Originalveröffentlichung in: Willée, Gerd/Schröder, Bernhard/
Schmitz, Hans-Christian (Hrsg.): Computerlinguistik. Was geht, was
kommt? Computational Linguistics. Achievements and Perspectives. -
Sankt Augustin: Gardez! Verlag, 2002. S. 198-202.
(Sprachwissenschaft, Computerlinguistik und neue Medien, Band 4)

Henning Lobin

Texttechnologie – eine neue Perspektive der Computerlinguistik

1 Einleitung

Obwohl sich die Computerlinguistik seit ihren Anfängen mit Texten befasst, kann erst seit den achtziger Jahren davon gesprochen werden, dass digitale Texte zu einem eigenen Forschungsgegenstand geworden sind. Dieses hat einerseits mit dem Siegeszug des PC und der Textverarbeitung zu tun, andererseits hat die Ausbreitung des World Wide Web zu einer unüberschaubaren Menge digital verfügbarer Texte geführt, die als Ressource für textbezogene computerlinguistische Forschung verfügbar sind und für die Tools und Systeme entwickelt werden. Parallel dazu wurden auch Texte, die in einem kommerziellen oder industriellen Zusammenhang stehen, zunehmend digitalisiert, wobei Methoden der Standardisierung von Repräsentation und Verarbeitung zum Einsatz kommen, die in letzter Zeit unter der Bezeichnung *Texttechnologie* zusammengefasst werden.

Das Gebiet der Texttechnologie kann als ein neuer Zweig der Computerlinguistik verstanden werden, der sich noch in seiner Ausformungsphase befindet und trotzdem schon Einfluss auf andere Teilgebiete der Computerlinguistik gewonnen hat. Standen bislang eher konkrete Anwendungserfordernisse im Vordergrund, wird mittlerweile der Blick auch auf die Grundlagen der Texttechnologie gerichtet und ihre Bezüge zur Computerlinguistik und zur Textlinguistik beleuchtet.

2 Texte in einem Wirtschaftskreislauf: besondere Anforderungen

Textuelle Daten bilden heute in vielen Bereichen einen zentralen Bestandteil im Produktionsprozess von Waren und Wissen – dabei stellt sich die Frage, wie große Textmengen im digitalen Medium effizient produziert, gepflegt und genutzt werden können. Der Einsatz entsprechender Technologien ist vor allem in solchen Anwendungsbereichen sinnvoll, wo 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 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 für unterschiedliche Medien zu vermarkten.

Man kann die Verwendung textueller Daten als das Durchlaufen eines Lebenszyklus konzeptualisieren. In der Phase der *Strukturierung* müssen die Daten analysiert und für sie eine formale Dokumentgrammatik spezifiziert werden. 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 automatischen 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 *Konvertierung* 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, 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.

3 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 Text 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. 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, zum Aufbau von Textbedeutung durch die Kombination von Sätzen und vieles mehr 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 mit der Linguistik.

4 Texttechnologie als Disziplin: Bereiche

Aus gegenwärtiger Sicht lässt sich eine Reihe von Teilbereichen nennen, in denen relevante Aspekte der Texttechnologie untersucht werden. Im Bereich der *Grundlagen* dreht es sich um die bei den Markup-Sprachen verwendeten Varianten formaler Sprachen und Grammatiken, um Hypertextmodelle, Systemarchitekturen und Dokumentmanagement und um die Bezüge zur Textlinguistik, etwa zu Fragen der Textkohärenz oder der Textsortenklassifikation. Zentral im Bereich der *Textstrukturierung* ist die Frage nach den Eigenschaften von Markup-Systemen und der anzuwendenden Methodologie bei der Festlegung des Annotationsvokabulars. Fragen der Informationserschließung aus Texten, der Hypertextualisierung und der Anfrage an Textdatenbasen lassen sich im Teilbereich *Textzugang* zusammenfassen. Die voll- oder teilautomatische Generierung von Texten in all ihren Aspekten bildet den Bereich der *Texterzeugung*, mit *Textderivation* kann der Bereich bezeichnet werden, in dem es um inhaltliche und strukturelle Transformationsprozesse geht, etwa die Erzeugung automatischer Textzusammenfassungen. Für die Linguistik von besonderem Interesse ist der Bereich der *Korpustechnologie*, der sich mit Methoden und Techniken befasst, die die Erstellung, Nutzung und Bearbeitung großer natürlichsprachlicher Korpora befasst.

Institutionell befasst sich die Gesellschaft für linguistische Datenverarbeitung seit langem mit Themen der Texttechnologie. Seit 1998 existiert ein eigener Arbeitskreis zu dieser Thematik (s. z. B. Lobin 1999, Mehler, Lobin im Erscheinen, Lobin, Lemnitzer im Erscheinen), die auch die GLDV-Frühjahrstagungen in den Jahren 1999 (vgl. Gippert 1999) und 2001 (vgl. Lobin 2001) geprägt hat. Verschiedene Projekte öffentlicher Drittmittelgeber befassen sich zurzeit mit Grundlagenproblemen der Texttechnologie, sodass mit einer Ausweitung der Forschungen auch bei der anwendungsorientierten Forschung in den nächsten Jahren zu rechnen ist. Zeitschriften wie "Text Technology" (McMaster) und "Markup Languages" (MIT Press) widmen sich zentralen Forschungsfragestellungen der Texttechnologie.

5 Ausblick

Die zukünftige Entwicklung der Texttechnologie ist eng mit der Weiterentwicklung des World Wide Web verbunden. Die Vision des *Semantic Web* (vgl. Berners-Lee 1999) steht und fällt mit der Verfügbarkeit standardisierter semantischer Repräsentations- und Verarbeitungsverfahren für textuelle Information. Methoden und Modelle der computerlinguistischen Kernbereiche werden dazu in stärkerem Maße zu nutzen sein als bisher. Zugleich gibt es Schnittstellen zu anderen Bereichen der Informationstechnologie und des elektronischen Publizieren. Da in diesem Bereich auch in der Zukunft mit großen Anwendungspotentialen zu rechnen ist, ist es an der Zeit, den Bereich der Texttechnologie als einen wichtigen Teilbereich der Computerlinguistik zu konturieren.

Literatur

Berners-Lee, Tim (1999): Weaving the Web. The Original Design and Ultimate Destiny of World Wide Web by Its Inventor. – San Francisco: Harper.

Gippert, Jost (ed.) (1999): Multilinguale Korpora. Codierung, Strukturierung, Analyse. – Prag: Enigma.

Lobin, Henning (ed.) (1999): Text im digitalen Medium. – Opladen, Wiesbaden: Westdeutscher Verlag.

Lobin, Henning, Lothar Lemnitzer (ed.) (im Erscheinen): Texttechnologie. Perspektiven und Anwendungen. – Tübingen: Stauffenburg (= Stauffenburg Einführungen).

Lobin, Henning (ed.) (2001) [Book on Demand]: Sprach- und Texttechnologie in digitalen Medien. Gießen: GLDV.

Mehler, Alexander, Henning Lobin (eds.) (im Erscheinen): Aspekte der Texttechnologie: Systeme und Methoden zur Analyse und Annotation von natürlichsprachlichen Texten. Opladen, Wiesbaden: Westdeutscher Verlag.