Fandrych, Christian; Meißner, Cordula; Wallner, Franziska (Hrsg.): **Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora.** Tübingen: Stauffenburg, 2017 (Deutsch als Fremd- und Zweitsprache, Schriften des Herder-Instituts (SHI), 11). – ISBN 978-3-95809-075-0. 213 Seiten, € 36,80.

Besprochen von **Gerhard Jaiser:** Amman / Jordanien

<https://doi.org/10.1515/infodaf-2020-0033>

Gesprochene Sprache stellt für die Korpuslinguistik in vielerlei Hinsicht eine besondere Herausforderung dar. Die Abweichungen der gesprochenen Sprache von der Standardsprache erschweren eine automatische Annotation ebenso wie die Auswertung durch Nutzer, denn „mündliche Daten [sind] bei exakter Transkription nur bedingt elektronisch nutzbar“ (16). Vor allem das Institut für Deutsche Sprache (IDS) hat in diesem Bereich in den letzten zwei Jahrzehnten Beachtliches geleistet. Daran anschließend und in Zusammenarbeit mit dem IDS widmet sich mittlerweile das Herder-Institut Leipzig intensiv der besonderen Herausforderung der gesprochenen Wissenschaftssprache. Für eine wissenschaftliche Bearbeitung dieses Bereichs ist das Herder-Institut durch langjährige Arbeit an Aufarbeitung und insbesondere auch didaktischer Vermittlung von Wissenschaftssprache prädestiniert. Stellvertretend für Letzteres sei das entsprechende Modul im Projekt *Dhoch3* für Masterprogramme im Ausland genannt, das unter Federführung des Herder-Instituts entstand.

Der vorliegende Band beschreibt in einer Einführung und acht Einzelbeiträgen verschiedene Aspekte der Erstellung und wissenschaftlichen Auswertung des Projekts *GeWiss* (Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen). Hierbei geht es einerseits um die enormen Herausforderungen der automatischen Annotation und der weiterhin notwendigen manuellen Nachbearbeitung, andererseits, dargestellt an exemplarischen Beispielen, um den Mehrwert dieser erweiterten Annotation bei der Analyse der Daten. Entsprechend ist der Band aus korpus-theoretischer Sicht interessant im Hinblick auf künftige Erstellung und Annotation von Korpora ebenso wie für Anwenderinnen und Anwender, die die Korpus-Daten für Analysen verwenden wollen.

Christian Fandrych stellt in seinem Beitrag das Projekt *GeWiss* vor und stellt die verschiedenen Bereiche vor, denen sich die folgenden Beiträge widmen. *GeWiss* umfasst notgedrungen nur einen Ausschnitt der fast unüberschaubaren Varianten der mündlichen Wissenschaftssprache, nämlich „wissenschaftlicher und studentischer Vortrag (einschließlich [...] Diskussionen) sowie Prüfungs-

gespräch“ (14). Fandrych fasst den Stand der Arbeit an *GeWiss* zusammen und verweist auf die detaillierteren Darstellungen der folgenden Beiträge. Einerseits werden dabei „exemplarisch zentrale Wege der Weiterentwicklung mündlicher Korpora“ aufgezeigt, andererseits „Desiderate und Perspektiven der Korpusentwicklung“ (13) diskutiert. Wesentliche Aspekte, die angesprochen werden, sind etwa die bereits erwähnte Gratwanderung zwischen getreuer Transkription und notwendiger Vereinheitlichung, die weiterhin hohe Fehlerquote bei automatischer Annotation und die dadurch notwendige manuelle Nachbearbeitung, Probleme der Annotation von Mehrworteinheiten sowie die noch kaum vorher versuchte Annotation von pragmatischen Kategorien. Die folgende Feststellung Fandrychs beschreibt hierbei wohl für mehr oder weniger alle vorgestellten Arbeiten die zentrale Herausforderung: „Aufwand und Nutzen der Annotation müssen in einem sinnvollen Verhältnis zueinander stehen“ (20).

Daniel Jettkas Beitrag über die Integration des *GeWiss*-Korpus in die CLARIN-Infrastruktur ist primär aus korpus-theoretischer Hinsicht interessant. Metadaten und Transkripte mussten vereinheitlicht und auf Konsistenz überprüft werden, um in die umfangreich genutzten Vorgaben von CLARIN (Common Language Resources and Technology Infrastructure) integriert werden zu können. Dies sollte dazu dienen, „die Aussichten auf die langfristige Verfügbarkeit des Korpus“ (46) zu verbessern.

Sechs weitere Beiträge widmen sich schließlich vertieft zwei Hauptaspekten der Annotation mündlicher Wissenschaftssprache, nämlich Wortarten (Teil 2) und pragmatischer Annotation (Teil 3). Zunächst stellt Franziska Wallner „[d]ie orthographische Normalisierung und Annotation von Wortarten gesprochen-sprachlicher Daten“ dar. Die Notwendigkeit hierfür ist offensichtlich, da nur so dem Nutzer ermöglicht wird, alle Varianten unter einer Suchanfrage aufzufinden. Das hierfür verwendete Programm (*OrthoNormal*) sollte durch die durchgeführte Studie soweit verbessert werden, dass mittelfristig „ein Verfahren zum automatisierten POS-Tagging gesprochener Sprachdaten ermöglicht“ (52) wird. Wallner stellt jedoch fest, dass aktuell noch ein enormer Nachbearbeitungsaufwand notwendig ist, so dass es notwendig ist, „die eingesetzten Verfahren zu optimieren, um die Fehlerquote und den damit einhergehenden Korrekturaufwand zu verringern“ (75).

Kornelia Bochniak, Karen Gräfe und Anna Iliash widmen sich der „Annotation von Modal-, Intensitäts- und Fokus/Gradpartikeln im *GeWiss*-Korpus“. Auch hier müssen die Autorinnen feststellen, dass die Nutzbarkeit der vorhandenen Daten bisher durch Desiderate bei der Annotation eingeschränkt bleibt: „Einerseits steht eine immer größere Menge gesprochen-sprachlicher Daten zur Verfügung, die linguistisch untersucht und (etwa nach Wortarten) annotiert werden könnte, jedoch fehlen noch für viele grammatische Phänomene adäquate Werk-

zeuge zur Berücksichtigung aller Spezifika“ (81). Die von ihnen durchgeführte manuelle Annotation „anhand von Kriterienlisten aus der einschlägigen Literatur“ (104) versteht sich entsprechend als Beitrag zur Diskussion über das weitere Vorgehen bei der Annotation von Partikeln.

In einem weiteren Beitrag untersucht Franziska Wallner die Kategorie der Diskursmarker, also „syntaktisch nicht oder nur schwach eingebundene [...] gesprächsorganisierende Einheiten“ (107). Die Untersuchung versteht sich als Ausgangsbasis für eine notwendige Untersuchung „auf der Basis einer größeren Datengrundlage“ (120). Franziska Wallner und David Stoppel widmen sich im letzten Beitrag des zweiten Teils Partikelverben. Hier liegt die besondere Schwierigkeit in der häufigen Distanzstellung von Partikelverben, die eine automatische Annotation bisher kaum möglich macht. Wie in mehreren der anderen Studien des Bandes liegt das Hauptaugenmerk auf einer weitgehend manuellen Nachbearbeitung, die einerseits die Nutzbarkeit des *GeWiss*-Korpus erheblich verbessert, andererseits aber auch Vorarbeit für die Verbesserung automatischer Verfahren sein will.

Aus Anwenderperspektive besonders interessant sind die letzten beiden Beiträge, die sich der Annotation pragmatischer Strukturen widmen. Die Schwierigkeiten hierbei sind enorm, „da eine solche Annotation ein sehr aufwendiges Verfahren darstellt, das zudem noch kaum auf gesicherte und einigermaßen breit akzeptierte pragmatische Kategorien zurückgreifen kann“ (22).

Sabrina Sadowskis Beitrag zur „Annotation von Zitaten und Verweisen“ erhebt den Anspruch, „die Grundlage für eine pragmatisch-funktionale Korpusanalyse [zu legen] und weitere Nutzungsmöglichkeiten für Forschung und Lehre [zu eröffnen]“ (146). Die Auswertung kann dabei jedoch lediglich „erste Teilergebnisse“ präsentieren, „die sich aus der Analyse eines Teils des Gesamtkorpus zusammensetzen“ (156). Cordula Meißner schließlich widmet sich im umfangreichsten Beitrag des gesamten Bandes der „automatischen Ermittlung von Belegen für sprachliche Handlungen auf der Basis manueller pragmatischer Annotation“. Der Fokus liegt hierbei auf Diskurskommentierungen in Konferenzvorträgen deutscher Muttersprachler. Meißner diskutiert ausführlich die Annotation metasprachlicher Handlungen, die mündliche Sprachhandlungen, im vorliegenden Fall wissenschaftliche Vorträge, gliedern. Besonderes Ziel hierbei war es, die umfangreich manuell annotierten Kategorien für automatische Abfragen anderer, nicht entsprechend annotierten Korpora nutzbar zu machen. Meißner stellt schließlich auch die Bedeutung für den Fremd- und Zweitsprachenunterricht heraus und erhebt den Anspruch, „die Kombination von manueller Annotation und darauf aufbauender automatischer Suchabfrage [...] für die Vermittlung rezeptiver und produktiver Fähigkeiten nutzbar [zu] machen“ (210).

Abschließend kann festgehalten werden, dass der Nutzen des *GeWiss*-Projekts für den Fremdsprachenunterricht potentiell groß ist, wenn er auch erzwungenermaßen im vorliegenden Band nur skizziert werden kann, denn „die Mündlichkeitsforschung [ist] bisher in der Sprachdidaktik noch kaum angekommen“ (26). Die Vermittlung mündlicher Wissenschaftssprache stellt eine enorme Herausforderung im nichtmuttersprachlichen akademischen Wissenschaftsbetrieb dar, wobei die Notwendigkeit korpusanalytischer Studien als Voraussetzung für entsprechende didaktische Fortschritte kaum bezweifelt werden kann. Gerade die pragmatische Annotation, die in den letzten beiden Beiträgen diskutiert wird, ist hierfür unabdingbare Voraussetzung. Den offensichtlich enormen Aufwand im Rahmen eines Projekts wie *GeWiss* mit dem Bedarf an noch weit größeren, für weitere Analysen abrufbaren Daten zu vereinbaren, stellt eine Herausforderung für die Zukunft dar. *Gesprochene Wissenschaftssprache – digital* ist zweifellos ein wichtiger Schritt auf dem Weg zur notwendigen Standardisierung in diesem Bereich. Gewünscht hätte man sich allenfalls noch eine kurze Zusammenfassung der Herausforderungen und Perspektiven am Schluss. Natürlich ist all das der einleitenden umfassenden Beschreibung des Projekts sowie den Querverweisen zu entnehmen. Gerade für Nutzer der Korpusdaten wäre eine abschließende Darstellung des Status quo dennoch nützlich gewesen.