Das Zusammenspiel automatisierter und interpretativer Verfahren bei der Aufbereitung und Auswertung mündlicher Daten
Ein Fallbeispiel aus der angewandten Wissenschaftssprachforschung

Cordula Meißner (Universität Leipzig)
Franziska Wallner (Universität Leipzig)

Obwohl die Erforschung der Wissenschaftssprache auch im Bezug auf das Deutsche in den letzten Jahren verstärkt Beachtung gefunden hat, stehen selbst für die geschriebene Modalität der Wissenschaftssprache nur in sehr begrenzten Umfang elektronisch verfügbare Datensammlungen als empirische Grundlage für diesbezügliche Untersuchungen zur Verfügung. Für die gesprochene Wissenschaftssprache fehlten sie lange Zeit vollkommen. Sowohl für den Bereich der Diskurs- und Variationsforschung als auch für den Bereich der Sprachlehr- und lernforschung stellt die Untersuchung der gesprochenen Wissenschaftssprache auf breiterer empirischer Basis ein noch weitgehend unbearbeitetes Desiderat dar.

Mit dem an der Universität Leipzig (Herder-Institut), der Aston University, Birmingham und der Universität Wrocław erarbeiteten Korpus GeWiss („Gesprochene Wissenschaftssprache kontrastiv – Deutsch im Vergleich zum Englischen und Polnischen") wurde 2013 für die gesprochene Wissenschaftssprache exemplarisch eine flexibel nutzbare Korpusressource der wissenschaftlichen Öffentlichkeit zur Verfügung gestellt (vgl. Fandrych/Meißner/Slavcheva 2012, 2014). Sie umfasst ca. 120 Aufnahmestunden von gesprochener deutscher, englischer und polnischer Wissenschaftssprache (mehr als 1 Mio Token, die als Transkripte vorliegen und mit Audiofiles synchronisiert einsehbar und analysierbar sind). Es handelt sich um ein Vergleichskorpus, welches zwei zentrale Genres der mündlichen Wissenschaftskommunikation umfasst – Vorträge/Referate und Prüfungsgespräche. Datengrundlage sind dabei zum einen Aufnahmen von L1-Sprecher/inne/n der drei Vergleichssprachen, zum anderen deutschsprachige Realisierungen dieser Genres von L2-Sprecher/inne/n in Deutschland, Großbritannien und Polen.

Ziel des aktuellen Folgeprojekts „Gesprochene Wissenschaftssprache digital" ist die Optimierung der Nutzungsmöglichkeiten des Gewiss-Korpus und die Erprobung weiterer methodischer Möglichkeiten zur Auswertung und Analyse der Daten. Im Fokus stehen dabei neben den Realisierungsmöglichkeiten wissenschaftlicher Sprachhandlungen wie beispielsweise Diskurskommentierungen, Verweisen und Zitaten auch der Wortschatz der allgemeinen gesprochenen Wissenschaftssprache sowie lexikalische, morphologische und syntaktische Besonderheiten der gesprochenen Wissenschaftssprache. Der Vortrag stellt

anhand eines Beispiels für die aktuelle Arbeit mit den Korpusdaten vor, wie hierbei interpretative und formbasiert automatisierbare Ansätze zusammenwirken.

Der erste Teil des Vortrags gibt einen Überblick über den Aufbau und das Design des GeWiss-Korpus und dokumentiert die aktuell erarbeiteten Aufbereitungsschritte, welche die orthografische Normalisierung, das Wortarten-Tagging, die Lemmatisierung sowie die pragmatische Annotation umfassen. Sowohl die Verfügbarkeit einer solchen Datenbasis zur gesprochenen Wissenschaftssprache als auch die genannten Formen der Aufbereitung eröffnen neue methodische Zugänge zum Untersuchungsgegenstand. In seinem zweiten Teil wird der Vortrag zeigen, wie diese Zugänge im Forschungsprozess genutzt werden können. Er will damit auch einem Beitrag zur Methodenreflexion leisten im Hinblick auf das Zusammenspiel von manuellen und automatischen Verfahren bei der Datenaufbereitung und zum anderen in Bezug auf das Ineinandergreifen von formbasiert-automatisierbaren Analyseverfahren und hermeneutisch-interpretativem Umgang mit den ermittelten Ergebnissen bei der Datenauswertung.

Exemplarisch sollen hierfür die Möglichkeiten zur Beschreibung der wissenschaftssprachlichen Handlung des Diskurskommentierens durch manuelle Annotation und darauf aufbauende korpusmethodische Ermittlung „guter Kandidaten" betrachtet werden.

Zur korpusbasierten Untersuchung pragmatischer Phänomene wurde bislang entweder der Weg der manuellen Annotation (vgl. Baur et al. 2013, Maynard/Leicher 2007, Alsop/Nesi 2012) oder der Weg eines automatischen Zugriffs auf der Formebene über konkrete Lexeme oder datengeleitet ermittelte N-Gramme (vgl. z.B. Scharloth/Bubenhofer 2012, Rühlemann 2010) verfolgt. Der erste Weg ermöglicht eine funktional orientierte, erschöpfende Erfassung des Phänomens. Eine derartige Korpusaufbereitung ist jedoch mit einem hohen Maß an Zeitaufwand verbunden und erfordert mehrstufige Korrekturdurchgänge sowie eine intensive Abstimmung unter den Annotierenden. Der zweite Weg ermöglicht zwar eine schnelle Datengewinnung, beschränkt die Ergebnisse jedoch auf angenommene, zuvor ausgewählte Formmerkmale oder datengeleitet ermittelte rekurrierende Wortfolgen, die anschließend der funktionalen Interpretation bedürfen.

Im Vortrag wird exemplarisch anhand einer Analyse zur sprachlichen Handlung der Diskurskommentierung ein Ansatz vorgestellt, der beide Wege kombiniert: Ausgehend von den im GeWiss-Korpus für ein Teilkorpus von Konferenzvorträgen annotiert vorliegenden Diskurskommentierungen (vgl. Fandrych 2014) werden über korpuslinguistische Analysen typische Formmerkmale dieser Sprachhandlung ermittelt. Daraus wird eine Suchabfrage abgeleitet, mit der in nicht-annotierten Korpora nach „guten Kandidaten" für die betrachtete

sprachliche Handlung gesucht werden kann. Es werden dazu einerseits von L2-Sprecher/inne/n gehaltene Konferenzvorträge und andererseits studentische Referate als Untersuchungskorpora herangezogen. Die hier ermittelten Belege geben Aufschluss über das Realisierungsspektrum der Handlung in anderen, sich sprachlich, kompetenz- und diskursartbezogen unterscheidenden Konstellationen.

Die Ergebnisse eröffnen weitere Möglichkeiten für kontrastive und kompetenzorientierte Untersuchungen: So lassen sich aus den Treffern, die der Suchausdruck (als Muster einer fachkompetenten L1-Realisierungsform) in Daten von fachkompetenten L2-Sprecher/inne/n und in Daten von fachlichen Novizen erzielt, Schlüsse über Variation und Kernmerkmale der Handlung in diesen sprachlichen Konstellationen ziehen. Die Untersuchung zeigt exemplarisch, wie ausgehend von manuell vorgenommenen pragmatischen Annotationen mittels formbasiert automatisierbarer korpuslinguistischer Analysemethoden weitergehende varianzbezogene Analysen durchgeführt werden können, welche die Generierung von Hypothesen für eine weiter ausdifferenzierende Beschreibung des Untersuchungsgegenstandes ermöglichen. Die Verbindung aus interpretativem und datenorientiertem Zugriff ermöglicht es zudem, aus den manuell annotierten Daten „typische Beispiele" zu extrahieren, die über häufige Formeigenschaften verfügen. Dies bietet eine weitere Anwendungsmöglichkeit für die Fremdsprachenvermittlung.

Literatur:

Baur, Benedikt/Gräfe, Karen/Lange, Daisy/Schmidt, Julia (2013). Dokumentation zur Annotation der Diskurskommentierungen. Abrufbar unter: https://gewiss.uni-leipzig.de/open.php?url=Annotationsdokumentation_GeWiss.pdf [Stand: 06.11.2014].

Fandrych, Christian (2014): Metakommentierungen in wissenschaftlichen Vorträgen. In: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation), 95-111.

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.) (2014): *Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen*. Heidelberg: Synchron-Verlag. (= Wissenschaftskommunikation).

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (2012): The GeWiss Corpus: Comparing Spoken Academic German, English and Polish". In: Schmidt, Thomas/Wörner, Kai (Hg.): Multilingual corpora and multilingual corpus analysis. Amsterdam: Benjamins. 319 – 337. (=Hamburg Studies in Multilingualism 14).

Maynard, Carson/Leicher, Sheryl (2007): Pragmatic Annotation of an Academic Spoken Corpus for Pedagogical Purposes. In: Fitzpatrick, Eileen (Hg.): *Corpus Linguistics Beyond the Word: Corpus Research from Phrase to Discourse*. Amsterdam: Rodophi. pp. 107-116

Alsop, Sian/Nesi, Hilary (2012): Annotating a corpus of spoken English: the Engineering Lecture Corpus (ELC). In: Mello, Heliana/Pettorino, Massimo/Raso, Tomaso (Hg.). *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*. Firenze: Firenze University Press, 58-62.

Rühlemann, Christoph (2010): What can the corpus tell us about pragmatics? In: O'Keeffe, Anne / McCarthy, Michael (Hg.): *The Routledge handbook of corpus linguistics*. London/New York: Routledge, 288-301.

Scharloth, Joachim/Bubenhofer, Noah (2012): Datengeleitete Korpuspragmatik. Korpusvergleiche als Methode der Stilanalyse. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analyse. Berlin / N.Y.: de Gruyter, 195-230.