Kratochvílová, Iva/Wolf, Norbert Richard (Hgg.) (2013): Grundlagen einer sprachwissenschaftlichen Quellenkunde (Studien zur Deutschen Sprache 66). Tübingen: Narr Francke Attempto Verlag GmbH + Co. KG, ISBN 978–3–8233–6836–6, 382 S.

Der Sammelband, der Beiträge zu allgemeinen korpuslinguistischen Problemen und spezifischen korpusbasierten bzw. -gestützten Studien enthält, ist dem Andenken Hans Wellmanns (1936–2012) gewidmet – sein letzter Beitrag befindet sich in diesem Band. Die 27 Beiträge sind in sieben Abteilungen gruppiert, deren Überschriften im Inhaltsverzeichnis und in der linken Kopfzeile zu finden sind; eine weitere Orientierungshilfe gibt das Sach- und Wortregister am Ende des Bandes. Aus dem Vorwort der Herausgeber geht hervor, dass es sich um Beiträge der im Oktober 2011 im Bildungshaus Sambachshof und an der Universität Würzburg gehaltenen II. Internationalen Konferenz „Korpuslinguistik Deutsch-Tschechisch kontrastiv" handelt. Der Band bildet eine Fortsetzung zum Sammelband *Kompendium Korpuslinguistik* (2010) von denselben Herausgebern.

Die Abteilung **Zur Einleitung** besteht aus zwei Beiträgen. Unter der Rubrik *DeuCze: Von der Struktur und Benutzbarkeit eines kleinen zweisprachigen Korpus* gibt **Iva Kratochvílová** einen Überblick über das in Zusammenarbeit zwischen den Universitäten Würzburg und Opava erstellte deutsch-tschechische kontrastive *DeuCze*-Korpus, das zum Zeitpunkt der Konferenz sein Zehnjahresjubiläum feierte. Das knapp eine Million Wortformen umfassende Korpus besteht aus einem bidirektionalen Übersetzungskorpus belletristischer Texte seit 1990 und einem kleinen Vergleichskorpus deutscher und tschechischer journalistischer Texte zum Irak-Krieg aus dem Jahr 2003. Über den quantitativen Zugang hinaus stehe die text- und kontextbezogene qualitative Auswertung des Korpus im Mittelpunkt. Kratochvílová geht es um die kontextuelle und diskursive Bedingtheit von Kollokationen, d. h. um Fälle, die „nicht mehr zufällige Kontextualisierungen" sind, sondern „[…] usuellen, frequenziell relevanten Sprachgebrauch [manifestieren] und […] wesenhafte semantische Beziehungen [realisieren]" (S. 20), sowie darum, ob im entsprechenden Diskurs in zwei Sprachen Unterschiede aufgedeckt werden können. Dies ist bei den Basen *Krieg/Irakkrieg – válka*, die im Irak-Krieg-Korpus die häufigste kollokationale Rekurrenz aufweisen, der Fall: Die Kollokationen weisen distributionelle Disparitäten auf. Die Ergebnisse sprechen dafür, dass, wie die Verfasserin hervorhebt, für kontrastive Zwecke schon kleinere Parallelkorpora aufschlussreich sind. Der Beitrag informiert zudem über bisherige Leistungen sowie aktuelle und künftige Forschungsthemen im Rahmen des Projekts, von denen einige in dem Band durch eigene Beiträge repräsentiert sind.

Während es im ersten Beitrag um ein kleines zweisprachiges Korpus ging, stellen **Marc Kupietz** und **Elena Frick** im zweiten Beitrag *Korpusanalyseplattform der nächsten Generation* eine sehr große einsprachige Korpussammlung vor. Zuerst wird über die rund 50jährige Geschichte der im Mannheimer IDS (Institut für Deutsche

Sprache) beheimateten Textkorpora berichtet. Heute ist das *Deutsche Referenzkorpus* (*DeReKo*) mit (zur Zeit der Konferenz) 5,4 Milliarden Wörtern aus unterschiedlichen Textarten seit 1956 und mit drei konkurrierenden morphosyntaktischen Annotationen, aus denen der Nutzer ein eigenes virtuelles Korpus zusammenstellen kann, das weltgrößte Korpus von geschriebenem Deutsch. Die Anfang der 1990er Jahre entwickelte Recherche- und Analysesoftware *COSMAS* (seit 2003 *COSMAS II*) hat sich jedoch bei der ständigen Zunahme der Textmenge als zu aufwendig erwiesen. Deswegen arbeitet man seit 2011 an einem moderneren, benutzerfreundlicheren Analysewerkzeug *KorAP* (*Korpusanalyseplattform der nächsten Generation*), das in den nächsten 15–20 Jahren den Anforderungen der linguistischen Forschung genügen soll. Die Kapazität soll für Korpusarchive mit 50 Milliarden Wörtern ausreichen. Aufgrund von Tests vorhandener Korpusrecherche- und Korpusanalysesoftware und Benutzerwünschen wird eine neue Abfragesprache mit einer intuitiveren Syntax entwickelt, die neben lexikologischen auch komplexe syntaktische Abfragen über mehrere Annotationsschichten erlaubt. Weitere Stichwörter sind multiple Tokenisierung, Abwärtskompatibilität zu *COSMAS II*, Orientierung am *Corpus Query Lingua Franca*, Kombinierbarkeit mit externen Analysemodulen, Berücksichtigung von Multimodalität sowie kollaborationsfördernde Schnittstellen für Forschungsinfrastrukturen (etwa *CLARIN*) und virtuelle Forschungsumgebungen (z. B. *TextGrid*). Die Plattform wird die Arbeitsbedingungen der Linguisten bedeutend verbessern.

Von **(Meta-)Lexikographie** handeln sechs Artikel. **Annette Klosa** geht auf *Primäre, sekundäre und tertiäre Quellen in der Lexikographie* ein. Es geht um die Nutzung von Quellen und korpuslinguistischen Verfahren für die Erarbeitung des korpusbasierten deutschen Online-Wörterbuchs *elexiko*, das ca. 300 000 Lemmata mit Angaben zur Bedeutung, Verwendung, Grammatik, Rechtschreibung und Wortbildung umfasst. Die Primärdaten entstammen dem virtuellen *elexiko*-Korpus mit 2,8 Milliarden Wortformen aus Pressetexten des *DeReKo*. Als Analysesoftware dient *COSMAS II*, als zentrale statistische Methode die Kookkurenzanalyse, die ggf. durch gezielte Korpussuchen ergänzt wird. Das Lesartenspektrum und die Verwendungsmuster eines Lemmas werden empirisch ermittelt und mit möglichst aussagekräftigen Beispielen veranschaulicht. Nur hinreichend belegte Lesarten werden aufgenommen. Wenn eine Lesart im Korpus fehlt, wird dies in einem Hinweis vermerkt. Häufiger soll es jedoch der Fall sein, dass im Korpus Lesarten ermittelt werden, die in anderen Wörterbüchern nicht vorkommen. Wörterbücher als Sekundärquellen dienen vor allem zur Überprüfung. Bei Angaben zu Synonymen und Etymologie wird durch Verweise und Links auf Einträge in entsprechenden Wörterbüchern hingewiesen. Auch tertiäre Quellen wie Grammatiken und linguistische Forschungen werden durch Verweise oder Links angegeben. Als Wörterbuchgrammatik dient die am IDS erarbeitete *Grammatik der deutschen Sprache* (1997) bzw. die darauf beruhende elektronische *grammis*. Mit Ausbau und Weiterentwicklung des *elexiko* wird sowohl den Forschern als auch dem breiten Publikum ein wertvolles, wissenschaftlich fundiertes Wörterbuch zur Verfügung stehen.

Sabine Kromes Beitrag *Digitale Datenflut: Chancen und Tücken eines Textkorpus zur deutschen Gegenwartssprache. Anforderungsprofil, Methoden und Instrumentarien zur Beobachtung des aktuellen Sprach- und Schreibgebrauchs* thematisiert die Einbeziehung primärer (digitaler), sekundärer und tertiärer Quellen bei der Aktualisierung von Wörterbüchern der deutschen Gegenwartssprache. Krome geht zuerst auf die Erweiterung der Lemmaliste durch Neologismen und dann auf orthographische Entscheidungen ein. Neologismen werden anhand des *WAHRIG Textkorpus* ^{digital}, das zwei Milliarden Wortbelege aus Pressetexten umfasst und Suchmöglichkeiten nach Sachbereich, Zeitung/Zeitschrift oder Jahrgang ermöglicht, demonstriert. Dabei ist die Frequenz in den einzelnen Jahrgängen ein zentrales Kriterium. Anhand von Beispielen wird aufgezeigt, wie sich Neologismen etablieren, wie sie von Okkasionalismen zu unterscheiden sind und wie Neubedeutungen allmählich Fuß fassen. Jugendsprachliche Neologismen wurden in einem kleineren literarischen Spezialkorpus des 20. Jahrhunderts ausgewertet. Durch diachrone Statistiken zu konkurrierenden Bezeichnungen wird u. a. der Werdegang der heute dominierenden *toll* und *cool* für ‚sehr gut' nachgezeichnet. Der Rechtschreibung wird anhand des *AG Korpus*, das auf den drei größten deutschen Korpora von IDS, *DUDEN* und *WAHRIG* basiert, nachgegangen. Am Beispiel der Getrennt- und Zusammenschreibung und der Groß- und Kleinschreibung wird beobachtet, wie gut den amtlichen Regeln gefolgt wird. Die Ergebnisse dienen der Vereinheitlichung der Schreibungen in den Wörterbüchern und der Schwerpunktsetzung im Schulunterricht. Um ein besseres Bild vom tatsächlichen Usus zu bekommen, sollten auch Blogs, E-Mails und andere Internettexte recherchiert werden. – Für die künftige Entwicklung der lexikographischen Basis seien spezielle Teil- und Subkorpora (historische Texte, gesprochene Quellen einschl. Mediensprache im Radio und Fernsehen, Jugendliteratur, Sprache im Netz usw.) nötig. Erwünscht sei auch eine engere Kooperation zwischen Institutionen, die über Korpora verfügen.

Hans Wellmann plädiert in seinen Beitrag *Muster der Adjektivderivation in alten und neuen Korpora – und ihre Reflexe im Wörterbuch* für einen – je nach der Forschungsfrage – flexiblen Umgang mit verschiedenen Korpora. Der Beitrag ist eine Metaanalyse, in der Ergebnisse vorhandener Untersuchungen zur Adjektivderivation aufeinander bezogen werden. Im Mittelpunkt stehen *-haft*-Derivate, die vom Mittelhochdeutschen bis zur deutschen Gegenwartssprache stark zugenommen haben. Der heutige Gebrauch wird mit Kookkurenzanalysen in *COSMAS II* erhellt, wobei Vorkommensfrequenzen okkasionelle Bildungen von usuellen unterscheiden lassen. Für die Ersteren werden spezielle Gebrauchssituationen, für die Letzteren Grade der Transparenz herauskristallisiert, wonach die Resultate mit den in gängigen Wortbildungslehren dargestellten semantischen und morphologischen Mustern verglichen werden. Die semantischen Klassen können u. a. auf konkurrierende Bildungsweisen hin untersucht werden. Diachrone korpusgestützte Untersuchungen zeigen, dass im Laufe der Zeit der im Mhd. noch sehr seltene komparative Bedeutungstyp den früher vorherrschenden possessiv-ornativen Typ völlig zurückgedrängt hat. Resümierend stellt Wellmann fest, dass die heutigen umfangreichen elektronischen Korpora

vielerlei variationslinguistische Fragestellungen ermöglichen. So können z. B. die Frequenzanteile der einzelnen Adjektivsuffixe diatopisch, diastratisch (bes. im Hinblick auf geschriebene vs. gesprochene Sprache), diachronisch und diaphasisch (d. h. im Hinblick auf Textsorten und Funktionalstile) untersucht werden. Neue Perspektiven öffnen Korpusanalysen auch für den interlingualen Vergleich, und zwar sowohl für kontrastive (*parole*-bezogene) als auch für typologische (*langue*-bezogene) Studien. Der facettenreiche Beitrag endet mit einem Vergleich von Darstellungen der Wortbildung mit *-haft* in DaF-Wörterbüchern.

In ihrem theoretisch und empirisch wohl untermauerten Beitrag *Phraseographie im Lichte sprachwissenschaftlicher Quellenkunde. Oder: Aus welchen Quellen kann ein zweisprachiges phraseologisches Lernerwörterbuch gespeist werden?* fokussiert **Hana Bergerová** auf die Auswahl von Kollokationen, eines Teilbereichs der Phraseologie i. w. S., für ein onomasiologisch-semasiologisch geordnetes phraseologisches Lernerwörterbuch mit dem Sprachenpaar Deutsch-Tschechisch. Da nur eine sehr begrenzte Auswahl von Einheiten aufgenommen werden könne, seien Repräsentativität, Frequenz und Geläufigkeit zentrale Kriterien. Über ein phraseologisches Optimum herrsche bisher jedoch keine Einigkeit. Die meisten phraseographischen Sekundärquellen sind für DaF-Lerner zu umfangreich und darüber hinaus stark idiomorientiert. Für Kollokationen sei Quasthoffs *Wörterbuch der Kollokationen im Deutschen* (2010) zwar ein gutes Hilfsmittel, aber bei 3.253 Grundwörtern und 192.000 Kollokationen sei der (potenzielle) Verfasser des Spezialwörterbuchs mit der „Qual der (Aus-) Wahl" (S. 90) konfrontiert. Da Kollokationen sprachspezifischen Kombinierbarkeitsrestriktionen unterliegen, sei die Auswahl in einem zweisprachigen Wörterbuch zudem sprachenpaarspezifisch, also auf kontrastiver Grundlage, zu treffen. Teiläquivalenz oder völlig fehlende Äquivalenz führe leicht zu einer Interferenz, sodass solche „ungleichen Paare" unbedingt Eingang finden sollen. Dies veranschaulicht Bergerová in einer onomasiologisch orientierten Fallstudie: Von den bei Quasthoff aufgenommenen 48 Basen des Emotionswortschatzes wird *Wut* als Beispielwort ausgewählt und in Verbindung mit einem adjektivischen Kollokator im *DeReKo* auf Kookkurenzen untersucht. Sodann werden die präferierten Kollokationen mit entsprechenden Kollokationen im *Tschechischen Nationalkorpus* und in dem tschechisch-deutschen elektronischen Wörterbuch *Lingea Lexikon 5* verglichen. Es ergab sich, dass insbesondere die deutschen Kollokationen *blanke* und *nackte Wut* aus tschechischer Sicht unerwartet sind. Laut der Verfasserin sollten aber neben unerwarteten Kombinationen auch hochfrequente deutsche Kollokationen mit totalen Äquivalenten aufgenommen werden.

Helge Goldhahn informiert über die *Grundlagen für das „Deutsch-tschechische Wörterbuch der Phraseologismen und festgeprägten Wendungen"*. Das im Jahr 2010 erschienene Wörterbuch ist mit 24.400 Stichwörtern eines der umfassendsten zweisprachigen phraseologischen Wörterbücher. Dank eines umfangreichen tschechischen Registers kann es auch in der umgekehrten Sprachrichtung konsultiert werden. Die Belege wurden hauptsächlich manuell gesammelt. Als Primärquellen dienten neuere

Belletristik sowie Presse- und Fachtexte, als Sekundärquellen phraseologische Wörterbücher aus allen deutschsprachigen Ländern. Die aufgefundenen Phraseologismen wurden von Muttersprachlern mit philologischer Ausbildung überprüft. Im Laufe der Zeit wurde für Recherche und Rückprüfung zunehmend auch das Internet und für Zweifelsfälle das IDS-Korpus benutzt; die Sekundärquellen wurden um das Internetlexikon *Redensarten-Index* erweitert. Das Ziel war eine möglichst umfassende Bestandsaufnahme deutscher Phraseologismen, wobei die Vorkommensfrequenz kein entscheidendes Kriterium war. Mündliche Primärquellen wurden nicht ausgewertet, aber laut dem Verfasser „[floss d]urch die Anpassung und teilweise vollständige Bildung der Kontextbeispiele durch Muttersprachler […] auch die gesprochene Sprache implizit in die Datenbankeinträge ein" (S. 101). Zum Schluss wird die Struktur der Lemmaartikel veranschaulicht: Sie bestehen „aus der Nennung des deutschen Phraseologismus, der Angabe zu seiner Stilebene, einer deutschen Definition, tschechischen Entsprechungen mit Stilebenenkennzeichnung und deutschen Kontextbeispielen" (ebd.). Gern hätte der Leser noch mehr erfahren, z. B.: Was war für die Festlegung der Nennform(varianten) des Phraseologismus entscheidend? Mithilfe welcher Primär- und Sekundärquellen wurden die tschechischen Äquivalente ermittelt?

Das Thema **Agnes Goldhahn**s lautet *Korpusgeleitete Lexikographie: „Das Häufigkeitswörterbuch der deutschen Gegenwartssprache"*. Heute seien korpusgeleitet erarbeitete Wörterbücher für Deutsch noch rar. Zu dieser Kategorie gehört das *Frequence Dictionary of German* von Jones und Tschirner (2006) mit den rund 4.000 häufigsten deutschen Wörtern. Die Lemmaliste basiert auf einer statistischen Auswertung des *Herder-BYU-Korpus*, das 4,2 Millionen laufende Wörter gesprochene Sprache, belletristische, journalistische und akademische Texte sowie Gebrauchstexte umfasst. In dem Artikel werden der Aufbau des Korpus und das korpusgeleitete Verfahren diskutiert. Im Folgenden steht das Letztgenannte im Mittelpunkt. Das automatische Tagging der Textwörter nach Wortarten wurde bei Bedarf manuell nachgebessert. Nach der Erstellung einer ersten Häufigkeitsliste wurden zum selben Lemma gehörende Wortformen zusammengefasst. Probleme bei der Wortartzuordnung bereiteten u. a. adjektivisch verwendete PII-Formen. Wenn diese mehr als 20% der verbalen Formen ausmachten und die Mindestfrequenz 16 pro Million erreichte, wurden sie als eigene Einträge kodifiziert. Im ersten Teil des Wörterbuchs sind die Wörter in ihrer Häufigkeitsreihenfolge aufgelistet und mit Rangangabe, Grundform, Wortart, Übersetzung ins Englische, Beispielsatz, Kennzeichnung der relativen Häufigkeit und ggf. Information zur Streubreite in den Subkorpora versehen. Zudem werden Mehrworteinheiten mit dem Grundwort oder dominante Wortformen angegeben. Zu speziellen Themen gibt es Infokästen. Der Häufigkeitsliste folgt eine alphabetische Liste, in der die Wörter mit Wortartangabe, Übersetzung ins Englische und Rangnummer versehen sind. Dazu gibt es Listen für verschiedene Wortarten, Abkürzungen und Eigennamen, die 100 häufigsten unregelmäßigen Verben und häufige Kollokationen. Im Schlusswort hebt die Verfasserin hervor, dass die Erstellung eines Häufigkeitswörterbuchs neben

quantitativen Methoden auch linguistisch fundierte qualitative Verfahren voraussetzt. Von Anwendungsmöglichkeiten der Häufigkeitslisten sei die DaF-Didaktik, speziell der auf dieser Grundlage verfasste neue *Grund- und Aufbauwortschatz Deutsch als Fremdsprache nach Themen* von Tschirner (2008) genannt.

Der **Gesprochene[n] Sprache** sind drei Beiträge gewidmet. **Ilka Mindt** beschäftigt sich mit *Gesprochene[n] Korpora des Englischen und ihre[r] Anwendung in der Grammatikforschung*. Damit in Datenbanken aufbewahrte gesprochensprachliche Korpora für die linguistische Analyse voll nutzbar gemacht werden können, soll die primäre Quelle anhörbar und von guter Aufnahmequalität sein. Transkripte stellen zwar eine große Hilfe dar, gelten aber als Sekundärquellen, die nur mit Vorbehalt als einziges Material zu nutzen seien. Weiter soll der ursprüngliche Kompilationszweck des Korpus beachtet werden, da sich daraus oft datenspezifische Restriktionen ergeben. Den Kern des Beitrags bildet eine Fallstudie zu Verbgefügen mit *come + to* + Infinitiv anhand von 27 anhörbaren, qualitativ ausreichend guten und von Muttersprachlern geäußerten Belegen im *Michigan Corpus of Academic Spoken English*. Mindt geht davon aus, dass sich Anzeichen grammatischer Funktionsänderungen in phonetischen Merkmalen bemerkbar machen. Auch wenn sie nicht alle ihre Hypothesen verifizieren kann, gelingt es ihr aufzuzeigen, dass sich die akustischen Parameter der final-additiven Gefüge mit *come* als Vollverb von denen der katenativen (den Beginn einer Situation bezeichnenden) Gefügen mit delexikalisiertem *come* deutlich unterscheiden. Man kann der Verfasserin zustimmen, dass weitere grammatische Studien anhand von primären gesprochenen Quellen ein Forschungsdesiderat sind, das Linguisten noch ein weites Arbeitsfeld eröffnet.

„Jede Datenbank zu Textkorpora […] ist ohne [Metadaten] nur eine amorphe Sammlung sprachlicher Daten." So leitet **Winfried Schütte** seinen Beitrag *Metadaten für Gesprächsdatenbanken: ein Überblick und ihre Verwaltung in der „IDS-Datenbank Gesprochenes Deutsch (DGD)"* ein. Zu diesen ‚Daten über Daten', die u. a. der Vorauswahl von Analysematerialien und der strukturierten Recherche dienen, zählen neben soziodemographischen Daten auch Angaben zum Ort und Medium der Speicherung, zu technischen Aufnahmeparametern, Transkription und sonstiger Datenaufbereitung sowie zu eventuellen Zusatzmaterialien wie Felddokumenten. Nach einer Besprechung konkurrierender Metadatenschemata geht Schütte auf die Erfassung und Verwaltung einer neuen Version der *Datenbank für Gesprochenes Deutsch* (*DGD 2.0*) ein, die unabhängig von spezifischen Forschungsansätzen sein und zwischen projektübergreifenden und -spezifischen Anforderungen vermitteln soll. Weitere wichtige Punkte sind detaillierte Datenstruktur, kalkulierte Redundanz, validierbare Datenerfassung, konsistente Datenspeicherung, benutzerfreundliche Darstellung mit Variationsmöglichkeiten, effektives korpusübergreifendes Retrieval sowie datenschutz- und -sicherheitsgerechte Benutzerverwaltung. Das generische Konzept des Dokumentationsmodells, neben dem projektspezifische Subschemata mit zusätzlichen fakultativen Komponenten erstellt werden können, besteht aus vier (XML-)Schemata: Ereignisdaten, allgemeine Sprecherdaten, Informationen zu Zusatzmaterialien

sowie Korpusbeschreibung. Diese werden detailliert vorgestellt und mithilfe von Diagrammen und Bildschirmfotos veranschaulicht. Eine wichtige Rolle spielt die Dokumentation der originalen Quellaufnahmen mit Angabe der Datenschutzvereinbarungen. Zum Schluss werden einige Suchoptionen und -operatoren demonstriert. Der Beitrag gibt einen guten Überblick über die Vielschichtigkeit einer systematischen Datenbankerstellung und kann jedem, der eine Datenbank plant, als Merkliste dienen.

In seinem mit *Tonband und Videokamera als Erkenntnisinstrumente zur Untersuchung mündlicher Kommunikation* betitelten Beitrag beleuchtet **Johannes Schwitalla** aus einer kultur- und wissenschaftsgeschichtlichen Perspektive die wissenschaftliche Nutzbarmachung von technischen Instrumenten, mit denen die von Natur aus flüchtige gesprochene Kommunikation festgehalten werden kann. In der Gesprächsforschung komme dem Tonbandgerät eine ähnliche Bedeutung zu wie dem Mikroskop in der Optik: Beide Geräte lassen Geordnetheit in Dingen erkennen, die früher als unwichtig und chaotisch galten, so etwa Abbrüche, Wiederholungen, Korrekturen oder ‚Flickwörter' in der gesprochenen Sprache. Heute ist das Bild von monologischen Sprechern, die fertig geplante Strukturen verlautbaren, vor dem von Dialogpartnern, die in einem kontinuierlichen interaktiven Prozess das Gespräch konstituieren, gewichen. Mit Tonträgern als primären Quellen und sog. reichen Transkripten als sekundären Quellen boten sich neue fruchtbare Forschungsobjekte wie funktionale Prosodie, Routinen der Kontaktaufnahme und -beendigung, Sprecherwechsel, Themaentwicklung und Reparaturen an, bei denen Musterhaftigkeit erkannt wurde. Als Beispiel führt Schwitalla Konstruktionsmuster der Wertung bzw. Modalisierung an. Eine zusätzliche Horizonterweiterung wurde durch visuelle Registrierung mit Kamera möglich, die dem Forscher auch nonverbale Signale zugänglich macht. Als Beispiel führt Schwitalla Reaktionen der Beteiligten an, wenn ein Sprecher ins Stocken gerät. Das visuelle Medium werde allerdings dadurch eingeschränkt, dass wegen des beschränkten Blickwinkels und subjektiver Entscheidungen bei der Kameraführung die Auslegung des komplexen kommunikativen Geschehens interpretationsabhängiger sei als bei Tonbandaufnahmen, auch wenn mehrere Kameras und die Split-Screen-Technik eine Abhilfe leisten.

Die Abteilung **Historische Sprachwissenschaft** umfasst drei Beiträge und wird von **Hans Ulrich Schmidt** mit dem Thema *Korpus und Korpuskel. Diachrone Onomasiologie am Beispiel von Modalverben* eingeleitet. Ziel ist nachzuweisen, dass auch ein Kleinkorpus (‚Korpuskel') relevante Ergebnisse liefern und bei onomasiologischer Fragestellung sogar sinnvoller sein kann als umfangreichere Korpora. Anhand von ausgewählten Texten – Bibelübersetzungen von Wulfila bis Luther, Bibeldichtung aus dem 9. Jh. und Bibelzitaten in Predigten aus dem 12. bis 14. Jahrhundert – werden, ausgehend vom Wortlaut der lateinischen *Vulgata*, alternative Ausdrücke für NOTWENDIGKEIT, ERLAUBNIS und FÄHIGKEIT diachron verfolgt. Neben den Modalverben sind in allen drei Funktionen unterschiedliche Verbgefüge im Gebrauch. Durch seinen onomasiologisch-funktionalen Ansatz kann der Autor Konkurrenzen und

Ablösungsprozesse demonstrieren und auch andere Ausdrucksweisen als Modalverben aufzeigen, die bei einer semasiologischen Modalverbanalyse nicht ins Blickfeld gekommen wären.

Überlegungen zum Erstellen von Korpora spätmittelalterlicher und frühneuzeitlicher Fachsprachen stellt **Lenka Vaňková** an. In den letzten Jahrzehnten haben Datenbanken für digitalisierte Handschriften und der flächendeckende Zugang zum Internet die Möglichkeiten zur Erforschung der historischen Sprachentwicklung enorm verbessert – bis zu einem Überangebot an Material, das zudem sehr heterogen sein kann. Deswegen sind gründliche Metadaten über zugängliche Handschriften und computerlesbare Korpora historischer Texte erforderlich. Bei der Erstellung solcher Korpora sei die sog. dynamische Edition mit minimal normalisierter Transkription zu empfehlen, die dann je nach dem Forschungszweck weiter bearbeitet werden könne. Besondere Herausforderungen stelle die Annotation, die bei großer graphematischer Varianz nicht automatisch erfolgen kann. In solchen Fällen empfehle sich die Herstellung von Listen von Textwörtern (Tokens). Dies veranschaulicht Vaňková anhand der mittels TUSTEP bearbeiteten Wortliste zum Olmützer medizinischen Korpus, die relevante dialektale Merkmale zu Tage förderte. Im letzten Abschnitt des Beitrags berichtet Vaňková vom Projekt zum Erstellen eines textsortenorientierten Korpus medizinischer handschriftlicher Texte des späten Mittelalters und der frühen Neuzeit, die in tschechischen Bibliotheken und Archiven lagern. Das Projekt ist hoch relevant: Fachtexte, insbesondere medizinische, waren weiter verbreitet als etwa belletristische Texte. Zudem zeigen sie viele Entwicklungstendenzen früher auf als andere Quellen, sind aber bisher nur wenig erforscht. Zum Schluss illustriert Vaňková den textsortengebundenen Zugang durch eine Sammlung von Monatsregeln aus dem 15. Jahrhundert und skizziert die weiteren Schritte zur Transkription der Texte und zu deren Umwandlung in eine maschinenlesbare Form.

Den Ausgangspunkt für **Vlastimil Brom**s Beitrag *Zur Quellenkunde in der Geschichtswissenschaft und Linguistik. Historiographische Werke als philologische Quellen* bildet ein Projekt zu deutschen spätmittelalterlichen historiographischen Texten aus den böhmischen Ländern. Quellenkunde und Quellenkritik spielten bei der Methodologie der historischen Wissenschaften schon immer eine zentrale Rolle, seien aber auch in der linguistisch-philologischen Forschung von Belang. Von der Quellenkunde wird erwartet, dass das verwendete Material katalogisiert, klassifiziert und dessen Quellenwert, etwa Entstehungsumstände und Objektivitätsgrad, eingeschätzt wird. Zur deutschen und europäischen Geschichte gibt es bereits einen Kanon quellenkundlicher Literatur, über die Brom einen Überblick gibt. Bei der sprachwissenschaftlichen Klassifikation von Quellen sei das Textsortenkonzept geeignet, bei der Abgrenzung historischer Sprachstufen sei aber die Kommunikationsgeschichte zu berücksichtigen, und in gewissen Bereichen mit starker illokutionärer Ausprägung, etwa in der Verwaltungs- und Rechtssprache, seien Sprechakttheorie und Soziopragmatik mit Gewinn einzusetzen. Der Beitrag ist ein gutes Beispiel dafür, wie verschiedene Wissenschaftsbereiche voneinander profitieren können.

Mit neun Beiträgen macht **Kleine und große Korpora – Spezialkorpora für Spezialfragen** die umfangreichste Abteilung des Bandes aus. **Gabriela Rykalová** geht es um *Kleine Korpora, große Korpora und Textsammlungen. Versuch einer korpustypologischen Zusammenschau*. In Hinblick auf die beiden gegensätzlichen Datengewinnungsarten, das in der ‚armchair linguistics' übliche introspektive Konstruieren von Daten und die empirische Korpuslinguistik, plädiert die Verfasserin für die letztere Alternative, weil nur authentische Belege in ihrem Gebrauchskontext die Sprachwirklichkeit widerspiegeln und die Forschungsergebnisse zur Revision von bisherigen Auffassungen führen können. In der tschechischen Germanistik habe sich insbesondere die korpus-gestützte Methode eingebürgert. Jede Korpusstudie sei aber mit der Frage konfrontiert, welches Korpus sich für welchen Forschungszweck eignet. Für große Wörterbücher und Grammatiken seien umfangreiche Korpora nötig, die den aktuellen Sprachgebrauch eines breiten Nutzerpublikums abdecken. Für kontrastive und übersetzungswissenschaftliche Studien eignen sich wiederum kleinere Parallelkorpora, die Originaltexte und ihre Übersetzungen – und dies bei einem Sprachenpaar in beiden Richtungen – enthalten, was u. a. beim *DeuCze*-Korpus der Fall ist. Weitere Beispiele für Bedingungen an Korpora gibt Rykalová aus den Bereichen der gesprochenen Sprache und der Sprachgeschichte. Für Studien zu speziellen Forschungsfragen reichen meistens kleine projektbezogene Korpora aus. Andererseits sei es nicht ungewöhnlich, dass bei speziellen Fragestellungen mehrere Korpora herangezogen werden. Dies exemplifiziert Rykalová mit einer Fallstudie zu adjektivisch gebrauchten partizipialen Komposita vom Typ *handgemalt*: Sie und ihre tschechischen Äquivalente wurden im *DeuCze*-Korpus, *COSMAS II* und *InterCorp* recherchiert, und schließlich konnte die Produktivität einzelner Wiedergabemöglichkeiten in einem großen tschechischen Korpus getestet werden.

In einer korpusbasierten Pilotstudie an einer kleinen Sammlung elizitierter Emails geht **Sven Staffeldt** der Forschungsfrage *Entschuldigungsmails – oder: Wie und wo findet man einen pragmatischen Standard?* nach. Mit der Fragestellung, ob Standard- und Nicht-Standard-Varianten von ansonsten vergleichbaren Texten nach gleichem Muster konstruiert werden, trägt er zur Tilgung einer Forschungslücke bei: Bisher sei die pragmatische Ebene von Nicht-Standard-Varietäten kaum untersucht worden. Staffeldts Korpuserstellung basiert auf der Annahme, dass sich Studierende in der Kommunikation mit einem Dozenten, den sie abgesehen von den Unterrichtskontexten nicht näher kennen, der Standardvarietät bedienen, während in der Kommunikation im Freundeskreis Nicht-Standard-Merkmale auftauchen. Dementsprechend formuliert er seine Email, in der er die zwei unterschiedlichen Probandengruppen – nicht näher bekannte Studierende und gute Freunde – um Entschuldigungsmails in einer vergleichbaren fiktiven Situation bittet, in der Standard- bzw. Nicht-Standardvarietät. In den analysierten 18 Entschuldigungsmails von Studierenden und den 14 Entschuldigungsmails von Freunden kommen Differenzen u. a. in der Textlänge (bei größerer Distanz kürzere Texte), im NP-Aufbau (komplexe NPs mit linksgestellten PII-Attributen nur in den Standard-Emails) und im Bedauerungsausdruck (schlichtes *leider* in

den Standard-Emails vs. emotionale Ausdrücke in den Nicht-Standard-Emails) vor. Die durchschnittliche Anzahl der Handlungen und die Haupthandlungsfolge sind in beiden Gruppen gleich; der Druck zur Versprechenshandlung (um den Anlass der Entschuldigung zu beseitigen), ist bei den Standard-Emails größer. Einige Teilergebnisse unterscheiden sich von denen früherer Email-Studien. Der Verfasser hebt hervor, dass seine Funde am ehesten Hypothesencharakter haben; ihre Signifikanz könne erst an größeren Datenmengen überprüft werden.

Jana Kusovás Artikel *Variation im Bereich der schwachen Substantive. Wege zur Korpuszusammenstellung und -auswertung* (S. 219–246) beginnt mit Überlegungen zu den Vor- und Nachteilen der als gegensätzlich geltenden Materialbeschaffungsmethoden der Linguistik, der *langue*-orientierten Introspektion und der *parole*-orientierten Empirie. Laut Kusová müssen diese Ansätze gegeneinander nicht ausschließen, denn neben der Beobachtung von empirischen Daten bleibe die Intuition des Linguisten u. a. bei der Aufstellung von Hypothesen zentral. Während die introspektive Methode beim Forscher ein natives Sprachgefühl voraussetzt, lassen empirische Korpusstudien es auch Nichtmuttersprachlern zu, quantitative und/oder qualitative Aussagen über den Gebrauch der untersuchten Sprache zu machen. Zur Datengewinnung für ihre gründliche und – trotz computergestützter Arbeit – aufwendige Studie über morphologische Schwankungen bei schwachen Substantiven im heutigen Standarddeutsch verwendet Kusová das *Langenscheidt Großwörterbuch Deutsch als Fremdsprache* (*LGDF*). Die dort kodifizierten 448 Substantive mit schwacher oder zwischen schwach und stark schwankender Deklination werden vorerst nach dem Wortausgang klassifiziert, um produktive Wortbildungsmuster ausfindig zu machen. Es stellt sich heraus, dass Maskulina mit den Suffixen *-ist* und *-e* und mit dem semantischen Merkmal ‚belebt' rund 45% der gelisteten Substantive ausmachen und als prototypische Fälle der schwachen Deklination mit einer geringen Tendenz zur Schwankung gelten können. Fälle, die dazu tendieren, von schwacher zu starker Flexion zu wechseln, werden nach Silbenzahl, Wortausgang und dem Merkmal ‚belebt/unbelebt' gruppiert. Sodann werden die einzelnen Formen des Deklinationsparadigmas der zu diesen Gruppen gehörenden sowie der im *LGDF* als Schwankungsfälle markierten Wörter mithilfe des *DeReKo* überprüft, die Belegzahlen in Tabellen erfasst und die Formvarianten mit Textbeispielen illustriert. Die Frequenzen und relativen Anteile der schwachen, starken und doppelt markierten Formen variieren von Lemma zu Lemma, sodass es letztendlich dem Forscher oder Lexikographen überlassen bleibe zu entscheiden, welche Frequenzen dazu berechtigen, von einer akzeptablen Deklinationsvariante zu sprechen.

Jana Valdrová thematisiert *Das unregelmäßige Verb und seine Bildungen im Definitionskorpus des Langenscheidt-Wörterbuchs*. Im Mittelpunkt steht die Frage, wie viele und welche der unregelmäßigen Verben DaF-Lernende auf dem Niveau B1–B2 des Europäischen Referenzrahmens für Sprachen beherrschen sollten; nach Schätzung von Lehrern wären es etwa 60 bis 80. Die Verfasserin führt eine Suche in dem 3.896 Wörter umfassenden Definitionswortschatz des *Langenscheidt-Taschenwörterbuchs*

(*LTWB*) und in zwei anderen kleineren Korpora durch. Als Ausgangspunkt für die Suche dient die Liste von 199 unregelmäßigen Verben in der *Duden-Grammatik* (2006). Die Suche ergibt, dass überraschenderweise ganze 133 (66%) der Verben der *Duden*-Liste im *LTWB*-Definitionswortschatz enthalten sind, also bedeutend mehr als die von den Lehrern geschätzte Zahl, und auch die beiden Vergleichskorpora zeigen ein ähnliches Resultat. Eine Stichprobe zu dem Anfangsbuchstaben *s-*, unter dem sich die meisten Grundwörter und Wortbildungen des deutschen Wortschatzes befinden, ergibt, dass im *LTWB*-Korpus 292 *s*-Wörter stehen und sogar 33 davon unregelmäßige Verben sind. Im Vergleich zu der *Duden*-Liste mit 61 Verben unter *s-* sind dies immerhin mehr als die Hälfte. Die meisten unregelmäßigen Verben gehören zum ältesten Wortschatzbestand und bezeichnen menschliche Grundtätigkeiten. Zudem werden zu ihnen viele Präfigierungen und weitere Wortbildungen gebildet. Fazit: „Die zentrale funktionale (semantische) Bedeutung der unregelmäßigen Verben liegt auf der Hand. Hinzu kommt die Bedeutung dieser Verben mit ihren Stammalternationen für den Ausbau des Wortschatzes durch Wortbildung." (S. 254). Das Resultat ist also durchaus relevant für den DaF-Unterricht.

Vít Dovalils Ziel ist, in seinem Beitrag *Zur Normativität als Problembereich der quantitativen und qualitativen Methodologie* ausgehend von den beiden im Titel erwähnten Zugängen „die sprachliche Normativität […] methodologisch adäquat zu verorten zu versuchen" (S. 259) und auf die Frage einzugehen, ob Belegbarkeit und hohe Vorkommensfrequenzen bestimmter (z. B. phonetischer, morphologischer, syntaktischer) Varianten als Beweis für Normgerechtheit betrachtet werden können. Der Verfasser beleuchtet zuerst die unterschiedliche Schwerpunktsetzung der quantitativen und qualitativen Korpuslinguistik: Der quantitative Zugang, der methodologisch den Naturwissenschaften nahe stehe, sei besonders für korpus-basierte Studien typisch; über Belegbarkeit und reine Häufigkeitszahlen der Varianten einer Variable hinaus interessieren die Distribution der Varianten und Korrelationen mit anderen inner- und außersprachlichen (etwa demographischen) Variablen. Im Unterschied dazu sei der korpusgeleitete Ansatz qualitativ orientiert; seine Explanationsgrundlage sei eher teleologisch als naturwissenschaftlich ausgeprägt. Als Beispiel für die qualitative Sprachforschung wird die der Ethnografie nahe stehende Sprachforschung angeführt: Im Mittelpunkt stehe das Funktionieren der Kommunikation, wobei der Interaktion nicht nur die Rolle als Spiegel sozialer Fakten, sondern auch als deren Mitgestalter zukommt. Normativität beim sprachlichen Verhalten komme mit ins Spiel durch die Frage, „was die Sprachbenutzer wissen müssen, um in konkreten Sprachgemeinschaften in angemessener Weise kommunizieren zu können" (S. 263). Dovalil erörtert das Wesen der Norm anhand der Auffassungen Klaus Gloys und Niklas Luhmanns: Normen seien das Ergebnis eines interpretierenden Schlussverfahrens und somit Bewusstseinsinhalte mit Regulationsfunktion; sie seien nicht an sich, sondern erst an ihrer Praxiswirkung empirisch beobachtbar. Empirisch feststellbare Regelmäßigkeiten seien aber kein hinlänglicher Beweis, sondern nur ein Hinweis auf *möglicherweise* existierende Normen. Nicht die Korpusbelegzahl an sich, sondern erst die diskursive

Praxis, in der die soziale Stellung und die Machtposition des Argumentierenden entscheidend sind, verleihe der Vorkommensfrequenz ihre normative Wirkung. Daher gelte: Die Normativität sei ein diskursives Konstrukt bzw. ein Prozess und sie sei primär qualitativ zu untersuchen.

Jiřina Malás Fallstudie *„Liebe auf den ersten Blick" oder „Wechselbad der Gefühle"? Phraseologismen in publizistischen (und literarischen) Texten korpusgestützt analysiert* geht auf ihr Forschungsvorhaben zur Emotionalität in Filmrezensionen zurück. Untersucht werden die Aufnahme und Darstellung der beiden im Titel enthaltenen, in Filmrezensionen recht üblichen Mehrwortverbindungen im *DUW* (2007) und in den idiomatischen Wörterbüchern *DUDEN 11* (2002) und Schemann (1993). Des Weiteren werden anhand des *DeReKo* und des *ZEIT*-Korpus des *DWDS* deren situations- und textsortenbedingter Gebrauch auch über Filmrezensionen hinaus erkundet sowie ihre Variierung und Modifizierung in verschiedenen Kontexten beschrieben. Schließlich wird ihren tschechischen Übersetzungsäquivalenten im (schönliterarischen) Parallelkorpus *InterCorp* nachgegangen. Der Beitrag ist ein gutes Beispiel dafür, wie zur Beantwortung spezieller Forschungsfragen schrittweise mehrere Korpora herangezogen werden und wie ein korpusgestützter Zugang die oft sehr knappen oder sogar fehlenden lexikographischen Informationen zu den Gebrauchsbedingungen von Phraseologismen ergänzen und vertiefen können.

Mit dem Emotionswortschatz beschäftigt sich auch **Eva Cieślarová**. In der ersten Hälfte ihres Beitrags *Korpuslinguistische Wege der Untersuchung von Emotionen im Deutschen und Tschechischen* geht es darum, wie im Projekt „Ausdrucksmittel der Emotionalität im deutsch-tschechischen Sprachvergleich" verschiedene Korpora zum Einsatz kommen, angefangen von einem (ggf. manuell ausgewerteten) einzelnen Werk über kleine zweckgebundene maschinenlesbare Korpora bis zu großen bzw. sehr großen digitalisierten Korpora, aus denen dann wieder zu bestimmten Forschungsfragen kleinere virtuelle Korpora maßgeschneidert werden können. Große Korpora eignen sich z. B. für Fragen der Belegbarkeit des Emotionswortschatzes. Komposita auf *Angst* beispielsweise können auf ihre semantische Struktur untersucht werden; ein soziolinguistischer Zugang könne wiederum Ängste in der Gesellschaft aufzeigen. Kleinere Korpora können zur Erforschung von Sprachvarietäten oder Textsorten und zu vergleichenden Studien dienen. Zu empfehlen seien auch „gestaffelte Korpora", d. h. eine schrittweise Kombinierung mehrerer Korpora (vgl. oben). Ein bewährtes Verfahren bestünde darin, von systematisch geordneten Daten, etwa Wörterbüchern oder Grammatiken (*langue*-Ebene), auszugehen und diese dann in einem authentischen Korpus (*parole*-Ebene) zu überprüfen, ob, wie häufig und in welchem Kontext die aufgelisteten Formen vorkommen. Durch Korpusrecherchen zu *brummen* und *zischeln* kann die Verfasserin Mängel in der Bedeutungsbeschreibung oder dem Äquivalentspektrum von Wörterbüchern aufzeigen. Für Phraseologismen bilden phraseologische Speziallexika einen guten Ausgangspunkt. Bei zwischensprachlichen Vergleichen können dann in einem zweiten Schritt ein- oder mehrsprachige Textkorpora ausgewertet werden. Zu deutsch-tschechischen Vergleichen eigne sich z. B. das

Teilkorpus *InterCorp* des tschechischen Nationalkorpus, das rund 12 Millionen deutsche Textwörter mit einem tschechischen Pendant enthält. Von den sehr großen einsprachigen Korpora seien das 1.300 Millionen Textwörter umfassende tschechische *SYN-Korpus* sowie die – vom Umfang her drei Mal so viel umfassenden – öffentlich zugänglichen Archive von *DeReKo* erwähnt. Die zweite Hälfte des Beitrags besteht aus einer Fallstudie über deutsche und tschechische Phraseologismen zu ‚Angst'. Für die ca. 70 diesbezüglichen deutschen Phraseologismen im *Synonymwörterbuch der deutschen Redensarten* von Schemann (1992) wurden im *InterCorp* Äquivalente gesucht. Auch wenn phraseologische Wiedergaben dominierten, wurde nicht immer zu phraseologischen Äquivalenten gegriffen, auch wenn solche existieren würden. Bei phraseologischer Wiedergabe stimmte der tschechische Ausdruck nur selten mit der in den tschechischen phraseologischen Wörterbüchern kodifizierten Nennform überein. Deshalb wurden die tschechischen Phraseologismen noch im *SYN-Korpus* auf ihre Variationen und Modifikationen untersucht. Es stellte sich heraus, dass einige phraseologische Varianten in den Wörterbüchern fehlen und dass nicht alles, was im Wörterbuch kodifiziert ist, tatsächlich im Gebrauch ist.

Am Beispiel seines Dissertationskorpus zur Erforschung der funktionalen Auslastung von *Grammatische[n] Mittel[n] der Informationskondensierung in Wirtschaftstexten* geht **Martin Mostýn** auf Kriterien bei der Erstellung von Kleinkorpora (mit maximal einer Million Tokens) ein. Große öffentlich zugängliche repräsentative Fachtextkorpora sind nach wie vor ein Desiderat; zurzeit existieren nur kleine Fachsprachenkorpora, z. B. zu einem Kommunikationsbereich oder zu einer Textsorte, die zudem i. d. R. nur forschungsgruppenintern zugänglich sind. Zweckgebundene Kleinkorpora seien typisch für kleinere Projekte und besonders dann nötig, wenn schon vorhandene große Korpora für den Forschungsgegenstand (etwa eine gewisse Sprachvarietät oder eine bestimmte Kommunikationsdomäne) nicht repräsentativ sind. Dies treffe z. B. für das IDS-Korpus zu, dessen Auswahl an Fachtexten bezüglich der Fachbereichs- und Textsortenpalette und der funktionalen Verteilung beschränkt ist. Mostýn zeigt, welche Rolle bei der Erstellung eines Spezialkorpus die genaue Definition der Forschungsfrage spielt: Nach ihr richten sich die Textauswahl, der Korpusumfang, die Annotierungsebenen und das korpuslinguistische Verfahren. Die Funktionen von Kondensierungsformen (satzwertige Infinitiv-, Partizip-, Nominal- und Präpositionalphrasen) seien am besten korpusgestützt erforschbar. Die nach textexternen und -internen Kriterien getroffene Textauswahl in der Studie gründet sich auf der Fachsprachengliederung: Horizontal geht es um den Fachbereich Wirtschaft, vertikal um drei Abstraktionsstufen, deren Einfluss auf die Kondensationsformen näher untersucht werden soll. Demnach wurden drei computerlesbare Teilkorpora (Theorie-, Handlungs- und Vermittlungssprache) mit einer Gesamtmenge von 480.000 Tokens erstellt. Es wird weiterhin berichtet, welche Softwares benutzt wurden, um die Primärdaten computerlesbar zu machen, die verschiedenen Kondensierungsmittel morphologisch, syntaktisch und semantisch (manuell) zu annotieren und ihre funktionale Auslastung zu analysieren. Zwischen den drei Teilkorpora konnten signifikante

Unterschiede festgestellt werden. Die Validität der Ergebnisse wurde noch an sehr großen Korpora (IDS-Archiv) überprüft, um herauszufinden, ob das Kleinkorpus zuverlässige Ergebnisse lieferte. Es konnte aufgezeigt werden, wie die Wahlmöglichkeit zwischen Infinitivphrasen und *dass*-Nebensätzen mit der Semantik des Matrixverbs und den sog. Subjekt-Kontrolle-Regeln zusammenhängt. Fazit des Verfassers ist, dass für spezielle Forschungsfragen ein Kleinkorpus ein ökonomischeres Mittel ist als ein großes Korpus, das sich wiederum eher für eine Gesamtdarstellung einer Sprache eignet.

Die Abteilung zu kleinen und großen Korpora endet mit **Thomas Schneider**s kritischem Blick auf die Anwendbarkeit digitalisierter Korpora in der Literaturwissenschaft. Gegenstand der Kritik in seinem Beitrag *Grundlosigkeit: Anmerkungen zum Problem der Quellen in der Literaturwissenschaft* ist die neohistorisch orientierte Projektskizze *KUWALU – Motivation und Grundzüge einer computergestützten Umgebung für die literatur- und kulturwissenschaftliche Recherche- und Analysearbeit* der Computerphilologen an der TU Darmstadt. Diese halten die textimmanente hermeneutische Interpretation des Textes als ‚quasisakrales Artefakt' für subjektiv und unwissenschaftlich. Im Gegensatz dazu könne der Text durch computerphilologische Mittel auf Intertexte geöffnet und in kulturelle und gesellschaftliche Wirkungszusammenhänge eingebettet werden, was eine kulturwissenschaftliche Neuorientierung der Literaturwissenschaft herbeiführen könne. Indem man bei der Betrachtung von Schnittstellen zwischen Texten das der Erschließung von Quellen zugrunde liegende Verfahren innerhalb eines Textkorpus digitalisiere, könne man es wissenschaftlich überprüfbar machen, was die Deskription und Interpretation inter(kon)textueller Dynamiken objektiviere. Laut Schneider wird hier allerdings der als religiös kritisierte hermeneutische Ansatz durch einen quasisakralen und quasinaturwissenschaftlichen Ansatz ersetzt, bei dem Textorientiertheit durch Kontextorientiertheit und qualitative Aspekte durch quantitative verdrängt würden. Mit einer Zunahme der Textquantität, die ja eigentlich die Objektivität der Erkenntnis garantieren solle, würden jedoch die Subjektivität und die Oberflächlichkeit der Analyse steigen, denn anstatt dass „die einzelnen Textereignisse sich gegenseitig in ihrem Sinn bestimmen, verliert sich dieser, indem der hermeneutische Horizont sich mit jedem neuen Textereignis nur weiter ins Unbestimmte verschiebt" (S. 323). Keine noch so große digitalisierte Textmenge könne die Geltung des hermeneutischen Zirkels außer Kraft setzen. Vielmehr verlangen schon die Konstituierung eines Archivs, die Festlegung der Kriterien für die Abgrenzung einer Epoche und die Auswahl der vom Archiv als Erkenntnisquellen bereitgestellten Textereignisse subjektive Entscheidungen. Um begründen zu können, welche Fakten die Parallelität beweisen können, bedürfe es einer „Vorverständigung über den jeweiligen Sinn der Vergleichsstellen" (S. 322) und ihrer sorgfältigen intra- und intertextuellen Interpretation. Dass Textstellen, an denen man der Sinn- oder Bedeutungsparallelität nachgeht, als bloßen Belegen noch keine Beweiskraft zukommt, exemplifiziert Schneider mit dem assoziationsreichen Wort *Grund*, das in Ernst Meisters Dichtung mehr als 80mal vorkommt, so auch in dem Gedicht „Der Grund

kann nicht reden“ (daher auch die Anspielung im Titel des Beitrags). – Schneiders Ausführungen richten sich nicht gegen die Aufstellung von digitalen Korpora literarischer Texte an sich, sondern gegen eine zu einseitige Auffassung von der Anwendung computerphilologischer Methoden. Trotz maschineller Verfahren darf man nicht aus dem Auge verlieren, dass literarische Texte ‚offene‘ Texte sind, die sich einer objektiven Interpretation entziehen – in Anlehnung an Peter Szondi: Der Subjektivität der Dichtung entspricht die Subjektivität der Deutung.

In dem ersten der drei Beiträge zu **Korpuslinguistik und/oder Datenbanklinguistik** behandelt **Wolf Peter Klein** das Thema *Datenbanklinguistik. Eine Weiterentwicklung der Korpuslinguistik?*. Es geht um die Frage, wie Sprachwissenschaftler angesichts der Vielfalt von Kommunikationsereignissen, der unmessbaren Gesamtmenge des Gesprochenen und Geschriebenen sowie des breiten Spektrums an unterschiedlichen sprachwissenschaftlichen Teilbereichen und Perspektiven „ihren empirisch übergroßen Gegenstand auf greifbare Maße zu reduzieren“ vermögen (S. 334). Neben Korpuslinguistik, die das Profil und die Zukunftsaussichten der Sprachwissenschaft revolutioniert hat, wird laut Klein die Datenbanklinguistik – ein neuer Terminus, den der Verfasser erst prägt und definiert – eine immer größere Rolle spielen. Die Datenbanklinguistik sei „eine spezifische, methodisch-technische Form der Sprachwissenschaft, bei der die Arbeit mit (Computer-)Datenbanken in methodischer, praktischer und wissenschaftstheoretischer Sicht einen hervorragenden Status einnimmt“; ihr Potential liege darin, „die Sprachwissenschaft in die Lage zu versetzen, mit ihren großen Datenmengen […] zurechtzukommen“ (S. 336). Dafür, wie die Sprachwissenschaft von der Datenbanklinguistik profitieren kann, bringt Klein drei Beispiele: 1) Eine Forschungsdatenbank ist eine Erweiterung traditioneller Forschungsbibliographien, die – wie etwa die IDS-Bibliographie zur deutschen Grammatik – den Zugriff auf Forschungsliteratur durch Suchfragen zu Titel, Person, Jahr, der untersuchten Sprache, Schlagwort und Objektwort erlaubt. Dank der Digitalisierung schon vorhandener gedruckter Forschungsliteratur und der elektronischen Herausgabe von Neuveröffentlichungen können mithilfe einer Forschungsdatenbank virtuelle Fachbibliotheken erstellt werden. 2) Bei der ständigen Zunahme von maschinellen Korpora aktueller und sprachgeschichtlicher Primärdaten wird es immer wichtiger, die Korpora mit systematischen Metadaten zu versehen sowie „für bestimmte, fest umrissene Forschungszwecke eigene Textdatenbanken zu erstellen“ (S. 338). 3) Bei metasprachlichen Forschungsthemen, etwa Spracheinstellungen und -bewertungen, verwendet man Fragebögen, Interviews, Sprachexperimente u. Ä. Für diese sollten eigene Datenbanken gegründet und datenbanktechnische Hilfstechniken entwickelt werden, sodass man mit wenig Aufwand z. B. die Sozialdaten und die metasprachlichen Aussagen auf Korrelationen untersuchen kann. Kleins Fazit ist, „dass Korpuslinguistik und Datenbanklinguistik gemeinsam das Erkenntnispotential der Sprachwissenschaft auf eine neue Stufe heben“ werden (S. 340). Als Anmerkung der Rezensentin kann noch hinzugefügt werden, dass dies desto eher Wirklichkeit wird, wenn auch noch Meta-Datenbanken, d. h. Datenbanken über Datenbanken, kreiert werden.

Dominik Banhold und **Claudia Blidschun** stellen *Die Datenbank „ZweiDat": Sprachliche Zweifelsfälle in historischer Perspektive* vor – ein Projekt, an dem seit 2011 an der Universität Würzburg gearbeitet wird. Das Ziel sei nicht, „die Zweifelsfälle der deutschen Gegenwartssparche zu erfassen, sondern […] die Texte, die diese sprachliche Problematik zum Inhalt haben, aufzubereiten" (S. 343). Durch die Analyse von Zweifelsfallsammlungen und Hinweisen in Grammatiken und Wörterbüchern könne man die zu einer bestimmten Zeit vorhandenen Zweifelsfälle linguistischen Ebenen zuordnen und Informationen zu ihrer Bewertung gewinnen. Als erste Texte wurden die Zweifelsfall-Klassiker von Theodor Matthias (Erstauflage 1892, 6. Auflage 1929) und Gustav Wustmann (Erstauflage 1890, 3. Auflage 1903) in Form von Bild- und Textdateien aufbereitet. Zu jedem Zweifelsfall werden in Tabellenform folgende Angaben gemacht: 1) sprachliche Systemebene (etwa Flexionsmorphologie), 2) Schlagwörter (beispielsweise Adjektiv, Komparation, Komposition), 3) Belege in der im Text genannten Form (etwa *besteingerichtetsten*); in Klammern werden die Wortart und das jeweilige Lemma angegeben (für das vorangehende Beispiel: Adj./ Lemma: *gut eingerichtet*), 4) Bezugsinstanz (Kennzeichnungen wie Person, Funktiolekt, diatopisch, diachron usw.), 5) Bewertung durch die Autoren. Darüber hinaus wird auf aktuelle Grammatiken hingewiesen, sofern sie das beschriebene Phänomen behandeln. Banhold und Blidschun geben drei Beispiele für Fragestellungen, die mit Hilfe von *ZweiDat* beantwortet werden können. Erstens kann man ermitteln, welche sprachlichen Systemebenen bzw. welche Sprachphänomene zu verschiedenen Zeitpunkten thematisiert werden. Zweitens kann man verfolgen, inwieweit dieselben Phänomene und Argumente rekurrieren (S. 349: „Schreiben die Autoren von Zweifelsfall-Literatur voneinander ab?"). Drittens bekommt man eine Antwort auf die Frage, welche Bezugsinstanzen zu verschiedenen Zeiten erwähnt werden, zum Beispiel: Sind Schriftsteller oder Politiker für den „guten Stil" tonangebend oder werden sie als warnende Beispiele zitiert? Werden die Zweifelsfälle aus einer diatopischen, diastratischen oder diaphatischen Perspektive betrachtet? Über die in dem Beitrag erwähnten Forschungsfragen hinaus wird *ZweiDat* auch für viele weitere Zwecke, z. B. die Erforschung der Sprachentwicklung, nützlich sein: Im Laufe der Zeit kann sich die Prioritätsreihenfolge der Varianten ändern, und in den allgemeinen Einstellungen gibt es Verschiebungen auf der Achse zwischen Purismus und Toleranz.

Um eine spezielle Textdatenbank geht es in dem Beitrag *Die Datenbank „Digitale Volltexte zur Geschichte der deutschen Fach- und Wissenschaftssprache". Eine bibliographische Sammlung digitalisierter deutscher Fachtexte vom Mittelalter bis zur frühen Neuzeit* von **Peter Stahl** und **Ralf Zimmermann**. Die Datenbank, in welche die frühere gleichnamige Linksammlung der Universität Erfurt integriert wurde, soll durch den Zugang auf digitalisierte deutschsprachige Texte aus dem 12. bis 17. Jahrhundert einen Überblick über die deutsche Wissenschaftssprache in einer Zeit geben, in der Lateinisch in der Wissenschaft dominierte. Stahl und Zimmermann stellen die wichtigsten Informationstabellen der Datenbank mit ihren Feldern vor und geben Beispiele dafür, welches Suchen die Tabellen ermöglichen. Die wichtigsten Tabellen

betreffen *Autoren*, von denen (neben dem obligatorischen Feld für die laufende Nummer des Datensatzes) Vor-, Nach- und eventueller Künstlername vermerkt werden, sowie *Werke* mit Angaben zum Titel, normalisierten Titel und zu Publikationsdaten, etwa zum Zeitpunkt des Erscheinens. Des Weiteren sind die Einträge nach 24 *Sachbereichen* sortiert. In der Tabelle *Datenbanken* wird u. a. der Anbieter des Digitalisats kenntlich gemacht und per Link auf den Online-Fundort der Quelle verwiesen. Die Tabelle *URL* enthält die Webadressen der Digitalisate. Die Online-Oberfläche der Datenbank steht frei zur Verfügung. Der Nutzer kann die Werke entweder komplett auflisten oder spezielle Suchen durchführen. Zum Konferenzzeitpunkt umfasste die Datenbank 408 Werke von 209 Autoren aus 32 verschiedenen Quellen. Am besten vertreten sind die Wissenschaftsbereiche Medizin, Astronomie und Astrologie. Die statistische Auswertung, etwa die Verteilung der Werke des jeweiligen Bereichs auf verschiedene Jahrhunderte, beleuchtet die Wissenschaftsgeschichte und die Praxisnähe. So ist es kein Wunder, dass gerade pharmazeutische Rezepte oder Informationen über Heilpflanzen in der deutschen Sprache schon früh gefragt waren, während man sich in wissenschaftlichen Auseinandersetzungen noch lange des Lateinischen bediente. Die kontinuierliche Ergänzung der Datenbank wird das Bild von der Geschichte der deutschsprachigen Fachkommunikation weiter verfeinern.

Die Abteilung **Die Probe auf's Exempel** besteht aus nur einem Beitrag, der den Band abrundet, indem er auf das im ersten Beitrag thematisierte *DeuCze*-Projekt zurückkommt. In **Norbert Richard Wolf**s Artikel *Text(e) lesen und (danach) Korpora analysieren: Grundlage einer verstehenden sprachwissenschaftlichen Textanalyse* stehen die Datenbasis und die Vorgehensweisen textlinguistischer Analysen im Mittelpunkt. Als Datenquelle der exemplarischen Analyse dient der im *DeuCze*-Korpus enthaltene Roman *Am kürzeren Ende der Sonnenallee* von Thomas Brussig (1999). Er wird in erster Linie aus einer raumlinguistischen Perspektive untersucht. Schon im Romantitel kommen Raumbezeichnungen vor, die dem Leser als Orientierungssignale für die Lokalisierung des Handlungsraums der fiktiven Figuren dienen. Neben diesen analysiert Wolf das Wort *Adresse* und die konträren Ausdrücke *(dort) drüben – hier* in ihren näheren Kontexten und im Hinblick auf die Gesamthandlung und den Handlungsraum des Romans. Auch areale Signale in der Figurensprache (Berliner Mundart) werden diskutiert. Die Beobachtungen veranschaulichen zugleich die von Wolf empfohlenen Analyseschritte, die auf eine fruchtbare Weise den korpusgeleiteten und der korpusbasierten Ansatz kombinieren: Zunächst soll eine diskursive Analyse des Ganztextes vorgenommen werden, denn jede interpretative Detailanalyse soll sich neben dem unmittelbaren Kontext auf den Ganztext stützen. Als wichtig erscheinende Phänomene sollen systematisch und frequenzorientiert im Ganztext überprüft werden. Auch Vergleiche mit anderen Texten seien nützlich, um idiolektale und individualstilistische Züge nicht zu generalisieren.

Summa summarum: Die Korpuslinguistik ist von der modernen Sprachwissenschaft nicht mehr wegzudenken. Der solide Band gibt ein vielseitiges Bild von ihren Möglichkeiten und Grenzen. Vor allem wird deutlich, dass jedes Korpus bzw.

jede Korpusstudie zweckorientiert ist und der Forscher demnach seine Korpuswahl und Vorgehensweise begründen muss. Die Beiträge zeugen von der Fruchtbarkeit der Zusammenarbeit von In- und Auslandsgermanisten, und was im Besonderen die Letzteren betrifft, von der Vitalität und dem hohen Niveau der tschechischen Germanistik. Wie der Vorgängerband *Kompendium Korpuslinguistik* kann der vorliegende Sammelband der empirischen Sprachforschung neue Anregungen geben, und seine Einbeziehung in den methodologischen Unterricht an Universitäten ist mehr als wünschenswert.

Irma Hyvärinen (Helsinki)