letzter Zugriff: 06.11.24)
- DTA (Deutsches Textarchiv) DDC Suchmaschine: https://web.archive.org/web/20241106155432/https://deutschestextarchiv.de/doku/software#ddc (Letzter Zugriff: 06.11.24)
- DWDS Homepage: https://web.archive.org/web/20241106155912/https://www.dwds.de/ (Letzter Zugriff: 06.11.24)
- DWDS Korpussuche: https://web.archive.org/web/20241106160030/https://www.dwds.de/d/suche#korpussuche (Letzter Zugriff: 06.11.24)
- DWDS Nutzungsbedingungen: https://web.archive.org/web/20240927130219/https://www.dwds.de/d/nutzungsbedingungen (Letzter Zugriff: 06.11.24)
- DWDS Recherche: https://web.archive.org/web/20241106160227/https://www.dwds.de/wb/Geschlecht (Letzter Zugriff: 06.11.24)
- DWDS Überblicksseite: https://web.archive.org/web/20241106160030/https://www.dwds.de/d/suche (Letzter Zugriff: 06.11.24)
- DWDS Zitationshilfe: https://web.archive.org/web/20240926033226/https://www.dwds.de/d/zitieren (Letzter Zugriff: 06.11.24)
- Wordcloud zum Begriff „Liebe“ im Verlauf des 20. Jahrhunderts: https://web.archive.org/web/20241106160605/https://ddc.dwds.de/dstar/kern/diacollo/?query=Liebe&format=cloud (Letzter Zugriff: 06.11.24)

Bibliographie

Barbaresi, Adrien. 2016. Efficient construction of metadata-enhanced web corpora. In: *Proceedings of the 10th Web as Corpus Workshop*, 7–16. Berlin: Association for Computational Linguistics. https://www.aclweb.org/anthology/W16-2602.

Barbaresi, Adrien und Kay-Michael Würzner. 2014. For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In: *Proceedings of NLP4CMC workshop (KONVENS 2014)*, 2–10. Hildesheim University Press.

Geyken, Alexander. 2014. Methoden bei der Wörterbuchplanung in Zeiten der Internetlexikographie. *Lexicographica* 30, Nr. 1: 77–111.

Herold, Axel. 2011. Retrodigitalisierung und Modellierung des Wörterbuchs der deutschen Gegenwartssprache. In: *Sprachliche Förderung und Weiterbildung – transdisziplinär*, hg. von Andreas Kraft und Carmen Spiegel. Frankfurt am Main: Peter Lang.

Horstmann, Jan und Alexandra Kern. 2024. Ressourcenbeitrag: Deutsches Textarchiv (DTA). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3791, https://fortext.net/ressourcen/textsammlungen/deutsches-textarchiv-dta.

Klappenbach, Ruth und Helene Malige-Klappenbach. 1980. Das Wörterbuch der deutschen Gegenwartssprache. Entstehung, Werdegang, Vollendung. In: *Studien zur modernen deutschen Lexikographie. Auswahl aus den lexikographischen Arbeiten. Erweitert um drei Beiträge von Helene Malige-Klappenbach*, hg. von Werner Abraham und Jan F. Brand, 3–58. Amsterdam: Benjamins.

Klein, Wolfgang und Alexander Geyken. 2010. Das Digitale Wörterbuch der Deutschen Sprache DWDS. *Lexicographica* 26: 79–96.

Schmidt, Thomas, Alexander Geyken und Angelika Storrer. 2008. Refining and Exploiting the Structural Markup of the eWDG. In: *Proceedings of the XIII EURALEX International Congress*, 469–481. Barcelona, Spain.

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.

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.

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.

Kollokation Als Kollokation bezeichnet man das häufige, gemeinsame Auftreten von Wörtern oder Wortpaaren in einem vordefinierten Textabschnitt.

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.

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.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

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.