Texte und linguistische Ebenen

Aus Sicht der Informatik ist ein Text zunächst nur eine Menge von Zeichenketten, also eine Datenmenge vom Typ string. Beim Text Mining müssen jedoch auch die linguistischen Strukturen eines Textes mit ausgewertet werden.

Linguistische Strukturen finden sich auf allen Ebenen eines Textes. In Anlehnung an Noam Chomsky wollen wir diese Ebenen als **linguistische Ebenen** bezeichnen. Damit werden (unabhängig von dem verwendeten linguistischen Theorierahmen) die Elemente einer hierarchischen Untergliederung von Texten bezeichnet, die mit den Elementen eines Alphabets (den Buchstaben) beginnt und mit der Ebene des Satzes endet (vgl. Chomsky [4]). Die wesentlichen Ebenen sind:

- Buchstaben eines Alphabets
- Morpheme als sinntragende Kombinationen von Buchstaben eines Alphabets
- Wortformen als sinntragende und morphologisch zulässige Kombination von Morphemen
- Phrasen als sinntragende und syntaktisch zulässige Kombination von Wortformen
- Sätze als sinntragende und syntaktisch zulässige Kombination von Phrasen.

Beispiel - Linguistische Ebenen bei der Zerlegung eines Satzes

Der Satz „Das Gewandhaus zu Leipzig befindet sich am Augustusplatz“ besteht aus:
1) einer Konkatenation von Buchstaben:
D-a-s-G-e-w-a-n-d-h-a-u-s-z-u-L-e-i-p-z-i-g-b-e-f-i-n-d-e-t-s-i-c-h-a-m-A-u-g-u-s-t-u-s-p-l-a-t-z
2) einer Konkatenation von Morphemen:
Das-Gewand-haus-zu-Leipzig-be-find-et-sich-am-Augustus-platz
3) einer Konkatenation von Wortformen:
Das-Gewandhaus-zu-Leipzig-befindet-sich-am-Augustusplatz
4) einer Konkatenation von Phrasen:
Das Gewandhaus-zu Leipzig-befindet sich-am Augustusplatz

Abhängig von den zu verarbeitenden Texten führt die Konkatenation von Phrasen nicht immer zur Ebene der Sätze. So finden sich z.B. in sozialen Medien im Unterschied zu Zeitungstexten nicht immer vollständige Sätze, sondern als vergleichbare Einheit sog. *Konversationen*. Für die Zwecke des Text Mining sollten die zu verarbeitenden linguistischen Ebenen daher je nach Aufgabe angepasst werden (vgl. Gründer et. al. [8]).

Linguistische Ebenen lassen sich durch die Kombination von einfachen Elementen (wie z. B. Wortformen) als komplexere Elemente (Wörter oder Sätze) darstellen. Hierbei werden die folgenden Bildungsprinzipien verwendet:

- Konkatenation (Aneinanderfügung) und
- Abstraktion (Bildung von Äquivalenzklassen)

Einzelne Ebenen sind auch Gegenstand von Teilgebieten der Linguistik: Gegenstand der Morphologie sind Morpheme, Gegenstand der Lexikologie sind Wörter, Gegenstand der Syntax sind Phrasen und deren Kombination zu Sätzen (vgl. Chomsky,[5]).

Die Berücksichtigung der linguistischen Ebenen ist für die automatische Ermittlung von inhaltlichen Zusammenhängen aus Texten eine wesentliche Voraussetzung. Hierbei wird linguistisches Wissen verwendet, welches die für eine Sprache typischen Gesetzmäßigkeiten auf

allen ihren linguistischen Ebenen zusammenfasst. Im nachfolgenden Abschnitt stellen wir zunächst einige Besonderheiten natürlicher Sprachen vor, welche für das Text Mining eine besondere Herausforderung darstellen und deren Verarbeitung linguistisches Wissen erfordert.

Warum erfordert die Verarbeitung natürlicher Sprache linguistisches Wissen?

Für die automatische Verarbeitung von Daten und Informationen werden in der Informatik künstliche, formale Sprachen, beispielsweise eine Programmiersprache oder eine Auszeichnungssprache wie XML, verwendet. Dabei wird für eine formale Sprache durch Wohlgeformtheitsregeln festgelegt, welche Zeichenketten in dieser Sprache zulässig sind. Nicht jede wohlgeformte Zeichenkette hat aber auch eine Bedeutung. Deshalb wird für jede formale Sprache mit Hilfe von Gültigkeitsregeln zusätzlich festgelegt, welche Zeichenketten gültig sind und welche Bedeutung eine gültige Zeichenkette in der intendierten Anwendung hat (Kleene,[15]). Dadurch wird sichergestellt, dass klar definiert ist, welches die zulässigen Zeichenketten einer formalen Sprache sind, und dass jede gültige Zeichenkette in dieser Sprache genau eine Bedeutung hat.
Natürliche Sprachen sind aber nicht künstlich geschaffen, sondern entwickeln sich nach eigenen Gesetzmäßigkeiten und eigener Dynamik. Insbesondere die für formale Sprachen charakteristische Eindeutigkeit, dass exakt definiert ist, welche Zeichenketten zulässig sind und welche Bedeutung die gültigen Zeichenketten haben, ist für natürliche Sprachen nicht gegeben. Natürlichsprachliche Texte sind deshalb mit Methoden und Verfahren der Informatik aus folgenden Gründen schwierig zu verarbeiten:

- Ausdrücke in natürlichen Sprachen können mehrdeutig sein,
- Texte unterliegen eigenen sprachstatistischen Gesetzmäßigkeiten,
- Texte spiegeln die Sprachdynamik einer Sprache wider.

Alle drei Aspekte **– Mehrdeutigkeit, Sprachstatistik** und **Sprachdynamik** – begegnen uns auf allen linguistischen Ebenen. Für ihre Beschreibung und Verarbeitung ist linguistisches Wissen erforderlich, das für die automatische Verarbeitung natürlicher Sprache in geeigneter Form zur Verfügung stehen muss. Welche Repräsentation linguistischen Wissens dabei verwendet wird, hängt zum einen von der zugrunde gelegten linguistischen Theorie ab, zum anderen von der zu bearbeitenden Aufgabe. Die wesentlichen Ansätze dazu stellen wir in Kapitel 2 vor. Zunächst aber sollen die genannten Phänomene genauer beschrieben werden.

Bei der Mehrdeutigkeit ist zwischen einer lexikalischen und strukturellen Mehrdeutigkeit zu unterscheiden. Ein Ausdruck ist lexikalisch mehrdeutig, wenn er mehr als eine Bedeutung oder Funktion hat. Meist werden Bedeutungsmehrdeutigkeiten in Form eines Lexikons beschrieben.

Beispiel – lexikalische Mehrdeutigkeit

Auf Wortebene haben die folgenden Ausdrücke mehr als eine Bedeutung:

Schloß (Nomen) – (a) Gebäude, (b) Schließvorrichtung als Teil einer Tür oder eines Tores
Aufstrich (Nomen) – (a) Lebensmittel, (b) Strichart bei Streichinstrumenten
Abstrich (Nomen) – (a) körpereigenes Material als Teil einer medizinischen Untersuchung, (b) Strichart bei Streichinstrumenten, (c) Einschränkung
leicht (Adjektiv) – (a) Gewichtsangabe, (b) Schwierigkeitsangabe
übersehen (Verb) – (a) alles im Blick haben, (b) etwas nicht bemerken

Oft hat ein Ausdruck mehr als eine Bedeutung, weil er verschiedenen syntaktischen Kategorien zuzuordnen ist, z.B. „Rauchen“ als Nomen oder Verb in dem Hinweis „Rauchen gefährdet die

Gesundheit“. Diese Art der lexikalischen Mehrdeutigkeit findet sich vor allem im Englischen wegen des Fehlens der einfachen Unterscheidung von Nomina und Verben bzw. Adjektiven durch Großschreibung, z.B. „light“ (Nomen) – Licht, „light“ (Verb) – anzünden, „light“ (Adjektiv) – leicht.
Auch Morpheme können mehrdeutig sein, z.B. die Endung „-en“ als Indikator des Indikativs eines Verbums („les-en“) oder der 1. Person Plural („wir les-en“).

Ein Ausdruck ist strukturell mehrdeutig, wenn er mehr als eine sinnvolle Zerlegung zulässt. Meist werden strukturelle Mehrdeutigkeiten in Form eines Strukturbaums dargestellt (vgl. Kapitel 2).

Die nachfolgenden Beispiele verdeutlichen leicht verallgemeinerbare, strukturelle Mehrdeutigkeiten auf Wort-, Phrasen- und Satzebene:

Beispiel – strukturelle Mehrdeutigkeit

Wort „Wachstube“ (Wach-stube, Wachs-tube)
„Druckerzeugnis“ (Druck-erzeugnis, Drucker-zeugnis)
„Urheberrechtsschutz“ (Urheber-Rechtsschutz, Urheberrechts- Schutz)
Phrase „zum Glück“ (als Einwortphrase: glücklicherweise, als Zweiwortphrase: für das Glück)
Satz „Peter sieht den Mann mit dem Fernrohr“ (Peter hat ein Fernrohr und sieht damit einen Mann, Peter sieht einen Mann, der ein Fernrohr hat)
„Meine sehr verehrten Damen und Herren“ (Meine sehr verehrten (Damen und Herren), Meine sehr verehrten Damen (und Herren))

Die Zerlegung einer Zeichenkette, beispielsweise ob die Phrase „zum Glück“ als Zweiwort- oder Einwortphrase verarbeitet werden soll, ist auch ein zentraler Bestandteil der sog. Tokenisierung, also der Festlegung, welche Zeichenketten in einem Text als kleinste zu verarbeitende Einheiten betrachtet werden (vgl. Kapitel 3).

Die Verteilung der Tokens in einem natürlichsprachlichen Text folgt auf allen linguistischen Ebenen den Gesetzmäßigkeiten der Sprachstatistik. Die Tokens sind nicht statistisch gleichverteilt, sondern folgen einer Potenzverteilung (engl. power law), die unter dem Namen Zipfsches Gesetz bekannt ist. Damit wird der Umstand bezeichnet, dass in einem natürlichsprachlichen Text einige wenige Tokens sehr häufig, die meisten aber selten vorkommen (vgl. Kapitel 5). Der Linguist George Kingsley Zipf (1902-1950), nach dem das Gesetz benannt ist, sieht in dem Zusammenhang ein „Prinzip des geringsten Aufwands“, weil die Tokens, die am häufigsten verwendet werden, meist sehr kurz sind und Tokens umso seltener auftreten, je länger sie sind Zipf ([27])[27].

Tab. 1 Beispiel – Häufigkeitsangaben aus dem Projekt Deutscher Wortschatz deu_newscrawl-public_2019_10K, 10.000 Sätze, , https://wortschatz.uni-leipzig.de/de/download

Token	Häufigkeit	Token	Häufigkeit	Token	Häufigkeit
der	4434	Jahren	184	Flüchtlinge	21
die	4208	können	180	Informationen	21
und	3166	Prozent	179	Möglichkeit	21
in	2610	habe	178	Unterstützung	21
den	1736	immer	175	tatsächlich	21

Das Zipfsche Gesetz gilt auch für Buchstaben sowie – mit Einschränkungen – auch für Buchstabenkombinationen (2 Buchstaben, 3 Buchstaben usw., sog. Buchstaben-n-Gramme) und Tokenkombinationen (2 Token, 3 Token usw. sog. Wort-n-Gramme). Welche Parameter des Zipfschen Gesetzes dabei für einen Text bzw. eine Sprache anzusetzen sind, muss empirisch bestimmt werden.

Natürliche Sprachen unterscheiden sich schließlich von formalen Sprachen durch eine evolutionären Prozessen in der Biologie vergleichbare Sprachdynamik auf Morphem-, Phrasen- und Wortebene. Auf Morphemebene führt diese Dynamik oft zu Problemen bei der Anwendung morphologischer Regeln, z.B. bei der Ableitung, ob ein Wort im Plural steht (E-Mails) oder in der Vergangenheitsform (gegoogled). Für eine angemessene Behandlung derartiger Sonderfälle – im konkreten Beispiel der Anwendung deutscher Flektionsregeln auf Wörter englischen Ursprungs – wird linguistisches Wissen benötigt.

Nicht nur finden in einer Sprache neue Wörter Verwendung, die es vorher in dieser Sprache nicht gegeben hat, sondern es verschwinden auch Wörter wieder oder nehmen eine andere Bedeutung an.

Sprachdynamik von Wörtern (vgl. Google Books NGram-Viewer [9])

Wort	**verwendet**	**Wort**	**verwendet**
googlen	seit ca. 2000	Frauenzimmer *(Frau)*	bis Mitte 19. Jh.
Laptop	seit ca. 1985	Kamisol *(Kleidungsstück)*	bis Mitte 19. Jh.
Entschleunigung	seit ca. 1985	Wangenstreich *(Ohrfeige)*	bis Mitte 19. Jh.

Zwischen dem Entstehungszeitpunkt eines Textes und der Auftretenshäufigkeit eines Wortes besteht ein enger Zusammenhang, der empirisch bestimmt werden kann (z.B. durch Auswertung von zeitbezogenen Metadaten). Ist nur ein Parameter bekannt, aber linguistisches Wissen über die Entwicklung einer Sprache vorhanden, kann anhand des Vorkommens bestimmter Wörter der Entstehungszeitpunkt eines Textes bestimmt werden oder es können auch Abschätzungen darüber gemacht werden, welche Wörter wie oft in einem Text zu erwarten sind (um beispielsweise Abweichungen davon für die Überprüfung von Urheberschaften zu verwenden) (vgl. Kapitel 7).

Zwei Ansätze für die Repräsentation und Verarbeitung linguistischen Wissens

Für die Beschreibung von linguistischen Gesetzmäßigkeiten können wir zwei grundlegende Ansätze unterscheiden, den **regelbasierten** und den **statistischen** Ansatz. Beide Ansätze schließen sich nicht gegenseitig aus, sondern werden bei konkreten Text Mining Anwendungen in der Praxis meist miteinander kombiniert.
Grundidee des regelbasierten Ansatzes ist es, die linguistischen Strukturen in den Zeichenketten eines Textes mit Hilfe von Mustern (engl. patterns) zu definieren. Mit Terry Winograd, dem Entwickler der sog. Blockswelt und Doktorvater des Google-Mitbegründers Larry Page, unterscheiden wir 5 Arten von Mustern (vgl. Winograd [22], S.35):

1. literal patterns (konkrete Zeichenfolge),
   z. B. „Leipzig“, „Volkswagen AG“,
   „weniger determiniert und weniger heteronom“
2. open patterns (Nutzung von Wildcards)
   z. B. „_ droht die Zahlungsunfähigkeit“, „_ droht _“,

„Der ontogenetische Prozess _, dass _ weniger determiniert und weniger heteronom _."

3. lexical patterns (auch lexikalische Kategorie möglich)
   z. B. „EN droht die Zahlungsunfähigkeit", „N droht PP",
   „Der ontogenetische Prozeß V, dass N PP weniger determiniert und weniger heteronom AUX."
4. variable patterns (Nutzung von Variablen; gleiche Variablen bezeichnen gleiches Wort)
   z. B. „X droht Y", „Rosenkrieg: X droht X"
5. Satzstruktur Patterns/Satzbaumuster
   z. B. „(Frageeinleitung) V Enom / Eakk <Alok> <Atemp>"
   Wieviel kostet eine Fahrkarte von Leipzig nach Dresden an einem Sonntag?

Für die Definition von Mustern können verschiedene Formalismen verwendet werden, beispielsweise reguläre Ausdrücke, kontextfreie Chomsky-Grammatiken (vgl. Vorlesung 3) oder um Attribute erweiterte kontextfreie Sprachen für die Informationsextraktion wie DIAL (Feldman&Sanger [7]) oder natürlichsprachliche Dialoge wie AIML ([1]). Der regelbasierte Ansatz ist besonders geeignet, wenn von der Textanalyse eine hohe Genauigkeit erwartet wird und bereits umfangreiche, qualitätsgesicherte linguistische Ressourcen für die Repräsentation linguistischen Wissens für eine Sprache vorliegen, beispielsweise in Form von Lexika und Grammatikregeln. Insofern sich regelbasierte Systeme immer auf konkrete Muster in einer Sprache beziehen, sind sie jedoch meist nur mit hohem Aufwand auf andere Sprachen zu übertragen.

Anders als der regelbasierte Ansatz berücksichtigt der statistische Ansatz die Häufigkeitsverteilungen von Zeichen und Zeichenketten im Text. Durch die Berechnung von statistisch signifikanten Verteilungen von Zeichen und Zeichenketten werden so ebenfalls auffällige Muster in Texten abgeleitet. Dadurch können linguistische Strukturen in Texten datengetrieben und ohne menschliche Interaktion in sog. unüberwachten Lernverfahren entdeckt und verarbeitet werden. Besonders geeignet ist der statistische Ansatz für die Erstellung von Listen von Wort-n-Grammen und Wörtern, die statistisch signifikant gemeinsam auftreten (Kookkurrenzen, vgl. Vorlesung 11). Statistische Verfahren lassen sich leicht von einer Sprache auf eine andere übertragen, bedürfen aber meist einer manuellen Nachbearbeitung, um Fehler oder Rauschen zu beseitigen.

Bei statistischen Verfahren ist zu unterscheiden zwischen sog. frequentistischen und Bayesschen Ansätzen (Howie [13]). Bei frequentistischen Verfahren wird die Wahrscheinlichkeit eines Ereignisses, also z.B. dem Auftreten einer Wortform in einem Text, als die relative Häufigkeit interpretiert, mit der es in einer großen Anzahl gleicher, wiederholter, voneinander unabhängiger Zufallsexperimente auftritt. Demgegenüber wird die Wahrscheinlichkeit eines Ereignisses beim Bayes‘schen Ansatz als Erwartungswert interpretiert, der sich aus einer Bewertung bisheriger Beobachtungen ableitet. Im Unterschied zu frequentistischen Verfahren wird daher bei Bayes‘schen Verfahren aus den gemachten Beobachtungen ein Modell abgeleitet, welches das Vorwissen und sog. A-Priori-Annahmen explizit ausdrückt. Im Text Mining bilden Bayes‘sche Ansätze die Grundlage für das Maschinelle Lernen und das semantische Clustern von Termen, sog. Topic Models (vgl. Vorlesung 9).

In der Entwicklung der Automatischen Sprachverarbeitung seit 1960 waren anfangs regelbasierte Verfahren bestimmend, wobei man sich stark an der Automatentheorie und Chomskys frühen Arbeiten zur Syntaxanalyse, insbesondere seinem Konzept von Transformationen (vgl. Chomsky [5]) orientiert hat. Musterbasierte Ansätze bildeten die Grundlage für erste natürlichsprachliche Dialogsysteme wie etwa Weizenbaums ELIZA, einem System, das oberflächlich das Gesprächsverhalten eines Psychotherapeuten simuliert ([24]) oder Winograds SHRDLU, einem System zur natürlichsprachlichen Steuerung eines Roboters in einer Spielwelt von Bauklötzchen unter Verwendung der Programmiersprache Planner ([22]). In den 70er Jahren wurde der

musterbasierte Ansatz zu einem regelbasierten syntaktischen Parsen unter Verwendung von Parsergeneratoren erweitert. Dadurch war es möglich, aus einer von Linguisten geschriebenen kontextfreien Grammatik für eine Anwendung (in einer natürlichen Sprache) (vgl. Abschnitt 2.3) automatisch einen Parser für diese Anwendung (in dieser Sprache) zu generieren. Erste größere Umsetzungen dieses Ansatzes waren Woods LUNAR, ein Dialogsystem für die Analyse von Gesteinsproben für Astronauten ([25], [26]) und HAM-ANS, einem generischen natürlichsprachlichen System mit Anwendungen in der Analyse von Strassenverkehrsszenen, der Unterstützung von Hotelreservierungen und der Abfrage einer relationalen Datenbank mit Fischereidaten unter Leitung von Walther von Hahn in Hamburg ([12]). Mit dem Aufkommen der Logikprogrammierung und der Programmiersprache PROLOG wurde dieser Ansatz für unifikationsbasierte Verfahren in der Sprachverarbeitung erweitert. Damit werden Grammatikmodelle bezeichnet, bei denen für jede linguistische Einheit eine Merkmalsstruktur im Sinne von morphologisch-syntaktischen Attribut-Wert-Paaren angenommen wird, wie beispielsweise in der Lexical Functional Grammar (LFG, [3]) oder Head Driven Phrase Structure Gramma (HPSG, [20]) (vgl. auch Abschnitt 6.1).
Anders als in der Sprachverarbeitung, bei der die Orientierung an menschlicher linguistischer Kompetenz lange Zeit als unhinterfragte Voraussetzung galt, wurden in der Spracherkennung (speech recognition) als Teilgebiet der Mustererkennung in der Informatik hauptsächlich statistische Verfahren eingesetzt (vgl. Jelinek [18]). Unter dem Einfluss der großen Erfolge der Spracherkennung, namentlich der Spracherkennungssysteme von IBM und Dragon Mitte der 90er Jahre, sind statistische Verfahren wie Hidden Markov Modelle und Bayes'schne Netze (siehe Kapitel 5 und 6) zunehmend auch in der Sprachverarbeitung eingesetzt worden. Der Paradigmenwechsel von regelbasierten zu statistischen Verfahren lässt sich besonders deutlich im Bereich maschineller Übersetzung nachvollziehen. Große regelbasierte vollautomatische Übersetzungssysteme wie z.B. Systran ([14]) sind in den 90er Jahren durch *translation memories*, als Übersetzungsunterstützungssysteme abgelöst worden. Dabei werden aus den von menschlichen Übersetzern übersetzten Sätzen, Absätzen oder Textabschnitten mit statistischen Verfahren Hypothesen für die Übersetzung von noch zu übersetzenden Texten generiert (vgl. Kugler et. al. [16], [17]). Damit dieser Ansatz erfolgreich angewendet werden kann, ist eine sehr große Menge von Texten und ihren Übersetzungen erforderlich. Die zunehmende Digitalisierung von Arbeitsprozessen und die Verbreitung des Internet schaffen dabei nicht nur die Voraussetzung für eine erfolgreiche Umsetzung dieses Ansatzes, sondern tragen auch entscheidend dazu bei, die Qualität der mit *translation memories* generierten Übersetzungen zu verbessern. Aufbauend auf diesem datengetriebenen Ansatz sind in den letzten Jahren verstärkt auch Verfahren des maschinellen Lernens, insbesondere konnektionistische Modelle, sowohl für die Übersetzungsunterstützung als auch das Retrieval von Fakten und Argumenten in monolingualen Dokumentkollektionen eingesetzt worden, wie es beispielsweise die von IBM unter dem Namen WATSON vermarktete Plattform für *cognitive computing* verdeutlicht ([11]). In Anbetracht der vielen Programme und Sprachdaten, die mittlerweile für kommerzielle und nicht-kommerzielle Zwecke öffentlich verfügbar sind, finden sich zunehmend auch integrierende Plattformen für die Bündelung von sprachverarbeitenden Programmen wie wir sie von den Marktplätzen und Plattformen im Internet kennen ([19]). Eine der ersten Plattformen auf Basis der Unstructured Information Management Architecture (UIMA, [21]) war die Plattform GATE (General Architecture for Text Engineering, [6]), eine gute Übersicht und Bündelung grundlegender Programmmodule für die Sprachverarbeitung auf Basis der Programmiersprache PYTHON findet sich im Natural Language Toolkit NLTK ([2]). Die Austauschbarkeit von Daten und Programmen für das Text und Data Mining wird aktuell im Zuge des Ausbaus von Forschungsinfrastrukturen vorangetrieben ([10]).

Nachfolgende Tabelle verdeutlich zusammenfassend, wie die Entwicklung der Sprachverarbeitung einhergeht mit der Entwicklung neuer Verfahren in der Informatik.

Dekaden	Technologie	Inspiration	Beispiel-anwendung	Implemen-tation
60er-70er	Mustererkennung/ Pattern matching	Transformations Grammatik	ELIZA, SHRDLU	Planner
70er-80er	Syntactisches Parsen und Dialog Management	TG, Speech acts	LUNAR, HAM-ANS	(kaskadierte) ATNs
80er-90er	Unifikation basierte Systeme	Logik Programming	LILOG	Prolog, LFG, HPSG
90er-2000er	Statistische Ansätze	Spracherkennung	Verbmobil, Translation Memories	HMMs, Bayes'sche Netze
seit 2010	Cognitive computing, konnektionistische Ansätze NLP Plattformen	Maschinelles Lernen, Workflow management	WATSON NLTK, GATE	Dependency parsing, neural networks UIMA

Übersicht: Historische Entwicklung derSprachverarbeitung

Literatur

[1] AIML – Artificial Intelligence Markup Language, http://www.aiml.foundation/doc.html

[2] Steven Bird, Ewan Klein and Edward Loper: Natural Language Processing with Python, Analyzing Text with the Natural Language Toolkit, O'Reilly Media 2009

[3] Joan Bresnan, Ash Asudeh, Ida Toivonen, Stephen Wechsler: *Lexical-Functional Syntax.* 2. Auflage, Wiley-Blackwell, Oxford, 2016

[4] Chomsky, Noam, The Logical Structure of Linguistic Theories, Plenum Press 1975, Nachdruck bei Springer „Classic Titles in Linguistics", auch verfügbar als Mikrofilm unter http://alpha-leonis.lids.mit.edu/wordpress/?page_id=466

[5] Chomsky, Noam, Syntactic Structures, Mouton 1957, Nachdruck bei Mouton - de Gruyter 2009

[6] Cunningham H., Maynard D., Bontcheva K. and Tablan V., GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, in: Proceedings. of the 40th Anniversary Meeting of the Association for Computational Linguistics 2002

[7] Feldman, Ronan/Sanger, James, The Text Mining Handbook, Cambridge University Press 2006

[8] Sabine Gründer-Fahrer, Antje Schlaf, Gregor Wiedemann, Gerhard Heyer, Topics and Topical Phases in German Social Media Communication during a Disaster, Natural Language Engineering 24 (2018), S. 221-264.

[9] Google Books Ngram Viewer, https://books.google.com/ngrams/info#

[10] Gerhard Heyer, Thomas Eckart und Dirk Goldhahn: *Was sind IT-basierte Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften und wie können sie genutzt werden?*. In: *Information - Wissenschaft & Praxis*, De Gruyter, 2015

[11] R High, The era of cognitive systems: An inside look at IBM Watson and how it works, IBM Corporation, Redbooks, 2012

[12] Hoeppner W., Morik K., Marburger H. (1986) Talking it Over: The Natural Language Dialog System HAM-ANS. In: Bolc L., Jarke M. (eds) Cooperative Interfaces to Information Systems. Topics in Information Systems. Springer, Berlin, Heidelberg

[13] David Howie: Interpreting Probability Controversies and Developments in the Early Twentieth Century Cambridge University Press 2002

[14] W. John Hutchins und Harold L. Somers, An Introduction to Machine Translation – 10. Systran, London, Academic Press, 1992

[15] S.C. Kleene, Introduction to Metamathematics, North-Holland Publishing Company 1971

[16] Marianne Kugler, Khurshid Ahmad, Gregor Thurmair (Hgs.), Translator's Workbench: Tools and Terminology for Translation and Text, Springer 1995

[17] Kugler, M., *Heyer*, G., Kese, R., v. Kleist-Retzow, B. und Winkelmann, G.: *The Translator's Workbench*: An environment for Multi-Lingual Text Processing and *Translation*, in: S. Nirenburg (Ed.), Progress in Machine Translation. IOS Press, 1993.

[18] Frederick, Jelinek. "Statistical methods for speech recognition.", MIT Press, 1997

[19] Geoffrey Parker, Sangeet Paul Choudary und Marshall W. Alstyne, Die Plattform-Revolution – Von Airbnb, Uber, PayPal und Co. lernen: Wie neue Plattform-Geschäftsmodelle die Wirtschaft verändern, MITP Fachbuchverlag, 2017

[20] Carl Pollard, Ivan A. Sag: *Head-Driven Phrase Structure Grammar*. (Studies in Contemporary Linguistics). University of Chicago Press, Chicago 1994

[21] UIMA, http://uima.apache.org/uima-specification.html

[22] Winograd, Terence, Procedures as a representation for data in a computer program for understanding natural language, Dissertation, MIT 1971, http://hci.stanford.edu/~winograd/shrdlu/AITR-235.pdf

[23] Winograd, Terence, Language as a Cognitive Process, Part 1, Addison-Wesley 1982

[24] Weizenbaum, Joseph, ELIZA A Computer Program For the Study of Natural Language Communication Between Man And Machine, Comnunications of the ACM, Volume 9 / Number 1, 1966

[25] Woods, W. A., Kaplan, R. M. and Nash-Webber, B. (1972), The *lunar* sciences natural language information system: final report, BBN Report 2378 (June 1972).

[26] *William A. Woods*, "Progress in Natural Language Understanding — An Application to *Lunar* Geology, AFIPS Conference Proceedings, Bd. 42, 1973

[27] George Kingsley Zipf, Human Behavior and the principle of least effort, Addison Wesley 1949