Originalveröffentlichung in: Lobin, Henning/Schneider, Roman/Witt, Andreas (Hrsg.): Digitale Infrastrukturen für die germanistische Forschung. - Berlin [u.a.]: de Gruyter, 2018. S. 199-218. (Germanistische Sprachwissenschaft um 2020, Bd. 6)

Ruxandra Cosma und Marc Kupietz

9 Von Schienen, Zügen und linguistischen Fragestellungen

Abstract: Das hier vorgeführte Schienenbild ist das in Anlehnung an Wittenburg (2009) als Erweiterungsinstrument gewählte Mittel in dem Versuch, Computertechnologie, linguistische Forschung und Vernetzung am Institut für Deutsche Sprache in deren rasch wachsenden Vielschichtigkeit zu beschreiben. Hier werden u. a. drei Blickwinkel, der des Technologie entwickelnden Wissenschaftlers, des entwickelnden Nutzers und des Nutzers von Informationstechnologie in der linguistischen Forschung vereint und um eine für den Sprachvergleich neue Dimension, die sprachspezifische Parameter von Analyseinstrumenten miteinander harmonisiert, erweitert.

Keywords: Empirie, Forschungswerkzeuge, Sprachkorpora, Vernetzung

1 Die Reise hin

In der Beobachtung sprachlicher Daten und Phänomene, deren Untersuchung, Beschreibung und Erklärung fährt man bereits seit Jahrzehnten auf linearen Transportbahnen, die sich immer mehr vernetzen. In kurzen Wegen zwischen einzelnen Knotenpunkten werden entfernteste Ziele miteinander verbunden. Die Fahrt erfolgt innerhalb der eigenen Sprache, überschreitet auch Grenzen, richtet sich sprachen- und netzvergleichend aus. Manchmal trifft man auf ein ähnlich ausgebautes Schienennetz, manchmal muss man sich die Reise anders überlegen, oder auch einen Gleisanschluss organisieren. Die Schienen zum *spurengebundenen Verkehr* (z. B. Ross 2001: 8) erlauben in der Linguistik mittlerweile staunenerregende Fahrgeschwindigkeit. Die komplementären Seiten des sich auf Forschung begebenden Nutzers und des Infrastruktur betreibenden, in Stand haltenden und selbst erstellenden Unternehmers führen in der germanistischen Linguistik-Forschung zusammen, anders als im Schienenverkehr,

Ruxandra Cosma, Universität Bukarest, Fakultät für Fremdsprachen, Abteilung für Germanische Sprachen, Str. Pitar Mos 7–13, RO-010451 Bucuresti, E-Mail: ruxandra.cosma@lls.unibuc.ro
Marc Kupietz, Institut für Deutsche Sprache, R5 6–13, D-68161 Mannheim, E-Mail: kupietz@ids-mannheim.de

Open Access. © 2018 Ruxandra Cosma und Marc Kupietz, publiziert von De Gruyter. Dieses Werk ist lizenziert unter der Creative Commons Attribution 4.0 Lizenz.
https://doi.org/10.1515/9783110538663-010

wo Transport zur Raumüberwindung und Schieneninfrastruktur in Eisenbahnunternehmen organisatorisch getrennt werden.[1]

Digitalisierung und Vernetzung verändern auf breiter Front die Welt, das Denken, die Wissenschaft. Digitalisierung und Vernetzung, sowie sich daraus ergebende Expansion und Diversifikation bestimmen auch das linguistische Arbeitsfeld. Im Kontext der in diesem Aufsatz gewählten Konzepte entfalten sich in dem linguistischen Bereich immer raffiniertere Fragen zu Datenerhebung, zu ihrer Verfügbarmachung, zum Umgang mit Daten, zu Datenmengen, zu Methoden und Interpretation, zur Nutzung von bestehenden technologischen Diensten, zu technischen Herausforderungen. Diese zunehmende Komplexität, im Sinne der Begriffsverfeinerung und trennung, zeigt, wie bedeutend Aufbau, Ausbau und Nutzung von Infrastruktur geworden sind. Eine die Linguistik begleitende Forschung, mit Hilfe derer schon lange bestehende linguistische Fragestellungen neu betrachtet und behandelt werden können, macht die Nutzung der Infrastruktur zum Bestandteil der linguistischen Wissenschaft und zur linguistischen Arbeit selbst. Aus der Nutzerperspektive sind Daten, Methoden, Technik und Dienste Teil der Forschungsinfrastruktur, wie sie aber auch Teil der eigenen wissenschaftlichen Auseinandersetzung mit Sprache sind.

Nach wie vor ist linguistische Forschung ohne institutionelle Infrastruktur möglich; der Sprachwissenschaftler kann selbst Daten sammeln, analysieren, interpretieren, Phänomene erklären. Voraussetzungen, Verfahrensweisen, Interessen, Blickrichtungen, Zielsetzungen, Grundbegriffe sind individuell begründet. Für die datenbasierte, empirische Linguistik sind aber die Ansprüche an die Validität und Nachvollziehbarkeit der Prinzipien der Theoriebildung sowie der Methoden und empirischen Verfahren und die Exaktheit ihrer Beschreibung, also letztlich an Wissenschaftlichkeit, erheblich gestiegen. Empirisch und theoretisch ausgerichtet, ist Linguistik heutzutage meist viel mehr als nur Erfahrungswissenschaft und theoretische Wissenschaft, die „nicht nur die sprachliche Materialbasis, sondern auch die Reflexion auf die die Materialbasis erfassenden Methoden zum Objekt ihrer Forschung erhebt" (van de Velde 1970: 7). Computertechnologie ergänzt und unterstützt sie nun, und diese ist aus der linguistischen Forschung in den auf authentische Daten angewiesenen Bereichen kaum noch wegzudenken, wie sie in Ergänzung dann auch auf quantitative Methoden der Datenanalyse, die deskriptive Statistik oder statistische Operationalisierung nebst qualitativen Untersuchungen immer mehr zurückgreifen muss. Diese methodischen Erweiterungen erweisen sich als notwendig, da die Erkenntnisse einer Korpusuntersuchung die Frage nach der Generalisierbarkeit

1 Vgl. Ross (2001: 8–9).

der Aussagen zum untersuchten Phänomen überstehen sowie eventuell anstehende Zweifel ausräumen müssen. Damit hängt die Wahl und Beschreibung des Sprachdaten-Ausschnitts zusammen. Und dies hängt seinerseits von der Größe des Korpus ab. Somit kommt man zurück auf das Thema der Methodenentwicklung und auf Wissenschaftlichkeitskriterien, auf Fragen nach der wissenschaftlichen Objektivität und der Repräsentativität, die nicht absolut beschrieben werden können. Die Technikdistanz, von Jakob (1991) für die Geisteswissenschaften noch als deutlich lang beschrieben, wobei Technik den Bereich der Kommunikation und Information miteinbezog, ist in der Linguistik allgemein, besonders aber in der Korpuslinguistik so viel kürzer geworden. So auch die nicht nur dadurch schrumpfende Distanz zu den *sciences*, zu den ‚harten' Wissenschaften (vgl. Busse 1989; Teubert & Belica 2014: 298). Denn Korpuslinguistik setzt in diesem Sinne viel mehr als nur Informationstechnologie ein, sie modelliert auch sprachliche Daten und Sachverhalte zu verschiedenen Verwendungszwecken (z. B. Jannidis, Kohle & Rehbein 2017: 13–14, 100–104).

Die Schienen-Analogie, die wir hier als Ausgangspunkt angenommen haben und zu eigenen Interessen ausbauen, geht auf Peter Wittenburg (2009: 7) zurück, der sie gerne zur Illustration des Potenzials der Vernetzung bestehender Forschungsressourcen verwendet hat. Für Wittenburg sind solche Schienen (und Signalanlagen) im Vergleich zu den komplexen Zügen, die in seiner Analogie Methoden und Werkzeuge darstellen, technisch betrachtet äußerst simpel und haben dennoch das Potenzial, die Produktivität von Sprachforschern deutlich zu verbessern.

Für den Sprachwissenschaftler und dessen Auseinandersetzung mit sprachlichen Phänomenen sind vor allem Reiseziele und der Weg hin von Bedeutung; diese mögen, je nach Vorgehensart, mit der Nutzung einer Infrastruktur, als Bestandteil ihrer Wissenschaft, einhergehen. Ist einem Reisenden das Reiseziel bekannt, so sucht dieser nach Reiserouten, gegebenenfalls nach Zugverbindungen, die zum Ziel führen sollen. Er vergewissert sich, ob das Ziel über Schienenverkehr erreichbar ist, informiert sich somit darüber, wie die bestehende Infrastruktur für eigene Ziele optimal genutzt werden kann. Auf der Reise zum Ziel fährt man durch die Landschaft, direkt oder über Verkehrsknoten, die ein Umsteigen möglich machen. Dadurch wird das bestehende Netzwerk noch intensiver genutzt. Es mag durchaus sein, dass sich der hierfür gewählte Vergleich und das In-Beziehung-Setzen (Hentschel 2010: 15–21) von Schienenverkehr und textbasierter Technologie zur Analyse von Sprache nicht gänzlich als mängelfrei erweist, dass im Kleinen noch Ungereimtheiten zu erkennen sind. Die Ähnlichkeitsbetrachtung ist hier aber breit angelegt und verfolgt keine disparaten Eigenschaften, sondern strukturelle Ähnlichkeiten und

Ähnlichkeiten zwischen Relationen.[2] In diesem Fall zwischen Objekt, Fragestellungen, Instrument und Methode.

Die bildliche Klammerung zum wissenschaftsergänzenden Unterbau bietet die Möglichkeit, Vernetzung sinnfällig zu machen. Nicht nur als Bild von Knoten und Kanten und vom kürzesten Weg, als Vernetzungstopologie, sondern auch im Sinne des Verbundes mit europäischen Projekten.

Am Institut für Deutsche Sprache (IDS) stehen Entwickler und Nutzer von Infrastruktur in engster Zusammenarbeit bzw. sind in der wissenschaftlichen Arbeit auf verschiedenste Weise vereint. Der vorliegende Aufsatz beschreibt einige dieser vielen Komponenten, die dem Aufbau und Nutzung von Sprachdaten und Analyseinstrumenten dienen. Heterogene Anforderungen an digitale Infrastrukturen im Dienste der Linguistik, wie die Frage nach der Zusammenbringung von Daten, Methoden, Diensten etc., um Forschungsaufgaben effektiver lösen zu können, oder die Frage, wie man Daten, Methoden etc. möglichst nachhaltig nutzbar machen kann, beschreiben einen Aufbruch, der am IDS schon lange Tradition hat.

Im Folgenden werden wir uns auf Datenerfassung, Modellierungsprozesse und Auswertungsverfahren zur geschriebenen Sprache am IDS beziehen. Die am IDS parallel verlaufenden Entwicklungen zu Korpora gesprochener Sprache werden z. B. in Stift & Schmidt (2014) beschrieben und in Kupietz & Schmidt (2015) zu denen der Korpora geschriebener Sprache in Beziehung gesetzt.

2 Digitalisierung

Die Dokumentation sprachlicher Phänomene am Korpus, in den ersten Monaten nach der Gründung des Instituts 1964 bereits als Aufgabenbereich des IDS umrissen, daraufhin Datenerhebung, d. h. Texterfassung und Kodierung zur grammatischen Analyse[3] oder zu lexikografischen Zwecken mittels datenverarbeitender Maschinen, haben sich im Wandel der Zeit und über die Zeit hinweg etabliert. Über die – selbst für die heutige Zeit über-ambitionierten – anfänglichen Erwartungen schreibt Engel (1968: 1), dass in der Zeit nirgends so viele Blütenträume zerstört worden seien wie auf dem Gebiet der maschinellen Sprachverarbeitung.

---

2 „Analogy denotes a resemblance not between things, but between the relations of things." (Hentschel 2010: 24).

3 Vorerst im Rahmen des IDS-Projekts zu Grundstrukturen der deutschen Sprache (siehe Teubert & Belica 2014: 301).

Die Anfänge wandten sich der Datenerfassung zur Erstellung des Mannheimer Korpus I und II, oder einer Datenbank zum Wortschatz von Ost und West als Zeitungskorpus zu. Zunächst durch Einsatz von Schreiblochern und Lochstreifen mittels weniger Programme zur Kodierung und Datenverarbeitung im Rahmen einer Kooperation mit Rechenzentren in Darmstadt oder in Bonn, Ende der 1960er Jahre bereits im Rahmen eines eigenen Rechenzentrums am Institut über Programme, die von wissenschaftlichen Mitarbeitern der linguistischen Datenverarbeitung erstellt wurden, ferner stark erweitert durch das 1998 gestartete Verbundprojekt Deutsches Referenzkorpus (DeReKo-I) (Teubert & Belica 2014: 308) wuchsen die Datenbanken über Zeit hinweg zu einer universellen Ur-Stichprobe des geschriebenen Gegenwartsdeutsch, dem Deutschen Referenzkorpus DeReKo (Kupietz et al. 2008). Die Programme zur Auswertung der Daten wurden ihrerseits zu immer komplexeren Analysesystemen, mit denen immer komplexere linguistische Ziele verfolgt werden konnten, entwickelt; hier seien Programme und Übergänge vorerst nur kurz erwähnt – REFER, COSMAS I, COSMAS II, KorAP. Das dynamische DeReKo (Kupietz & Keibel 2009a) bot von vornherein die Möglichkeit zur Erstellung eines eigenen (virtuellen) Korpus, das individuellen Forschungsinteressen und Untersuchungsintentionen entgegenkommt, indem sich für die speziellen Fragestellungen und zu untersuchende Grundgesamtheiten maßgeschneiderte und möglichst repräsentative Unterstichproben approximieren lassen. Dokumentiert werden diese Entwicklungen in ihren Anfängen und in ihrem zeitlichen Verlauf von Engel (1968, 2014), Zint (1968), Stickel (2007, 2014), Berens (2014), Hellmann (2014), Kupietz (2014), Teubert und Belica (2014) und vielen anderen. Den Übergang von der maschinellen zur elektronischen und digitalen Datenerfassung, wählt man differenzierende Bezeichnungen, unterstützten technische und institutionelle Rahmenbedingungen, die bereits sehr früh vorwegnahmen, dass die Zukunft einem Schienenverkehr und der Vernetzung, anfangs nur als Computernetz betrachtet, gehört. Anfänge der Vernetzungsprozesse zwischen den Ressourcen oder auch zwischen Personen, die Wandlung von Informationstechnologien zu Informations- und Kommunikationstechnologien sind in der Mitte der 1980er Jahre angesiedelt (Jannidis, Hubertus & Rehbein 2017: 9). Ein institutionalisierter Raum für die Grundlagenforschung zur linguistischen Methodik basierend auf Korpora, d. h. für einen eigenen Programmbereich Korpuslinguistik, wurde am IDS ab 2004 geboten (Kupietz 2014, Teubert & Belica 2014: 315). 2009 ging dann aus diesem, angestoßen durch das BMBF-geförderte Projekt *Aufbau eines Zentrums „Digitale Forschungsressourcen für die germanistische Sprachwissenschaft“* ein eigener Programmbereich Forschungsinfrastrukturen hervor, der zunächst vor allem Informationsangebote, Informationstechnik und Informationswissenschaft, Projekte „mit Schnittstellen-

charakter“ wie CLARIN-D, WissGrid, TextGrid etc. sowie juristische Expertise bündelte (Schonefeld & Witt 2014).

Jeder der Aspekte des medialen Transfers von gedruckten Texten in eine elektronische Fassung, zusammenhängend mit sich rasch und selbstständig entwickelnden Rekonstruktions- und Kodierungsfragen, Fragen zum Umgang mit *Big Data* mittels eines Analysesystems, bereitete eine nächste Phase vor: die der Vernetzung, nicht nur im Sinne des *global village*, das durch das immer noch verbesserbare Internet und somit durch den Online-Zugriff möglich wird, sondern im Sinne einer Sammlung von Ressourcen, die den Traum von einem vereinfachten gebündelten Zugriff auf diese Ressourcen erlaubt. Und hier wird wiederum die eingangs genannte Analogie zu den linearen Transportbahnen merklich: Durch die Landschaft der Forschungsdaten fährt man mit ICE-Zügen, die wie im eigentlichen Sinne spur- und schienengebundene Verkehrsmittel sind. In diesem Sinne seien hier die am IDS entwickelten und eingesetzten Auswertungssysteme COSMAS II und KorAP vorerst nur kurz genannt; deren Rolle wird im Folgenden auf einer allgemeineren Ebene diskutiert.

3 Wissenschaftliche Werkzeuge

Inzwischen steht schon längst fest, dass „Korpuslinguistik nicht auf eine Senkung der theoretischen, sondern auf eine Hebung der methodischen Ansprüche hinausläuft“ (Lehmann 2007: 27). Damit hängt jedoch die noch nicht geklärte Frage nach der Rolle der Infrastruktur zusammen, ob diese nur ein Instrument im Dienste der Linguistik darstellt, oder vielmehr ein elementarer Bestandteil der eigentlichen Forschung wie die Interpretation von Daten und Analyseergebnissen ist. Zur Beantwortung der Frage ist zunächst vorauszuschicken, dass die Unterscheidung zwischen Infrastruktur und Forschung sowie zwischen Infrastruktur und Wissenschaft meist vom jeweiligen Betrachtungswinkel und -abstand abhängt.[4] So kann der Large-Hadron-Collider (LHC) mit etwas Abstand sicherlich als Bestandteil der CERN-Infrastruktur oder besser, da diese zur Forschung verwendet wird, als Bestandteil der CERN-Forschungsinfrastruktur betrachtet werden. Auf der anderen Seite sind Teilchenbeschleuniger im Allgemeinen natürlich Gegenstand von Forschung und der LHC selbst auch Ergebnis von Forschung. Zwischen Forschung und Infrastruktur(-Bestandteil) muss also grundsätzlich kein Widerspruch bestehen.

4 Eine weit gefasste Definition des Begriffs Forschungsinfrastruktur findet sich in Wissenschaftsrat 2011: 17.

Interessanter ist vielleicht die Frage, inwieweit solche wissenschaftlichen Werkzeuge, die der Produktion oder der Auswertung von Forschungsdaten dienen, derselben fachwissenschaftlichen Forschung zuzurechnen sind. Zwar kann auch diese Frage einfach darauf zurückgeführt werden, wie eng man diese Kategorie fasst, jedoch hat es diesbezüglich in vielen Disziplinen eine rasante Entwicklung gegeben. Dies betrifft generell solche Disziplinen, in denen anhand von großen Mengen an Forschungsdaten statistische Modelle gewonnen werden können, die konventionelle Erklärungsmodelle zunehmend ergänzen oder gar ersetzen, also keineswegs nur die Teilchenphysik, sondern auch z. B. auch die Medizin und solche Geisteswissenschaften, die auf größere Datenmengen zurückgreifen können und deren Gegenstand zu komplex ist oder sich aus anderen Gründen (noch) einer *Theory of Everything* verschließen (siehe z. B. Lehmann 2007, Keibel & Kupietz 2009). In besonderem Maße ist dabei die Linguistik betroffen, bei deren Forschungsprimärdaten vom Typ „beobachtbare Sprache" es sich gewissermaßen um ein bewegliches Ziel handelt, das zumindest Aspekte eines dynamischen Artefakts aufweist, da es sich zwar eingeschränkt aber permanent und unvorhersagbar verändert (vgl. Keller 1994; Kupietz & Keibel 2009b). Hinzu kommt, dass man von diesen und anderen Forschungsdatentypen in der Linguistik nur sehr indirekt auf allgemeinere Prinzipien – seien sie sprachsystemischer, sprachpsychologischer oder anderer Art – schließen kann, um zu neuen Hypothesen und Erkenntnisgewinn zu gelangen. Wie ein solcher Schluss gelingen kann, welche Eigenschaften der Daten dabei relevant und welche theoretischen Konstrukte dabei hilfreich sind, ist daher selbst elementarer Gegenstand der sprachwissenschaftlichen Forschung. Dies manifestiert sich z. B. in der großen Zahl unterschiedlicher Grammatiktheorien und -formalismen in der Linguistik, die sich mit unterschiedlichen Blickwinkeln ihrem Gegenstand nähern, parallel verfolgt werden und deren Anzahl die in der Teilchenphysik bei weitem übersteigt. Einen Grund für die Annahme, dass dieser Theorienpluralismus sich durch die verstärkte Einbeziehung von Korpora und anderen Daten, vielleicht durch eine Art Objektivierung, zwangsläufig wesentlich reduzieren würde, gibt es kaum. Es ist lediglich zu erwarten und bereits zu beobachten, dass Theorien, die sich überhaupt nicht anhand von Daten falsifizieren lassen, es schwer haben werden. Die Bandbreite möglicher Ansätze – und entsprechend die Unsicherheit über den jeweils richtigen – beginnt auch noch in einer eng gefassten Korpuslinguistik bereits bei der Auswahl und Erhebung der Daten. Sie setzt sich fort z. B. bei der Kodierung der Daten bzw. der Kodierung der Rekonstruktion der ursprünglichen Sprachereignisse, über die verwendeten Analysemethoden (jeweils mit unterschiedlichem Skopus, unterschiedlichen Bewertungsmaßen und anderen Parametern) bis hin zur Darstellung der Such- oder Analyseergebnisse. Für die

Ergebnisse einer abschließenden Interpretation haben alle auf diesem Weg getroffenen Entscheidungen einen potenziell ausschlaggebenden Einfluss. Daten und Werkzeuge zu ihrer Aufbereitung, Analyse und Interpretation ausschließlich unter dem Aspekt einer nicht weiter aufschlüsselbaren und einer idealerweise austauschbaren Infrastruktur zu betrachten, birgt daher die Gefahr, die Relevanz eines ganzen Bereichs ausschlaggebender Entscheidungen zu unterschätzen. Dies würde nicht nur die Suche nach möglichen Fehlschlüssen unnötig einschränken, sondern generell den Suchraum für Wege zum Erkenntnisgewinn.

Die Beantwortung der Frage, ob man diesen Bereich innerhalb einer jeweiligen Fachwissenschaft verorten sollte, hängt davon ab, in welchem Maße ein Werkzeug für die jeweilige Fachwissenschaft konzipiert wurde. Zwar kann es sein, dass auch eher fachfremde Werkzeuge sich als nützlich erweisen (hier besteht eher das Problem nur scheinbare Nützlichkeit zu erkennen). Dass aber etwa speziell linguistische Werkzeuge nur zufällig nützlich für die linguistische Forschung sind, ist eher unwahrscheinlich. Vielmehr muss man davon ausgehen, dass die oben genannte Reihe von Entscheidungen jeweils mit dem Ziel getroffen wurden, linguistische Probleme lösen zu können. Hinzu kommt, dass durch den Einzug von Korpora in verschiedene Bereiche der Linguistik, in denen sie sehr unterschiedliche Rollen spielen können, derzeit viele noch wenig erforschte Stellen gibt, an denen quantitative auf qualitative Methoden aufeinandertreffen. Diese Stellen betreffen Analysewerkzeuge nicht nur hinsichtlich ihrer verwendeten statistischen Methoden, sondern z. B. auch hinsichtlich der Darstellung, d. h. Visualisierung der Ergebnisse dieser, da sie einen erheblichen Einfluss auf die aus ihnen abduzierbaren qualitativen Hypothesen haben.

Die Frage, ob eine Darstellung sinnvoll ist, lässt sich auch hier nicht allgemein mit ja oder nein beantworten. Vielmehr ist die Antwort typischerweise abhängig vom konkreten Forschungsziel. Sie ist also mehr als nur disziplinspezifisch. Ein weiterer Aspekt der Integration von quantitativen und qualitativen Methoden betrifft die Unterstützung von Arbeitsabläufen durch Analysewerkzeuge. Hier haben Analysewerkzeuge ein großes Potenzial, quantitative Methoden mit qualitativen Methoden zu engen Zyklen miteinander zu verzahnen und in explorativen Kontexten die Abduktion von Hypothesen stark zu beschleunigen (vgl. Kupietz & Schmidt 2015: 307; Kupietz et al. 2017a: 326 f.). An dieser Stelle können verlässliche wissenschaftliche Werkzeuge auch dazu beitragen, dass in der Linguistik auch einzelne Wissenschaftler weiterhin empirisch fundierte Forschung zumindest zu einem großen Teil eigenständig betreiben können. Eine zunehmende Spezialisierung und Arbeitsteilung, wie sie z. B. in der Psychologie (s. a. Haider 2015) schon lange gang und gäbe ist, ist zwar aufgrund des Wachstums des notwendigen Wissens unumgänglich, aber

zumindest begrenzt auf die Exploration von Daten und Hypothesen können Werkzeuge z. B. Experten zur Datenanalyse ersetzen.

Wittenburgs klare Unterscheidung zwischen Schienen und (ICE-)Zügen ist also nicht nur hilfreich, um das Potenzial einer genuin infrastrukturellen und technisch vergleichsweise simplen Vernetzung zu verdeutlichen, sondern auch, um mögliche Missverständnisse bezüglich der Rolle wissenschaftlicher Werkzeuge zu vermeiden.

4 Vernetzung

Bereits in den 1990er Jahren gab es zahlreiche paneuropäische Projekte zur Förderung von *human language technology* (HLT), zum Aufbau von Ressourcen und Technologien sowie zum Aufbau von Infrastrukturen, wie insbesondere das Projekt TELRI (*Trans-European Language Resources Infrastructure*) (siehe Teubert & Belica 2014). Zwar haben sich diese Initiativen positiv auf die Zusammenarbeit europäischer Sprachzentren und auf die Entwicklung der Korpuslinguistik auch außerhalb der Anglistik und über die Lexikologie hinaus ausgewirkt, aber letztlich „wenige Ressourcen von bleibendem Wert" (Teubert & Belica 2014: 306) und insbesondere wenig greifbare Strukturen oder Infrastrukturen von bleibendem Wert geschaffen. Sprachressourcen und Technologien wurden vielmehr außerhalb solcher koordinierender Verbundinitiativen und im Zuge zunehmender Diversifikation und Spezialisierung auseinander divergierender Subdisziplinen entwickelt.

Die Ausgangslage für die nächste große paneuropäische Infrastrukturinitiative CLARIN (*Common Language Resources and Technology Infrastructure*) war 2006 folglich, dass durchaus in großer Anzahl vorhandene Sprachressourcen und Technologien (Language Resources and Technology – LRT) zum weitaus größten Teil untereinander nicht interoperabel waren. Das heißt, es war nicht ohne weiteres möglich, ein Korpus vom Zentrum A mit einem Korpus von Zentrum B zu vereinen, beide zusammen mit einem Wortart-Tagger vom Zentrum C zu annotieren, um dann die neu entstandene Ressource mit einem Werkzeug vom Zentrum D zu analysieren. Um dies zu bewerkstelligen, mussten nach dem sogenannten „*download first, then process*"-Paradigma (Kemps-Snijders et al. 2008) zunächst alle Ressourcen und Werkzeuge ggf. lizenziert und heruntergeladen werden, um dann die Auszeichnungsformate der Ressourcen und die Ein- und Ausgabeformate der Werkzeuge aneinander anzupassen und einen größten gemeinsamen Nenner aller Lizenzbedingungen zu finden. Der damit verbundene Aufwand wäre allerdings so groß gewesen, dass er in der Regel unrealistisch und höchstens teilweise zu bewerkstelligen war, so

dass es für solche Projekte meist einfacher war, neue, spezialisierte Ressourcen aufzubauen und Werkzeuge zu entwickeln, die allerdings selbst wiederum typischerweise nicht in anderen Kontexten nachnutzbar waren.

Die Hoffnung von CLARIN war, dass sich durch eine Verbesserung der Interoperabilität bestehender und zukünftiger LRT ihre Wiederverwendbarkeit exponentiell steigern ließe und damit ein regelrechter Boost in den mit Sprachressourcen arbeitenden geistes- und sozialwissenschaftlichen Disziplinen ausgelöst würde. Um eine solche Interoperabilitätsverbesserung zu erreichen, hat CLARIN vor allem an den folgenden Teilen einer besseren und stabileren Vernetzung von Zentren, Ressourcen und Nutzern gearbeitet[5]:

1. Sozialer Teil
   a) Interoperabilität als Gegenstand von Kooperationen (über unmittelbare fachliche Ziele hinaus)
   b) Erweiterung der Kontakte von Diensteanbietern über ihre unmittelbaren Zielgruppen hinaus, durch so genannte Facharbeitsgruppen zu verschiedenen Themenbereichen und durch Kurationsprojekte
   c) Dissemination z. B. durch Sommerschulen und durch Help-Desks
2. Normativer Teil
   a) Standards und Best-Practices zu Auszeichnungs- und Kodierungsformaten
   b) Best-Practices z. B. zum Aufbau und zur Kuration von Korpora
   c) Schnittstellendefinitionen
   d) Muster-Lizenzvereinbarungen
3. Rechtlich/vertraglicher Teil
   a) z. B. Verträge zwischen Diensteanbietern
4. Informationstechnischer Teil
   a) Basisdienste wie PID-Service und AAI
   b) Muster-Applikationen

Hinsichtlich einer Basisinfrastruktur hat CLARIN nun zehn Jahre später in der Tat sehr gute Dienste geleistet. Für die allermeisten Anwendungen liegt zwar ein web-basiertes Paradigma, wie es noch in Kemps-Snijders et al. (2008) skizziert wird, bei dem Ressourcen und Tools per Mausklick virtuell kombiniert werden können, in weiter Ferne, aber es sind die Grundlagen dafür geschaffen, dass auch im Rahmen von Abschlussarbeiten und kleinen Projekten Daten und Methoden aufgebaut, genutzt und nachnutzbar gemacht werden können. Der Grund, warum die Vision eines Mausklick-Szenarios nicht erreicht wurde und

5 Eine detaillierte Darstellung der verschiedenen Infrastrukturkomponenten findet sich in Fiedler et al. 2014.

wahrscheinlich auch nicht erreicht werden kann, ist dabei einfach die bei der Entwicklung von Ressourcen und Technologien mit begrenzten Mitteln unabdingbare Verpflichtung zur Sparsamkeit. Dies hat zur Folge, dass Features, die für die primäre Anwendung nicht notwendig sind gegenüber Features, die für die primäre Anwendung wichtiger sind, zurückstehen müssen, so dass sich LRT typischerweise nur bedingt und nicht auf dem gleichen Niveau in nicht primär intendierten Anwendungskontexten wiederverwenden lassen. Im CLARIN-Kontext trifft dies in besonderem Maße zu, da der CLARIN-Skopus nicht an bestimmten Anwendungen orientiert ist, sondern am Datentyp Sprachressource im Allgemeinen, so dass das Anwendungsspektrum innerhalb der Sozial- und Geisteswissenschaften sehr breit ist und völlig unterschiedliche Forschungsgegenstände und -ziele umfasst, obwohl mit DARIAH[6] zeitgleich eine weitere, komplementäre Infrastruktur für die Geistes- und Kulturwissenschaft aufgebaut wurde. Dass CLARIN trotzdem viel erreicht hat, liegt vor allem daran, dass sich die Arbeiten auf solche basalen Infrastrukturschichten konzentriert haben, die zwar für den Anwender nicht unmittelbar sichtbar sind, aber tatsächlich praktisch immer benötigt werden, um Sprachressourcen für die eigene Forschung zu erschließen und die Ergebnisse für andere nachnutzbar zu machen. So hat CLARIN ein recht vollständiges infrastrukturelles Basisfundament geschaffen, auf dem ein breites Spektrum fachwissenschaftlich orientierter Projekte aufsetzen kann.

Dass CLARIN sich auf, aus Anwenderperspektive meist weit entfernte, basale Komponenten konzentriert hat, war auch richtig und unvermeidbar im Hinblick auf eine andere für Infrastrukturen unabdingbare Eigenschaft: ihre Nachhaltigkeit. Wenn nicht sichergestellt ist, dass Schienen noch in vielen Jahren benutzbar sind, ist es ein Fehler, in ihren Ausbau oder ihre langfristige Nutzung zu investieren. Diese Herausforderung betrifft in erster Linie die technischen Infrastrukturkomponenten, deren permanente Instandhaltung und Wartung dauerhaft gesichert sein muss. Die Frage was und wie viel gerade an technischer Infrastruktur entwickelt werden sollte – also wie viel eine Disziplin sich leisten kann und leisten sollte, ist demnach weniger von den initialen Entwicklungskosten abhängig als vielmehr von der Frage, wie viel dauerhaft finanzierbar ist. Auch wenn die Infrastruktur-Finanzierung in den Geisteswissenschaften noch nicht so stark bedarfsgekoppelt ist wie in den Naturwissenschaften, müssen die Kosten natürlich auch hier in angemessener Relation zur Größe der Anwenderschaft stehen, so dass auch aus diesem Grund einfache Basisdienste, angefangen bei der Vergabe persistenter Identifikatoren

6 https://de.dariah.eu/ (letzter Zugriff: 20.10. 2017).

bis hin vielleicht zu einer einfachen föderierten Suche (Stehouwer et al. 2012) im Fokus stehen müssen.

Durch eigene Initiativen wird der Vernetzungsgedanke hausintern weiterentwickelt, das IDS wächst durch Ausbau zur Mitte. Einige dieser Initiativen werden in den folgenden Abschnitten beschrieben.

5 EuReCo

Die Idee der EuReCo-Initiative (siehe Kupietz et al. 2017b), die 2013 im Kontext von CLARIN und im Kontext der CMLC-Workshops (Bański et al. 2013, 2015; Kupietz et al. 2014a) entstanden ist, besteht darin, ein virtuelles Europäisches Referenzkorpus zu schaffen, das ausschließlich auf bereits existierenden oder im Aufbau befindlichen Referenz- und Nationalkorpora beruht, und zu diesen einen gemeinsamen Zugang anzubieten. EuReCo verfolgt dabei im Wesentlichen zwei Ziele: Zum einen soll so mit Hilfe der gemeinsamen technischen Plattform sprachvergleichende Forschung mit dynamisch definierbaren, vergleichbaren Korpora der verschiedenen Sprachen ermöglicht werden. Zum anderen soll durch die Verwendung einer gemeinsamen Plattform eine neue bzw. erweiterte Infrastruktur geschaffen werden, die auf folgende Synergieeffekte abzielt:

1. Vermeidung von Redundanzen bei der Entwicklung von Korpusanalyse- und Rechercheplattformen: Funktionalitäten werden nur einmal entwickelt und in verschiedenen Kontexten genutzt.
2. Vermeidung von Redundanzen bei Aufbau und Kuration cross-linguistischer und vergleichbarer Korpora: Bestehende Korpora werden dynamisch nachgenutzt. Sie werden weiterhin von den jeweils verantwortlichen Institutionen erweitert und kuratiert und über EuReCo nur weiteren Verwendungsarten zugeführt.
3. Vermeidung redundanter Lizenzierung: Die einzelsprachlichen Korpora verbleiben an ihren Standorten und müssen daher für die neuen Verwendungsarten typischerweise nicht neu lizenziert werden.
4. Verbesserte Nachhaltigkeit insbesondere der technischen Infrastrukturelemente durch die Unterstützung mehrerer Institutionen.
5. Schnellerer methodischer Fortschritt in den einzelsprachlichen Philologien durch die gemeinsame Verwendung von Weiterentwicklung von Analysemethoden und Workflows.
6. Leichtere Identifikation zusätzlich möglicher Synergien durch engere Kooperation unter den National- und Referenzkorpusanbietern.

Das oben skizzierte Problem, dass Forschungsinfrastrukturen typischerweise entweder sehr basal sind oder nicht mächtig genug, um stark divergierenden Endnutzeranforderungen zu genügen, soll dabei gelöst werden, indem die technische Infrastruktur als Ergänzung und zusätzliche Alternative zur möglicherweise bereits bestehenden Infrastruktur verwendet wird. Darüber hinaus soll der technische Teil der Infrastruktur als Open-Source-Software mit Entwicklungszweigen realisiert sein, die, falls notwendig, voneinander abweichen können.

5.1 DRuKoLA

EuReCo existierte lange Zeit nur als Idee und als Teil verschiedener Förderanträge. Dies änderte sich Anfang 2016 mit dem Start des Projektes DRuKoLA (ursprünglich *Sprachvergleich korpustechnologisch – Deutsch-Rumänisch* genannt), einer Kooperation zwischen dem IDS, der Universität Bukarest und den Forschungsinstituten der Rumänischen Akademie in Bukarest und Iași.[7] In seinem Mittelpunkt stehen das Deutsche Referenzkorpus (DeReKo) (Kupietz et al. 2010) und das Referenzkorpus der Rumänischen Gegenwartssprache (*Reference Corpus of Contemporary Romanian Language*, CoRoLa) (Tufiș et al. 2015, 2016). DRuKoLA ist ein transdisziplinäres Projekt mit Zielen im Bereich der Korpuslinguistik, der Grammatik und der Forschungsinfrastrukturentwicklung, indem über die Entwicklung und Harmonisierung von Forschungsinfrastruktur auf Ziele in der Grammatik und in der Korpuslinguistik hingearbeitet wird. Aufgrund neu entwickelter Instrumente sollen probeweise explorative linguistische Studien durchgeführt werden Somit wird eine Grundlage für empirische und technische Forschung im Sprachvergleich (hier Deutsch-Rumänisch) geschaffen, die sich gleichwohl einem höheren Ziel, der Entwicklung einer EuReCo-Grundstruktur und einer Blaupause zur Integration weiterer Korpora bzw. weiterer Institutionen in die Infrastruktur, unterwirft.

Neben dem ersten Schritt zur Verwirklichung der Idee eines virtuellen Europäischen Referenzkorpus ist also der Sprachvergleich das zentrale Thema des DRuKoLA-Projekts. Während im Wechsel der Zeiten der Sprachvergleich in Verbindung zur Sprachtechnologie im Bereich automatischer Übersetzung, der Erstellung von zwei- oder mehrsprachigen elektronischen Wörterbüchern oder

7 DRuKoLA wird von der Alexander-von-Humboldt-Stiftung als Institutspartnerschaft zwischen dem IDS, der Universität Bukarest und der beiden Forschungsinstitute für Computertechnologie der rumänischen Akademie in Bukarest und in Iași (Mihai Drăgănescu Research Institute for Artificial Intelligence in Bukarest und Institute for Computer Science in Iași) gefördert.

von multilingualen und multifunktionalen Sprachdatenbanken (z. B. Rovere & Wotjak 1993: 2) angesiedelt war, verbindet das IDS über DRuKoLA dessen lange Tradition des Sprachvergleichs mit der Vorreiter-Rolle im Bereich der Gewinnung und Auswertung von Forschungsdaten zur deutschen Sprache.

Am Institut für Deutsche Sprache wurde dem Deutschen im Sprachvergleich schon immer ein besonderer Stellenwert zugemessen. In unterschiedlicher Weise war das Deutsche Gegenstand sprachvergleichender grammatischer Forschungsprojekte, sei es in Form von zweisprachig kontrastiven Grammatiken, über das den sprachlichen Nachbarraum des Deutschen untersuchende GDE-Projekt (Projekt *Grammatik des Deutschen im europäischen Vergleich*) mit der 2017 fertiggestellten *Grammatik des Deutschen im europäischen Vergleich. Das Nominal* (Gunkel et al. 2017), gegenwärtig fortgesetzt mit dem Thema *Verbgrammatik* (u. a. Trawiński 2016; Wöllstein 2016), über mehrsprachige Forschungsprojekte wie das *Eurogr@mm-Projekt* (z. B. Augustin & Fabricius-Hansen 2012; Dalmas, Fabricius-Hansen & Schwinn 2016), über die Erforschung komplexer Sätze im Deutschen, Portugiesischen und Italienischen (u. a. Blühdorn & Ravetto 2014) oder komplexer Argumentstrukturen im Deutschen und im Rumänischen (Cosma et al. 2014), um hier nur einige Projekte der jüngsten Zeit zu nennen. Dabei zeigt sich immer wieder, dass eine sprachübergreifende Perspektive auf die untersuchten Phänomene das Sichtfeld ergänzt, so wie Analyseansätze zu grammatischen Phänomenen in anderen Sprachen neue Einblicke für die Analyse entsprechender Phänomene in der eigenen Sprache bringen können. Um mit den Worten der Autoren des GDE-Projekts den wissenschaftlichen Ertrag des Sprachvergleichs zu beschreiben, werde dadurch ersichtlich, „wo eine Sprache eigene Wege geht und wo ihre Strukturen mit denen anderer Sprachen konvergieren" (Gunkel & Zifonun 2012, Klappentext).

Jedoch sind die sich immer komplexer gestaltenden anwendungsbezogenen Ziele sprachvergleichender Untersuchungen nicht zu übersehen. Am Beispiel von DRuKoLA als wegeröffnenden Versuch des anspruchsvollen hausinternen Zieles, das neue Analysesystem KorAP „für die Allgemeinheit zu öffnen" und „auf mehrere Schultern zu verteilen" (Kupietz 2014: 326), daher Ressourcen aus unterschiedlichen Sprachen unter einem System zwecks Weiterentwicklung zu bündeln, wird vielmehr gezeigt, dass im Sinne der gleichen, in diesem Aufsatz verwendeten Analogie zum Ziel des Sprachvergleichs, aber auch über den Sprachvergleich hinaus, aus einer technologischen Perspektive Gleisanschlüsse organisiert werden können, um Knotenpunkte grenzübergreifend miteinander zu verbinden. Im Falle des als Vergleichssprache gewählten Rumänischen fällt der Gleisanschluss mit der Erstellung des rumänischen Referenzkorpus CoRoLa zusammen, einem Projekt, mit dem ein Jahr vor dem Start des DruKoLA-Projekts begonnen wurde und das ein Jahr vor diesem enden

soll, daher optimale Bedingungen für Harmonisierungsverfahren zur Daten- und Prozessmodellierung bietet. Zum anderen aber bietet die Möglichkeit, mit demselben Analysesystem zwei Sprachen z. B. syntaktisch anhand virtuell definierbarer Korpora zu untersuchen, ein herausragendes Beispiel für die Art, in der jeweils einzelsprachliche Untersuchungen zum selben Phänomen über die gleiche Ausgangsbasis und dieselben technologischen Möglichkeiten im Sprachvergleich sprachtechnologische und infrastrukturelle Entwicklungen begünstigen können, die ihrerseits wiederum u. a. zur Weiterentwicklung der Linguistik im engeren Sinne beitragen. Als nächste Sprache soll das Ungarische im Rahmen des Projekts *DeutUng*[8] in EuReCo integriert werden.

5.2 KorAP

Die technische Basis des DRuKoLA-Projekts bildet die Korpusanalyseplattform KorAP (Kupietz & Frick 2013; Bański et al. 2013; Diewald et al. 2016), die als Nachfolgesystem zum Korpusrecherche- und Analysesystem COSMAS II (Bodmer 2005; Bodmer Mory 2014) entwickelt wird. KorAP ist zwar ein in erster Linie wissenschaftliches Werkzeug und bietet in dieser Hinsicht einige Neuerungen (Kupietz et al. 2017a), es bietet aber auch unter dem infrastrukturellen Aspekt Designeigenschaften, die für die Verwirklichung der DRuKoLA-Projektziele und der EuReCo-Visionen unabdingbar sind. Die in dieser Hinsicht wichtigste Eigenschaft ist, dass KorAP es erlaubt, mit Daten zu arbeiten, die physikalisch auf verschiedene Standorte verteilt sind. Dies setzt nicht nur das EuReCo-Konzept um, dass jedes beteiligte Zentrum weiterhin für die Kuration seiner Korpora verantwortlich ist, sondern löst auch das in der Linguistik omnipräsente rechtlich-ökonomische Problem, dass Korpustexte normalerweise nur so weit lizenziert werden können, dass sie zwar auszugsweise, aber nicht vollständig die Standorte der unmittelbaren Lizenznehmer verlassen dürfen (Kupietz et al. 2014b: 4; Schonefeld & Witt 2014: 332). Ein weiteres damit eng verbundenes Problem ist, dass die (Nach-)Nutzbarkeit von Daten und Recherchesoftware in Zusammenhang mit extern entwickelten Analysemethoden typischerweise sehr eingeschränkt ist. KorAPs Lösungsansatz folgt hier Jim Grays (2003) Postulat *put the computation near the data.* D. h., wenn die Daten aus rechtlichen oder informatischen Gründen nicht bewegt werden können, müssen Möglichkeiten geschaffen werden, dass vielmehr die Methoden zu den Daten bewegt und auf diese angewendet werden können. Nach der ursprüng-

8 Langtitel: Deutsch-ungarischer Sprachvergleich: korpustechnologisch, funktional-semantisch und sprachdidaktisch – ebenfalls als Institutspartnerschaft (zwischen der Universität Szeged und dem IDS) gefördert von der Alexander von Humboldt-Stiftung.

lichen KorAP-Konzeption sollte dieses Prinzip durch eine spezielle, gesicherte Umgebung für extern entwickelte Methoden (*mobile code sandbox)* realisiert werden (vgl. Kupietz 2014: 325). Dieser Ansatz wurde jedoch zugunsten einer besseren langfristigen Wartbarkeit durch einen einfacheren, auf Standardtechniken der Open-Source-Softwareentwicklung beruhenden Ansatz ersetzt. So können neue Algorithmen an den bestehenden KorAP-Schnittstellen ansetzen, die z. B. auch von KorAPs Weboberfläche verwendet werden. Falls dies nicht möglich ist, können die neuen Methoden aber auch als KorAP-Erweiterungen realisiert werden, die entweder in eigenen Entwicklungszweigen gepflegt werden oder, idealerweise, in den Hauptentwicklungszweig von KorAP eingehen. Bei aller gebotenen Skepsis bezüglich der Nachnutzbarkeit von spezialisierten Forschungswerkzeugen in neuen Kontexten, legt KorAP so zumindest eine auch aufgrund des institutionellen Engagements des IDS realistische infrastrukturelle Basis für weitergehende externe Entwicklungen. Inwieweit diese genutzt wird und so vielleicht auch zur Bündelung von Ressourcen und zur Kanonisierung von Methoden sowie damit vielleicht auch zur Beantwortung der Fragen, wie viel Infrastruktur sich die Linguistik leisten kann und leisten sollte, beitragen kann, bleibt allerdings abzuwarten.

6 Rundblick

Im Unterschied zum wirklichen Schienenverkehr, in dem Streckenausbau mit Fahrplan-Umstellungen einhergeht[9], gibt es in der hier beschriebenen Entwicklung von Forschungsinfrastruktur in der linguistischen Forschung am IDS trotz Verkürzung der Fahrten über Schienenwege und Höchstgeschwindigkeitszüge keinen Zeitzwang. Dafür gibt es aber immer mehr neue Schienen, Gleise, Anschlüsse, Knotenpunkte, Weichen und Übergänge, die von dem Stellwerk am IDS entwickelt, durchgeführt, angeboten werden. Das Netzwerk wächst in dem inneren Bau weiter, öffnet sich nach außen, schließt sich an. Der Weg hin verläuft aber in eigenem Tempo.

Der vorliegende Text berichtet letztendlich von Erfahrungen, von Ergebnissen, von Träumen und Neuerungen. Wir übernehmen hier den Werbeslogan des neuen Porsche-Panamera: *Ideen alleine verändern nichts. Sondern der Mut, sie umzusetzen.*[10] Hiermit sei für den Mut zum Wandel gedankt.

---

9 Z. B. https://web.de/magazine/reise/streckenausbau-deutsche-bahn-plant-grosse-fahrplan-aenderungen-32381716 (letzter Zugriff: 20. 10. 2017).

10 http://www.porsche-club-deutschland.de/PcLife/16-2/PC-Life/PDFs/032_Porsche.pdf (letzter Zugriff: 20. 10. 2017).

Literatur

Augustin, Hagen & Cathrine Fabricius-Hansen (Hrsg.) (2012): *Flexionsmorphologie des Deutschen aus kontrastiver Sicht*. Tübingen: Groos.

Bański, Piotr, Joachim Bingel, Nils Diewald, Elena Frick, Michael Hanl, Marc Kupietz, Piotr Pęzik, Carsten Schnober & Andreas Witt (2013): KorAP: The new corpus analysis platform at IDS Mannheim. In: Vetulani, Zygmunt & Hans Uszkoreit (Hrsg.): *Human language technologies as a challenge for computer science and linguistics*. Proceedings of the 6th Language and Technology Conference, 586–587. Poznan: Uniwersytet im. Adama Mickiewicza w Poznaniu. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-32617 (letzter Zugriff: 20. 10. 2017).

Bański, Piotr, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen, Andreas Witt (Hrsg.) (2015): *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, Lancaster, 20 July 2015. Mannheim: Institut für Deutsche Sprache.

Berens, Franz Josef (2014): Zur Frühgeschichte des IDS. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 52–60. Mannheim: Institut für Deutsche Sprache.

Blühdorn, Hardarik & Miriam Ravetto (2014): Satzstruktur und adverbiale Subordination. Eine Studie zum Deutschen und zum Italienischen. In *Linguistik online* 67, 3–44.

Busse, Dietrich (1989): Sprachwissenschaftliche Terminologie Verständlichkeits- und Vermittlungsprobleme der linguistischen Fachsprache. In *Muttersprache* Jg. 99, 27–38.

Bodmer, Franck (2005): COSMAS II. Recherchieren in den Korpora des IDS. In *Sprachreport* 3, 2–5.

Bodmer Mory, Franck (2014): Mit COSMAS II „in den Weiten der IDS-Korpora unterwegs“. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 376–385. Mannheim: Institut für Deutsche Sprache.

Cosma, Ruxandra, Stefan Engelberg, Susan Schlotthauer, Speranța Stănescu & Gisela Zifonun (Hrsg.) (2014): *Komplexe Argumentstrukturen. Kontrastive Untersuchungen zum Deutschen, Rumänischen und Englischen*. Berlin, München, Boston: de Gruyter. (Konvergenz und Divergenz 3).

Dalmas, Martine, Cathrine Fabricius-Hansen & Horst Schwinn (Hrsg.) (2016): *Variation im europäischen Kontrast. Untersuchungen zum Satzanfang im Deutschen, Französischen, Norwegischen, Polnischen und Ungarischen*. Berlin, Boston: de Gruyter. (Konvergenz und Divergenz 5).

Diewald, Nils, Michael Hanl, Eliza Margaretha, Joachim Bingel, Marc Kupietz, Piotr Bański & Andreas Witt (2016): KorAP architecture – Diving in the deep sea of corpus data. In Nicoletta Calzolari et al. (Hrsg.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3586–3591. Portorož, Slovenia, Paris: European Language Resources Association (ELRA).

Engel, Ulrich (1968): Vorbemerkung zum Forschungsbericht des Instituts für deutsche Sprache. In Hugo Moser, Peter von Polenz, Hans Glinz & Paul Grebe (Hrsg.): *Forschungsberichte des Instituts für deutsche Sprache*. Mannheim: Institut für deutsche Sprache.

Engel, Ulrich (2014): Das Institut für deutsche Sprache 1965–1976. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 64–79. Mannheim: Institut für Deutsche Sprache.

Fiedler, Norman, Antonina Werthmann, Maik Stührenberg, Oliver Schonefeld, Joachim Bingel, Andreas Witt (2014): *Forschungsinfrastrukturen in außeruniversitären Forschungseinrichtungen*. Forschungsbericht. Mannheim: Institut für Deutsche Sprache.

Gray, Jim (2003): *Distributed computing economics*. Technical Report MSR-TR-2003–24, Microsoft Research. https://www.microsoft.com/en-us/research/publication/distributed-computing-economics/ (letzter Zugriff: 24. 11. 2017).

Gunkel, Lutz, Adriano Murelli, Susan Schlotthauer, Bernd Wiese & Gisela Zifonun (2017): *Grammatik des Deutschen im Europäischen Vergleich. Das Nominal*. Berlin, Boston: de Gruyter. (Schriften des Instituts für Deutsche Sprache 14).

Gunkel, Lutz & Gisela Zifonun (Hrsg.) (2012): *Deutsch im Sprachvergleich. Grammatische Kontraste und Konvergenzen*. Institut für Deutsche Sprache. Jahrbuch 2011. Berlin, Boston: de Gruyter.

Haider, Hubert (2015): *Zwischen Sortieren und Prädizieren – Grammatikforschung & Grammatikbeschreibung auf dem Sprung vom 19. in das 21. Jahrhundert*. Vortrag im Rahmen der Tagung *ars grammatica*. Institut für Deutsche Sprache, Mannheim. Abstract: http://arsgrammatica.ids-mannheim.de/abstracts/Abstract_Haider.pdf (letzter Zugriff: 20. 10. 2017).

Hellmann, Manfred W. (2014): Die Bonner „Forschungsstelle für öffentlichen Sprachgebrauch“ (F.ö.S.) 1964–1980. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache. Mannheim*. 80–98. Mannheim: Institut für Deutsche Sprache.

Hentschel, Klaus (2010): Die Funktion von Analogien in den Naturwissenschaften. In *Acta Historica Leopoldina* Nr. 56, 13–66.

Jakob, Karlheinz (1991): *Maschine, mentales Modell, Metapher: Studien zur Semantik und Geschichte der Technikspracher*. Tübingen: Niemeyer.

Jannidis, Fotis, Hubertus Kohle & Malte Rehbein (Hrsg.) (2017): *Digital Humanities. Eine Einführung*. Stuttgart: J. B. Metzler.

Keibel, Holger & Marc Kupietz (2009): Approaching grammar: Towards an empirical linguistic research programme. In Makoto Minegishi & Yuji Kawaguchi (Hrsg.): *Working Papers in Corpus-based Linguistics and Language Education*, No. 3, 61–76. Tokyo: Tokyo University of Foreign Studies.

Kemps-Snijders, Marc, A. Klassmann, C. Zinn, P. Berck, A. Russel & Peter Wittenburg (2008): Exploring and enriching a language resource archive via the web. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 807–810. Paris: European Language Resources Association (ELRA).

Kupietz, Marc & Holger Keibel (2009a): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In Makoto Minegishi & Yuji Kawaguchi (Hrsg.), *Working Papers in Corpus-based Linguistics and Language Education*, No. 3, 53–59. Tokyo: Tokyo University of Foreign Studies.

Kupietz, Marc & Holger Keibel (2009b): Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit. In Marek Konopka & Bruno Strecker (Hrsg.), *Deutsche Grammatik – Regeln, Normen, Sprachgebrauch*, 33–50. Berlin, New York: de Gruyter. (= Jahrbuch des Instituts für deutsche Sprache 2008).

Kupietz, Marc & Elena Frick (2013): Korpusanalyseplattform der nächsten Generation. In Iva Kratochvílová & Norbert Richard Wolf (Hrsg.), *Grundlagen einer sprachwissenschaftlichen Quellenkunde*, 27–36. Tübingen: Narr, 2013. (= Studien zur Deutschen Sprache 66).

Kupietz, Marc (2014): Der Programmbereich Korpuslinguistik am IDS. In: *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim. 320–328.

Kupietz, Marc, Hanno Biber, Harald Lüngen, Piotr Bański, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt & Jani Takhsha (Hrsg.) (2014a): *Proceedings of the LREC 2014 Workshop Challenges in the Management of Large Corpora (CMLC-2)*. Paris: European Language Resources Association (ELRA).

Kupietz, Marc, Harald Lüngen, Piotr Bański, Cyril Belica (2014b): Maximizing the potential of very large corpora. In Marc Kupietz et al. (Hrsg.), *Proceedings of the LREC-2014-Workshop Challenges in the Management of Large Corpora (CMLC-2)*, 1–6. Paris: European Language Resources Association (ELRA).

Kupietz, Marc & Thomas Schmidt (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In Ludwig M. Eichinger (Hrsg.), *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven*, 297–322. Berlin/Boston: de Gruyter. (Jahrbuch des Instituts für Deutsche Sprache 2014).

Kupietz, Marc, Nils Diewald, Michael Hanl, Eliza Margaretha (2017a): Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In Marek Konopka & Angelika Wöllstein (Hrsg.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*, 319–329. Berlin, Boston: de Gruyter. (Jahrbuch des Instituts für Deutsche Sprache 2016).

Kupietz, Marc, Andreas Witt, Piotr Bański, Dan Tufiş, Dan Cristea & Tamás Váradi (2017b): EuReCo – Joining forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In *Proceedings of the 5th Workshop on Challenges in the Management of Large Corpora (CMLC-5)*, 15–19. Mannheim: Institut für Deutsche Sprache.

Lehmann, Christian (2007): Daten – Korpora – Dokumentation. In Werner Kallmeyer & Gisela Zifonun (Hrsg.), *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. Berlin, New York: de Gruyter (Jahrbuch des Instituts für Deutsche Sprache 2006) http://www.christianlehmann.eu/publ/daten.pdf (letzter Zugriff: 20. 10. 2017).

Ross, Sebastian (2001): *Strategische Infrastrukturplanung im Schienenverkehr. Entwicklung eines Planungs- und Entscheidungsmodells für die deutsche Bahn AG*. Diss. Deutscher Universitätsverlag. Wiesbaden: Gabler.

Rovere, Giovanni & Gerd Wotjak (Hrsg.) (1993): *Studien zum romanisch-deutschen Sprachvergleich*. Tübingen: Niemeyer. (Linguistische Arbeiten 297).

Schonefeld, Oliver & Andreas Witt (2014): Forschungsinfrastrukturen am IDS: Gegenwart und Zukunft. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 329–336. Mannheim: Institut für Deutsche Sprache.

Stehouwer, Herman, Matej Durco, Eric Auer & Daan Broeder (2012): Federated search: Towards a common search infrastructure. In N. Calzolari et al. (Hrsg.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*, 3255–3259. Paris: European Language Resources Association (ELRA).

Stickel, Gerhard (2007): Die Gründerjahre des IDS. In Heidrun Kämper & Ludwig M. Eichinger (Hrsg.), *Sprach-Perspektiven. Germanistische Linguistik und das Institut für Deutsche Sprache*, 23–41. Tübingen: Narr.

Stickel, Gerhard (2014): Schlechte und bessere Zeiten für das IDS. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 122–143. Mannheim: Institut für Deutsche Sprache.

Stift, Ulf-Michael & Thomas Schmidt (2014): Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch. In *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 360–375. Mannheim: Institut für Deutsche Sprache.

Teubert, Wolfgang & Cyril Belica (2014): Von der linguistischen Datenverarbeitung am IDS zur „Mannheimer Schule der Korpuslinguistik". In: *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 298–319. Mannheim: Institut für Deutsche Sprache.

Trawiński, Beata (2016): Linguistic data in contrastive studies. Addressing the need for a multilingual parallel resource annotated with semantic-functional information, In Domínguez Vázquez, María José & Silvia Kutscher (Hrsg.), *Interacción entre gramática, didáctica y lexicografía. Estudios contrastivos y multicontrastivos*, 85–98. Berlin, Boston: De Gruyter.

Tufiş, Dan, Verginica Barbu Mititelu, Elena Irimia, Ştefan D. Dumitrescu, Tiberiu Boroş, Horia N. Teodorescu, Dan Cristea, Andrei Scutelnicu, Cecilia Bolea, Alex Moruz & Laura Pistol (2015): CoRoLa starts blooming – An update on the Reference Corpus of Contemporary Romanian Language. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3)*, 5–10.

Tufiş, Dan, Verginica Barbu Mititelu, Elena Irimia, Ştefan, D. Dumitrescu, Tiberiu Boroş (2016): The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language. In Nicoletta Calzolari et al. (Hrsg.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz, Slovenia.

Velde, Roger G. van de (1970): Zur Wissenschaftlichkeit der Linguistik. *Studia philosophica Gandensia* vol. 8, 7–33.

Wissenschaftsrat (2011) *Übergreifende Empfehlungen zu Informationsinfrastrukturen*, Berlin. http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf (letzter Zugriff: 20.10. 2017).

Wittenburg, Peter (2009): *Service centers as backbone of a D-SPIN infrastructure – requirements and current state*. Vortrag vom 17. Oktober 2009 im Rahmen des D-SPIN-Beiratstreffens. https://weblicht.sfs.uni-tuebingen.de/Beirat/2-MPI-service-centers.pdf (letzter Zugriff: 20.10. 2017).

Wöllstein, Angelika (2016): Was ein Strukturmodell für die (kontrastive) Sprachbetrachtung im DaF-Bereich leisten kann. In Juan Cuartero Otal, Larreta Zulategui, Juan Pablo & Christoph Ehlers (Hrsg.): *Querschnitt durch die deutsche Sprache aus spanischer Sicht. Perspektiven der Kontrastiven Linguistik*, 211–231. Berlin: Frank & Timme.