Probleme der Textauswahl für einen elektronischen Thesaurus: Tagungsbericht

Michael Schlaefer, *Deutsches Wörterbuch,*
Akademie der Wissenschaften zu Göttingen, Deutschland

Abstract: Problems Concerning the Choice of Text Materials for an Electronic Thesaurus: Conference Report. At the invitation of the Akademie der Wissenschaften in Göttingen a colloquy was held on 1 and 2 November 1996. Experts in lexicographical studies and single word research participated in a discussion of problems concerning the choice of text materials for an electronic thesaurus.

Keywords: DEUTSCHES WÖRTERBUCH, CORPORA, TEXT SELECTION, GERMAN LANGUAGE

Zusammenfassung: Auf Einladung der Göttinger Akademie der Wissenschaften wurde am 1. und 2. November 1996 ein Kolloquium abgehalten. Lexikographen und Wortforscher diskutierten Probleme der Textauswahl für einen elektronischen Thesaurus.

Stichwörter: DEUTSCHES WÖRTERBUCH, CORPORA, TEXTSELEKTION, DEUTSCHE SPRACHE

Das im Jahr 1838 von den Brüdern Jacob und Wilhelm Grimm begonnene *Deutsche Wörterbuch* wird seit 1960 in der Trägerschaft der Akademien in Göttingen und Berlin im Buchstabenbereich A - F neu bearbeitet.

Dem Göttinger Neubearbeitungsteil des *Grimmschen Wörterbuchs* liegt eine Quellenauswahl zugrunde, die etwa 6000 Texte des 8. - 20. Jahrhunderts umfaßt. Dieses Quellenmaterial ist etwa zwischen 1960 und 1970 über Mikroverfilmung erfaßt und in ein Zettelarchiv für die Buchstaben D - F umgesetzt worden. Dieses Zettelarchiv umfaßt einen Bestand von etwa 2,5 Millionen Belegen. Durch technische Veränderungen bedingt, ist seit etwa 1980 eine finanziell vertretbare künftige Weiternutzung der Quellenfilme fraglich geworden. Aus diesem Sachverhalt ergeben sich weitreichende Konsequenzen für Überlegungen zur Fortführung der Neubearbeitung des *Grimmschen Wörterbuchs* über F hinaus.

Die Göttinger Kommission für das *Deutsche Wörterbuch* hat daher seit Mitte der achziger Jahre nach Wegen gesucht, den Zugang zum Göttinger Quellenmaterial zu erhalten, um für die historische Lexikographie und Wortforschung auch über den engeren Planungsbereich des laufenden Neubear-

beitungsvorhabens hinaus eine wissenschaftliche Arbeit in den Standards des *Grimmschen Wörterbuchs* zu ermöglichen. Ab dem Jahr 1986 beauftragte die Kommission die Göttinger Arbeitsstelle des Grimmschen Wörterbuchs, Quellen des Göttinger Neubearbeitungsteils in eine maschinenlesbare Form umzusetzen und daraus Belegmaterial für den historischen deutschen Wortschatz zu gewinnen. Das Arbeitsvorhaben erhielt die Bezeichnung "Sicherungsmaßnahme". Für eine erste Phase, die vor allem der Entwicklung und Erprobung der theoretischen Grundlagen, der praktischen Vorgehensweise und der Auswertbarkeit gewonnener Materialien dienen sollte, wurde das in der Sicherungsmaßnahme zu erfassende Textvolumen auf einen Bestand von 50 Texten begrenzt. Als erreichbares Belegvolumen wurde etwa eine Million elektronisch verfügbarer Stichwortnachweise angenommen.

Dieses erste Arbeitsphase der Sicherungsmaßnahme kam im Jahr 1994 weitgehend zum Abschluß, so daß es wünschenswert erschien, die gewonnenen Erfahrungen und Ergebnisse im Rahmen eines Arbeitsgesprächs mit Fachleuten für deutschsprachige Textcorpora zu diskutieren und Perspektiven für die Weiterarbeit zu entwickeln. Am 1. und 2. November 1995 fand daher auf Einladung der Kommission für das Deutsche Wörterbuch (Göttingen) in der Trägerschaft der Göttinger Akademie der Wissenschaften ein Kolloquium zu "Problemen der Textauswahl für einen elektronischen Thesaurus" statt.

Dieses Kolloquium wurde zugleich als Einleitung zu einer Reihe von wissenschaftlichen Veranstaltungen verstanden, die sich mit den Problemen der historischen deutschen Wortforschung als einem Teil der Rahmenbedingungen für die "Sicherungsmaßnahme" befassen sollten.

Außer den im weiteren genannten Referenten nahmen der Präsident und der Sekretär der Göttinger Akademie, Mitglieder der Göttinger und der Berliner Wörterbuchkommissionen, das Kollegium der Göttinger Arbeitsstelle sowie einzelne Gäste an der Veranstaltung teil.

Das Programm des Arbeitsgesprächs für den 1. November 1996 umfaßte drei Themenbereiche. Unter dem Schwerpunkt "Zielsetzungen und Vorarbeiten" war aus dem Projekt der "Sicherungsmaßnahme" zu berichten. Daran schlossen sich die Themenblöcke "Corpuserfahrungen in sprachhistorischen und gegenwartssprachlichen Projekten" und "Sprachgeographische Corpusaspekte" an.

Im ersten Themenschwerpunkt stellte der Vorsitzende der Göttinger Kommission für das Deutsche Wörterbuch, R. Bergmann (Bamberg), Geschichte und Arbeitsstand des Göttinger Digitalisierungsvorhabens dar. Er erläuterte die unternehmensstrategischen Grundzüge und ging in diesem Zusammenhang auf die in den letzten Jahren stark veränderten Rahmenbedingungen für die Wortforschung ein. Eine zukunftsorientierte Gestaltung der Grundlagen für historische deutsche Wortforschung könne nicht mehr in isolierten Einzelvorhaben, sondern nur in Kooperation vieler Wissenschaftler und Institutionen des gesamten deutschsprachigen Raumes erreicht werden. Der Leiter der Arbeitsstelle, M. Schlaefer (Göttingen), referierte über den aktuellen Stand der

Arbeiten an der "Sicherungsmaßnahme". Er zeigte Nutzungsmöglichkeiten der verfügbaren Materialien auf und stellte schwerpunktmäßig Grundlagen der Textauswahl bzw. der Lemmatisierung und der Belegauswahl für die Sicherungsmaßnahme vor.

Die Diskussion zu den Einführungsvorträgen ließ erkennen, daß für einen Teil der Anwesenden die "Sicherungsmaßnahme" einen Schritt in Richtung auf einen sprachnationalen Thesaurus darstellte. Die unter einer solchen Annahme zu behandelnden wissenschafts- und fachpolitischen Perspektiven traten daher gegenüber den konzeptionellen arbeitspraktischen Fragen des Göttinger Projekts in den Vordergrund. Angesichts der knapp bemessenen Diskussionszeit wurde vereinbart, am Ende des Kolloquiums das Thema "Gesamtthesaurus" im Rahmen einer Grundsatzdiskussion zu behandeln.

Im Themenschwerpunkt "Corpuserfahrungen" gingen zunächst die Referenten K.P. Wegera (Bochum) und H.-J. Solms (Halle) unter verschiedenen Gesichtspunkten auf Erfahrungen mit dem Bonner Corpus frühneuhochdeutscher Quellen ein, das seinerzeit als eines der ersten maschinenlesbaren Corpora für sprachhistorische Arbeiten erstellt worden war. Dabei standen für Herrn Wegera Fragen der Textauswahl und der Textumfänge im Vordergrund, für Herrn Solms Fragen des Zusammenhangs von sich wandelnden Erkenntnisinteressen und Corpusstrukturen sowie Aspekte der polyfunktionalen Nutzbarkeit elektronischer Corpora.

Mit unterschiedlichen Problemen der Strukturierung historischer Quellencorpora befaßten sich auch die Referate von U. Goetz (Bamberg), H. Kämper-Jensen (Mannheim) und U. Haß-Zumkehr (Mannheim). U. Goetz stellte am Beispiel des seit 1990 laufenden und kurz vor dem Abschluß stehenden Bamberg-Rostocker Gemeinschaftsprojekt "Die Entwicklung der Großschreibung im Deutschen von 1500 bis 1700" Möglichkeiten für eine straffe raum-zeitliche und quantitative Corpusgliederung dar. H. Kämper-Jensen referierte über das "IdS-Fremdwörtercorpus", das aufgrund der langen Entstehungszeit sowie der unterschiedlichen Schwerpunktsetzung in seiner Struktur heterogene Züge aufweist. Den Gesichtspunkt der Polyfunktionalität betonte U. Haß-Zumkehr in ihrem Bericht über das im Aufbau befindliche sogenannte Historische Corpus des Instituts für deutsche Sprache in Mannheim. Der Beitrag von M. Wermke (Mannheim) über Vorüberlegungen zum Aufbau elektronischer Textcorpora in der Dudenredaktion machte in sehr anschaulicher Weise den Unterschied zwischen Textcorpusorientierung und Belegcorpusorientierung deutlich. In diesem Zusammenhang hob er die Bedeutung des Modells einer lemmatisierten exemplarischen Belegsammlung gegenüber einem unbegrenzten Wortformenreservoir für die Wörterbucharbeit hervor.

Im zweiten Themenschwerpunkt gingen W. Bauer (Wien), R. Ris (Zürich) und P. Ott (Zürich) auf unterschiedliche Zusammenhänge der sprachgeographischen Schichtungen des Deutschen mit vorhandenen oder künftigen Corpusbildungen ein. Zuerst berichtete W. Bauer über "Historische Quellen des Wörterbuches der bairischen Mundarten in Österreich zu den bairisch-öster-

reichischen Sprachvarietäten des 14. bis 19. Jahrhunderts". Über "Möglichkeiten eines schweizerischen Corpusteils" sprachen R. Ris und P. Ott. R. Ris erörterte die zentrale Frage nach der Tragfähigkeit der dem [1]DWB zugrundeliegenden Quellenbasis für eine geschichtlich zuverlässige Abbildung des Schweizerischen. Er wies eine starke Einseitigkeit der DWB-Quellenbasis im Bereich protestantischer Autoren aus Zürich nach und entwickelte eine alternative Corpusstruktur für die jüngere schweizerische Varietät des Deutschen. P. Ott stellte anhand detaillierter sprachgeschichtlicher, sprachsoziologischer und sprachgeographische Merkmale einzelner Quellen und Quellengruppen ein exemplarisches Corpus für das schweizerische Deutsch des 16. Jahrhunderts vor.

Für den zweiten Tag des Kolloquiums waren Beiträge zu exemplarischen fachsprachlichen Aspekten eines historischen Corpus sowie zur Frage von Wörterbüchern als Quellen vorgesehen. K. Jacob (Dresden) befaßte sich in diesem Zusammenhang mit "Techniksprachlichen Quellen des 17.-19. Jahrhunderts". Er machte deutlich, daß gerade unter lexikalischen Gesichtspunkten eine Ausblendung der jüngeren Techniksprache, wie sie auch noch das *Grimmsche Wörterbuch* zeigt, kaum vertretbar ist. G. Wagenitz (Göttingen) stellte in seinem Vortrag "Wortbildung und Textarten in der Biologie" Übergänge von fachsprachlich-botanischen Bildungen in die Gemeinsprache dar und zeigte eine Reihe von Textsorten auf, die als Vermittler zwischen den beiden sprachsoziologischen Bereichen wirkten.

Am Beispiel der "Wörterbücher des 16. Jahrhunderts als Thesaurusquellen" erläuterte P.O. Müller (Erlangen) die besondere Problematik, die sich in diesem metasprachlichen Quellensektor durch immanente philologische Bedingungen und Traditionen ergeben.

H. Henne (Braunschweig) ging unter dem Thema "Ein Deutscher Thesaurus und die deutschen Wörterbücher des 17. und 18. Jahrhunderts" über den engeren thematischen Rahmen ausgreifend auf Möglichkeiten und Erwartungen ein, die er mit elektronischen Textcorpora und Wörterbüchern verbunden sah.

Für die Weiterarbeit in der Sicherungsmaßnahme haben die Beiträge eine Reihe wichtiger Anregungen gegeben. Auf konzeptioneller Ebene wird die Abwägung zwischen Text- und Belegarchiv ebenso kritisch vorzunehmen sein, wie die ausschließliche Stützung auf das Göttinger Quellencorpus zu hinterfragen ist. Eine noch größere Aufmerksamkeit dürfte künftig der Auswahl und der Umfangsbegrenzung des digitalisierten Quellenmaterials zukommen. Für den arbeitspraktischen Bereich stellt sich die Aufgabe der Bündelung der Ressourcen. Angesichts der Vielfältigkeit der Corpusinitiativen liegen hier große Aufgaben für die Koordination, aber auch Chancen dafür, kooperativ eine tragfähige Plattform für die historische Wortforschung zu schaffen. Angesichts der Bedeutung, die den Referaten für die Sicherungsmaßnahme, aber auch für die Erstellung historischer Corpora allgemein zukommt, strebt der Vorsitzende der Göttinger Kommission für das Deutsche Wörterbuch die Veröffentlichung

eines Sammelbandes für 1997/98 an.

Die zum Abschluß des Kolloquiums angesetzte Grundsatzdiskussion über einen Gesamtthesaurus für die deutsche Sprache führte gedanklich weit über den Rahmen hinaus, der mit der Göttinger "Sicherungsmaßnahme" und dem Thema des Kolloquiums gegeben war. In der Diskussion und einzelnen Statements wurde die Frage nach der Aktualität und Dringlichkeit eines Thesaurusaufbaus ebenso angesprochen, wie die Stellung eines künftigen Gesamtthesaurus gegenüber den in Bearbeitung befindlichen Wörterbüchern. Dabei spielte das Problem der Bearbeitungszeit gerade des *Grimmschen Wörterbuchs* eine besondere Rolle. Ein weiterer Diskussionsschwerpunkt betraf Überlegungen zum Zusammenwirken der deutschsprachigen Länder und deren Forschungseinrichtungen. Schließlich wurden mehrfach Entwürfe über verschiedene Thesauruskonzeptionen mit Text- bzw. Belegarchiven, aber auch mit verschiedenen Wörterbuchprojekten gedanklich angerissen. Der Wunsch, den Plan für einen solchen Thesaurus der deutschen Sprache in weiteren Kolloquien zu diskutieren, wurde allgemein unterstützt. Als eine fokusartige Bündelung der mit der Grundsatzdiskussion eröffneten fachpolitischen und konzeptionellen Dimensionen läßt sich die mehrfach benutzte Formulierung betrachten, es gehe um die deutsche Lexikographie des 21. Jahrhunderts.