

Einleitung: Text(e) technologisch

Henning Lobin und Lothar Lemnitzer

Als *Texttechnologie* bezeichnen wir die Entwicklung von Verfahren, die präzise und deshalb algorithmisierbar sind und die auf Texte als semistrukturierte Daten angewendet werden. Die Texttechnologie ist in diesem Sinne wissenschaftlich begründete Praxis. Texttechnologie ist dadurch abzugrenzen

- von anderen sprachtechnologischen Domänen, deren Gegenstand zumindest nicht in erster Linie Texte sind,
- von der Datenbanktechnologie, die sich mit der Verwaltung und Abfrage von strukturierten Daten befasst, die Texttechnologie hingegen mit semistrukturierten Daten,
- von textorientierten Verfahren wie der theologischen, philologischen und juristischen Exegese, denen bezüglich des gleichen Untersuchungsgegenstandes andere Methoden und Untersuchungsziele zugrunde liegen.

Am nächsten liegt die Texttechnologie den Gebieten des Dokumentenmanagement und des Elektronischen Publizierens. Bei diesen steht jedoch das Dokument und insbesondere dessen äußere Eigenschaften im Mittelpunkt des Interesses, textlinguistische Fragestellungen, die geradezu konstitutiv sind für die Texttechnologie, spielen dort höchstens eine untergeordnete Rolle.

Zweifellos ist das Feld der Texttechnologie als sehr junge und hybride Disziplin in starkem Wandel begriffen und kann daher nur in Form einer Momentaufnahme dargestellt werden. Dennoch hoffen wir mit den Beiträgen in diesem Band zu zeigen, dass sich das Feld nicht nur über seinen Gegenstand konstituiert, sondern auch über einen Kanon von Methoden, über den sich in der Forschungsgemeinschaft zumindest partielle Einigkeit erzielen lässt. Die Anwendungen, die die Autorinnen und Autoren dieses Bandes er-

wähnen oder beschreiben, haben dabei Beispielcharakter: Das Potenzial der Texttechnologie ist in ihnen sicher noch nicht ausgeschöpft.

Damit sich das Forschungs- und Arbeitsgebiet Texttechnologie überhaupt etablieren konnte, musste eine Reihe von Voraussetzungen erfüllt sein, die uns heute zum Teil als selbstverständlich erscheinen.

Da es sich bei allen mittlerweile etablierten Verfahren um zumindest teilautomatische Prozesse handelt, müssen Rechner verfügbar sein, die einen textverarbeitenden Prozess in akzeptabler Zeit bewältigen können. Manche der in diesem Buch beschriebenen Verfahren sind sehr daten- oder rechenzeitintensiv. *Krüger-Thielmann* und *Paijmans* zeigen in ihrem Beitrag am Beispiel des „Latent Semantic Indexing“, dass auch heute noch die Grenze der Leistungsfähigkeit von Rechnern eine Rolle spielt. Wenn sich texttechnologische Verfahren im Alltag der Benutzer etablieren sollen, dann müssen sie sich mit ihrem Speicher- und Rechenzeitbedarf am Zuschnitt des durchschnittlichen Arbeitsplatzrechners orientieren. Dies spricht keinesfalls gegen die Erprobung sehr aufwändiger Verfahren, wie sie auch in diesem Band beschrieben werden: Es steht schließlich zu vermuten, dass sich auch in Zukunft die durchschnittliche Leistung von Rechnern nach oben verschieben wird.

Eine weitere Voraussetzung ist die Verfügbarkeit von Texten in digitaler Form. Dies ist nicht nur eine unabdingbare Voraussetzung, sondern geradezu der Motor für die Entwicklung texttechnologischer Verfahren. Hier nun kommt ein grundlegender Wandel zum Tragen. Nahezu alles, was seit etwa Mitte der achtziger Jahre gedruckt wurde und wird, geht durch eine digitale Vorstufe. Nahezu alle Texte, die wir im Alltag auf gedrucktem Papier zu Gesicht bekommen, liegen so auch in einem digitalen Format vor (vgl. Zimmer, 2000, S. 26). Dazu kommt die schier unüberschaubare Menge an Texten, die ausschließlich in digitaler Form zur Verfügung gestellt werden (Homepages, digitale Zeitungen, Referenzwerke seien hier als drei populäre Textsorten genannt). Drittens wird eine Vielzahl von Texten, die man als kulturelles Erbe oder aus sonstigen Gründen für bewahrenswert erachtet hat, nachträglich digitalisiert (man spricht von Retrodigitalisierung, vgl. Altrichter, 2001). Es ist also die Fülle des digital verfügbaren Schriftguts, die uns nach Verfahren zu ihrer expliziten Strukturierung und zu ihrer effizienten Erschließung suchen lässt.

Eine zumeist unterschätzte Voraussetzung texttechnologischer Verfahren ist die Verfügbarkeit von Zeichensystemen als Kodierungsvorschriften, die garantieren, dass entsprechende Programme alle verwendeten Schriftzeichen dieser Welt jederzeit auf verschiedenen Trägermedien (Bildschirm,

Ausdruck auf Papier) als solche darstellen. Eine texttechnologische Fragestellung, die noch keineswegs als definitiv gelöst gelten kann, ist, wie verschiedene Zeichensysteme in einem Text(korpus) parallel kodiert werden können bzw. sollten (näheres dazu bei *Sasaki* und *Witt*, in diesem Band). Die Kombination bzw. parallele Verwendung verschiedener Schreibsysteme und Schreibrichtungen erfordert einen Grad an Flexibilität von Annotations- und Analyseverfahren, die in der alltäglichen Praxis der digitalen Textverarbeitung noch nicht erreicht ist.

Textuelle Daten bilden heute in vielen Wirtschaftsbereichen einen zentralen Bestandteil im Produktionsprozess, so dass sich die wichtige Frage stellt, wie diese effizient produziert, gepflegt und genutzt werden können. Der Einsatz entsprechender Technologien ist vor allem in solchen Anwendungsbereichen sinnvoll, in denen mit sehr großen Textmengen und schnellen Publikationszyklen zu rechnen ist. In einem Zeitungsverlag werden die Texte oftmals bereits wenige Minuten, nachdem sie verfasst worden sind, im Web publiziert, bevor sie, u. U. in abgewandelter Form, für die Print-Ausgabe der Zeitung verwendet werden. Zugleich werden die Texte archiviert und über Datenbank-Schnittstellen mit Recherche-Funktionalität den Redakteuren und Online-Nutzern wieder zur Verfügung gestellt. Sachbuchverlage, die etwa Lexika oder Wörterbücher herstellen, generieren ihre Produkte heutzutage vielfach aus Datenbanken, in denen die redaktionell erstellten Texte ohne spezifischen Produktbezug gepflegt werden. Spezielle Klassifikationsmerkmale erlauben es, Teile derartiger Datenbanken als neue Produkte zusammenzufassen und in unterschiedlichen Medien zu vermarkten.

Man kann die Verwendung textueller Daten als das Durchlaufen eines Lebenszyklus konzeptualisieren. Danach folgen auf die Phasen der Strukturierung und der Datenerfassung die der Bearbeitung und Konvertierung, bevor sich der Zyklus mit der Phase der Revisionierung schließt. In der Phase der *Strukturierung* müssen die Daten analysiert und eine formale Dokumentgrammatik spezifiziert werden. Dieser Vorgang kann mit dem Programmieren in einer Programmiersprache verglichen werden, da ein ähnlicher Unterzyklus (Spezifikation, Implementation, Testen, Modifikation) zu durchlaufen ist wie beim Software Engineering. Die *Datenerfassung* kann darauf aufbauend entweder die durch die Dokumentgrammatik unterstützte Eingabe neuer Daten sein oder die teil- oder vollautomatische Konvertierung von Altdaten. In der Phase der *Bearbeitung* werden aus dem textuellen Datenbestand verschiedene Textversionen abgeleitet. Der Textbestand kann beispielsweise mehrere Sprachversionen in sich vereinen, möglicherweise auf der Ebene der kleinsten Texteinheiten parallelisiert. Die Aufgabe eines automa-

tischen Verarbeitungsprozesses ist es dann, die verschiedenen einzelsprachlichen Versionen aus dem Textbestand herauszufiltern und dabei ggf. noch weitere notwendige Umstellungs- oder Auswertungsprozesse durchzuführen. In der Phase des *Viewing* werden die textuellen Daten über sog. Style Sheets mit Darstellungsinformationen kombiniert, um sie in geeigneten Browsern anzeigen zu können. Die *Transformation* der Textbestände in andere Zielformate ähnelt der Festlegung von Style Sheets zu Zwecken des Viewings. Der Unterschied besteht darin, dass die strukturierten Textbestände in andere Auszeichnungsformate unwiderruflich überführt werden, um von dort aus mit anderen Verfahren weiterbearbeitet zu werden. In der Phase der *Revisionierung* werden die in den Test- und Anwendungsläufen der Dokumentgrammatik sowie der Bearbeitungs- und Viewing-Subsysteme gewonnenen Erfahrungen evaluiert, mit der Spezifikation neuer Anforderungen verbunden und bilden dann den Ausgangspunkt für einen neuen Lauf durch den Lebenszyklus.

Texttechnologie systematisch

Texte können wie auch andere sprachliche Erscheinungsformen als Zeichen verstanden werden. Zeichen kann man als die Vereinigung eines Inhaltskonzepts mit einer bestimmten Ausdrucksform definieren. Bei Texten besteht der Inhalt aus der durch den inhaltlichen Textaufbau bedingten Verbindung der Satzbedeutungen; der Ausdruck ist die äußere Form des Dokuments, vom Schrifttyp bis zur Seiten- oder Bildschirmgestaltung.

Die technologische Entwicklung hat schrittweise zur Entkopplung und Abstraktion dieser beiden Aspekte textueller Zeichen geführt. Sind beim handschriftlichen Verfassen eines Textes die inhaltliche und die ausdrucksseitige Realisierung noch untrennbar miteinander verbunden, entstehen gedruckte Texte aus der Kombination zweier getrennter, den Inhalt und den Ausdruck betreffender Arbeitsphasen. Sowohl Inhalt als auch Ausdruck eines Textdokuments weisen in sich Regularitäten auf, die in allgemeiner Form beschrieben werden können. Charakterisiert man etwa, wie ein Brief oder eine Gebrauchsanweisung normalerweise inhaltlich aufgebaut ist, spricht man nicht über den Inhalt als solches, sondern über die *Struktur des Inhalts*. In gleicher Weise kann man anstatt von einem bestimmten Ausdruck von der *Struktur des Ausdrucks* eines Textes sprechen. Das bedeutet, dass es nicht um konkrete Gestaltungsmerkmale geht, sondern um die Ausdrucksfunktion, die damit verbunden ist. Kursivdruck beispielsweise kann für verschiedene Zwecke verwendet werden, wird er aber für die Schreibung eines wichtigen

Begriffs im Text verwendet, geht es um die abstrakte strukturelle Funktion der Hervorhebung.

Im Zuge der Digitalisierung der Texterstellung und -bearbeitung sind zunächst Methoden und Techniken entwickelt worden, mit der Ausdrucksseite eines Textes flexibler umzugehen, als es beim Druck oder beim handschriftlichen Verfassen möglich ist. Texteditoren, die wie etwa Microsoft Word die unmittelbare Manipulation der Ausdrucksseite eines Textes, also seiner Gestaltung in den Vordergrund stellen, werden heute meistens nach dem „What you see is what you get“-Prinzip entworfen – danach entspricht das Aussehen des Textes auf dem Bildschirm weitgehend seiner Gestalt im Ausdruck. Mit der Idee des generischen Markup ist es allerdings auch möglich geworden, die Struktur des Ausdrucks formal zu beschreiben und damit Texte unabhängig vom konkreten Ausdruck so zu strukturieren, dass mit allgemeinen, einfachen Verfahren unterschiedliche Medien und Erscheinungsformen bedient werden können. Strukturierte Texte bilden deshalb eine Art informationellen Rohstoff im Publikationsprozess, der für verschiedene Zwecke eingesetzt und „raffiniert“ werden kann.

Gegenwärtig ist zu beobachten, wie auch die Aufarbeitung der inhaltlichen Struktur von Texten verstärkt ins Blickfeld des Interesses rückt (vgl. z. B. Berners-Lee, 1999). Geht es bei einer auf die Ausdrucksseite bezogenen Textstrukturierung vor allem um effizientes Publizieren, erschließt die inhaltliche Strukturierung den Text weitergehenden wissensverarbeitenden Prozessen. Die inhaltsbezogene Textstrukturierung ist allerdings mit ungleich größeren Problemen konfrontiert als die ausdrucksbezogene. In sie müssen Theorien zur grammatischen Struktur von Sätzen und zum Aufbau von Textbedeutung durch die textuelle Bedeutung von Sätzen einfließen. Da auf der Inhaltsseite viel mehr Information anfällt als auf der Ausdrucksseite, muss auch die Frage beantwortet werden, aufgrund welcher Prozesse die inhaltliche Struktur eines Textes explizit gemacht werden kann. Im Bereich der inhaltlichen Textstrukturierung konvergieren somit die Gebiete des Elektronischen Publizierens und der Linguistik.

Bereiche der Texttechnologie

Wenn auch zunächst einmal die Texttechnologie eher als Konglomerat denn als festumrissenes Forschungsfeld erscheinen mag, so werden doch bei näherem Hinsehen Sichtweisen auf Gegenstand, Fragestellungen und Methoden erkennbar, die von einer gewissen Stabilität dessen zeugen, was wir *Texttechnologie* nennen. Wir wollen deshalb im Folgenden den Blick zunächst auf

den gemeinsamen Gegenstand richten und anschließend die Verfahren, die wir der Texttechnologie zurechnen, weiter klassifizieren.

Den Blick auf den Gegenstand Text hat die Texttechnologie in großen Zügen von der Textlinguistik übernommen. *Storrer* (in diesem Band) zeigt, dass die Begriffsbestimmung von de Beaugrande und Dressler (1981), bei aller Kritikwürdigkeit im Detail, eine solide textlinguistische Arbeitsgrundlage für die Texttechnologie bilden kann. Sie zeigt zugleich, dass man das ideelle Objekt Text von seinem bislang typischen Trägermedium trennen muss, um sich von konzeptuellen Beschränkungen, die einzig und allein dem Trägermedium geschuldet sind, freizumachen. Dies führt zum Konzept des Hypertextes, der eine geeignetere Form der Präsentation von Texten im digitalen Medium darstellt. *Storrer* zeigt, wie eine textlinguistisch fundierte Textgrammatik dazu verwendet werden kann, um konventionelle Texte in Hypertexte zu überführen. Über die genuin textlinguistische Darstellung hinaus geht die Beschreibung quantitativer Aspekte von Texten, wie sie *Mehler* in seinem Beitrag „Quantitative Methoden“ präsentiert. Aus der Perspektive der Informatik wiederum sind Texte semistrukturierte Daten. Was das heißt, und wie Texte in dieser Sichtweise an die Theorie(n) formaler Sprachen angebunden werden können, das zeigt der Beitrag von *Morawietz* und *Mönnich* in diesem Band. Dieser Beitrag verlangt vom Leser allerdings mehr als nur grundlegende Kenntnis in der Theorie formaler Sprachen und zudem einen sicheren Umgang mit mathematischen Formalismen. Der Titel „Formale Grundlagen“ ist deshalb so zu verstehen, dass er dem Leser ein vertieftes Verständnis der formalen Eigenschaften von Markupsprachen erlaubt. Die Funktion von Markupsprachen ist die explizite Markierung der Struktur textueller Daten. Während also *Morawietz* und *Mönnich* das Konzept und die Formalisierung semistrukturierter Daten darstellen, sind die formalen Eigenschaften der Markupsprachen selber Gegenstand des Beitrags von *Lobin* in diesem Band.

Zu einem Text gehören Daten, die dieses Objekt beschreiben und es somit einer gezielten (thematischen) Suche zugänglich machen. Die Erstellung dieser Metadaten ist Teil der bibliothekarischen bzw. dokumentarischen Arbeitspraxis. Die Diskussion um Metadaten, um ihre Form, Funktion und Reichweite, ist im Zusammenhang mit Texten, die (ausschließlich) im digitalen Medium existieren, neu entfacht. Die sich dabei herausbildenden Standards beschreibt *Schmidt* in ihrem Beitrag.

Texttechnologischer Verfahren spielen auch eine Rolle im Software Engineering, beim Design komplexer Softwarearchitekturen. *Wolff* zeigt dies in seinem Beitrag anhand von Systemen für das elektronische Publizieren und

anhand sog. *Web Services*: „Web Services [...] bieten für sprach- und texttechnologische Anwendungen eine neue Perspektive, da mit ihnen die bisher nur durch aufwändige Adaption an proprietäre Schnittstellen mögliche Nutzung texttechnologischer Ressourcen und Dienste standardisiert und damit vereinfacht werden kann.“

Neben dem Gegenstand spielen die verwendeten Methoden bzw. Verfahren eine wichtige Rolle für die Definition des Faches. Wir können die folgenden Klassen von texttechnologischen Verfahren unterscheiden:

1. *Verfahren, bei denen Texte mit zusätzlichen Informationen angereichert werden* – Die Annotation sprachlicher Einheiten – vor allem Wort, Satzteil und Satz – nach linguistischen Kriterien spielt hier eine besondere Rolle. Durch die Identifikation und Klassifikation dieser Einheiten erhalten annotierte Texte eine zusätzliche Schicht an Informationen, die die Arbeit anderer texttechnologischer Verfahren erleichtern kann. Der Beitrag von *Ule* und *Hinrichs* in diesem Band stellt den Stand der Technik auf diesem Gebiet dar.
2. *Verfahren, mit denen Informationen aus Texten extrahiert¹ werden* – Dazu zählen zum einen Informationen, die für den Aufbau lexikalischer Ressourcen benötigt werden. *Lemnitzer* und *Wagner* stellen anhand einiger Fallstudien dar, welche für die Beschreibung lexikalischer Einheiten relevanten Informationen (teil-)automatisch aus einer genügend großen Textbasis extrahiert werden können. Anhand der beschriebenen Anwendungen wird zugleich der Nutzen annotierter Textkorpora deutlich. Im Bereich des Information Retrieval wird viel mit Dokument-Term-Matrizen gearbeitet, was die Extraktion (relevanter) Terme aus den zu beschreibenden Texten voraussetzt (siehe *Krüger-Thielmann* und *Paijmans*, in diesem Band).

Andererseits sind Informationen zu berücksichtigen, die für die Klassifikation dieser Texte hilfreich sind. In welcher Form die mehr oder weniger explizite Auszeichnung von Elementen der Textstruktur verwendet werden kann, zeigt *Rehm* in seinem Beitrag zu diesem Band. Dessen Titel, „Texttechnologie und das World Wide Web“, benennt einen wichtigen Antrieb für die texttechnologische Forschung und Entwicklung. Eine wichtige Lehre dieses Beitrags ist, dass wir nicht so lange warten können, bis die meisten Texte explizit strukturiert und durch Metadaten

¹ Genauer gesagt: Es werden Daten extrahiert, aus denen wiederum ein Benutzer in einem bestimmten Kontext Informationen erschließen kann. Solange es nicht zu Missverständnissen kommen kann, belassen wir es aber bei dieser verkürzten, im Übrigen aber auch eingespielten Redeweise.

optimal erschlossen sind, sondern unsere Verfahren auf den leider alles andere als idealen Ist-Zustand ausrichten müssen.

3. *Verfahren, mittels derer Texte zugänglich gemacht werden, um ein bestimmtes Informationsbedürfnis zu befriedigen* – Das klassische Information Retrieval antwortet dabei auf ein Informationsbedürfnis, das präzise formuliert werden kann. Die Beschreibungen der zur Verfügung stehenden Dokumente und die Beschreibung der Anfragen werden soweit synchronisiert, dass die wahrscheinlich interessanten Zieldokumente auf Grund einer Ähnlichkeitsbeziehung zwischen Anfrage und Dokument ermittelt werden. Grundlage der Ähnlichkeitsbeziehung ist typischerweise die lexikalische Form von Dokument und Anfrage. *Krüger-Thielmann* und *Paijmans* zeigen in ihrem Beitrag aber auch, dass es bereits zumindest im experimentellen Stadium Verfahren gibt, die diese einfache lexikalische Ebene verlassen. Das Textmining hingegen ist eine stärker explorative Technik, die relevante Information auf unscharfe Anfragen hin zu ermitteln versucht. Die hier verwendeten grundlegenden Techniken stellt *Mehler* in seinem Beitrag „Textmining“ vor. Während Information Retrieval und Textmining auf das Auffinden kompletter Texte als Antwort auf ein Informationsbedürfnis hin ausgelegt sind, entwickeln sich in jüngster Zeit Abfragesprachen, die auf einzelne Elemente oder Gruppen von ihnen zielen. Diesen Abfragesprachen muss die Struktur der angesprochenen Texte zugänglich sein, was deren explizite Strukturierung voraussetzt. Der Aufsatz von *Schulz* und *Meuss* in diesem Band – „Abfrage strukturierter Dokumente“ – beleuchtet dieses Feld. Da es sich hierbei um Abfragen über semistrukturierten Daten handelt, stellt dieser Beitrag eine praktische Anwendung der in dem Beitrag von *Morawietz* und *Mönnich* skizzierten Theorie dar.
4. *Verfahren, mittels derer aus Texten abgeleitete Objekte erstellt werden* – Dieses Objekt kann wiederum ein Text sein, der eine andere Funktion erfüllt als der Primärtext. Dies ist der Fall bei Zusammenfassungen, die heute nicht mehr ausschließlich intellektuell erstellt werden. Welche wissenschaftlichen und statistischen Verfahren dazu verwendet werden, um zumindest zu einem funktionsgerechten Textsurrogat zu kommen, das beschreibt *Endres-Niggemeyer* in ihrem Beitrag. Bei Text-to-Speech-Systemen schließlich soll aus einem Text eine lautliche Repräsentation erstellt werden, die einer von einer Person gesprochenen Äußerung so weit wie möglich ähnelt. Einen Überblick hierüber geben *Stöber*, *Schröder* und *Hess* in ihrem Beitrag zu diesem Band.

Zu der Klasse der erzeugenden Verfahren gehören auch diejenigen, bei denen ein Text auf Grund von in anderer Form repräsentierten Propositionen entsteht. Hierzu gehört die Generierung geschriebener Texte, zum Beispiel aus Datenbankeinträgen. *Stede* zeigt in seinem Beitrag anhand eines Fallbeispiels, wie bei der Generierung von Texten auch komplexe Kohärenzbeziehungen modelliert werden können.

Der vorliegende Band gliedert sich in drei Teile. Im ersten Teil „Grundlagen und Methoden“ werden die theoretischen Grundlagen der Texttechnologie systematisch abgehandelt. Der zweite und dritte Teil des Bandes rücken Anwendungsaspekte in den Vordergrund. Während die Beiträge im zweiten Teil die Anwendung der Texttechnologie auf die Bereiche der Linguistik und der Sprachtechnologie illustrieren, wird im dritten Teil ihre Nutzung für die praktischen Zwecke der Informationserschließung fokussiert.

Literaturverzeichnis

- Altrichter, Helmut (2001): “Retrodigitalisierung in Deutschland – Versuch einer Zwischenbilanz”. <http://www.bsb-muenchen.de/mdz/forum/altrichter/>.
- Berners-Lee, Tim (1999): *Weaving the Web. The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco: Harper.
- de Beaugrande, Robert-Alain und Dressler, Wolfgang Ulrich (1981): *Einführung in die Textlinguistik*. Tübingen: Niemeyer.
- Zimmer, Dieter (2000): *Die Bibliothek der Zukunft: Text und Schrift in den Zeiten des Internet*. Hamburg: Hoffmann und Campe.