Original und erstmals erschienen in:
Börner, Ingo, Wolfgang Straub & Christian Zolles (eds.). 2018.
Germanistik digital: Digital Humanities in der Sprach- und Literaturwissenschaft, 135–152.
Wien: Facultas UTB. ISBN: 9783708917191

Ludwig Maximilian Breuer, Melanie E.-H. Seltmann

Sprachdaten(banken) – Aufbereitung und Visualisierung am Beispiel von SyHD und DiÖ

Einleitung

Es gibt unzählige Definitionen, die versuchen, zu beschreiben, was Digitale Geisteswissenschaften oder Digital Humanities sind.[1] Einer der vielen Ansätze, der für den vorliegenden Beitrag als Grundlage dienen soll, fokussiert die Beantwortung von Forschungsfragen mithilfe digitaler Methoden (vgl. Jannidis/Kohle/Rehbein 2017: XI). Diese digitalen Methoden sollen dabei als Prozesse zur Wissensgenerierung aufgefasst werden, d.h. dass das Wissen nur oder zumindest leichter mithilfe digitaler Methoden generiert werden kann. Unter digitalen Methoden muss dabei natürlich wiederum fachspezifisches digitales Werkzeug verstanden werden und nicht die bloße Anwendung von Office-Programmen. Die Ergebnisinterpretation bleibt geisteswissenschaftlich, sprich: eingebettet in theoretische (Erkenntnis-)Modelle der jeweiligen geisteswissenschaftlichen Disziplin.

Wenn dies nun grob zusammengefasst Digitale Geisteswissenschaften sind, wäre jede Person, die sich oben genannter digitaler Methoden bedient, einE DigitaleR GeisteswissenschaftlerIn; in der aktuellen Praxis ist dies u .E. allerdings (noch) nicht der Fall. Vielmehr nimmt einE DigitaleR GeisteswissenschaftlerIn eine Art Vermittlungs- oder Schnittstellenposition zwischen GeisteswissenschaftlerInnen und TechnikerInnen ein. Die Einbindung solcher VermittlerInnen in geisteswissenschaftliche Arbeitsprozesse bzw. Projekte erscheint insbesondere aus zwei Gründen wichtig: 1. sind die technischen Erfahrungen des geisteswissenschaftlichen Personals sehr unterschiedlich, 2. sind die geisteiswissenschaftlichen Erfahrungen des technischen Personals, z.B. in Bezug auf Arbeitsweise und Terminologie, sehr unterschiedlich und tendenziell eher gering. Genau zwischen diesen beiden Positionen kann nun das DH-Personal eine entscheidende Vermittlungsposition einnehmen.

Der vorliegende Beitrag gibt anhand zweier größerer sprachwissenschaftlicher Projekte Einblicke in die Grundlagen digitaler

geisteswissenschaftlicher Arbeit. Betrachtet werden das Projekt „Syntax hessischer Dialekte“ (SyHD) (Fleischer/Lenz/Weiß 2017) und des Spezialforschungsbereichs (FWF F 60) „Deutsch in Österreich. Variation – Kontakt – Perzeption (DiÖ)“ (DiÖ 2017). Da dies häufig mit der Erarbeitung von Sprachdatenbanken einhergeht, werden deren technischen Hintergründe und Modellierung vorgestellt. Daraufhin werden die Transkriptionen und Annotation als Prozesse der Datenübertragung in die Datenbank vorgestellt sowie abschließend die visuelle Datenverarbeitung und -aufbereitung angeschnitten.

SyHD und DiÖ – Konzeption und Daten

Im Folgenden wird betrachtet, wie Sprachdaten aufbereitet und modelliert werden können, was Sprachdatenbanken sind und inwiefern bereits die Datenbankmodellierung als geisteswissenschaftlicher Prozess aufgefasst werden kann. Die wichtigsten Aspekte der gewählten Beispielprojekte werden zunächst kurz vorgestellt.

Syntax hessischer Dialekte (SyHD)

Das Projekt „Syntax hessischer Dialekte“ (SyHD) wurde von 2010 bis 2016 an den Universitäten Marburg, Frankfurt und Wien durchgeführt, wobei die computerlinguistische bzw. digitale Arbeitsgruppe in Wien die Forschungsplattform SyHD-info[2] entwickelt und implementiert hat. SyHD untersucht syntaktische Variation in der Horizontalen, also die diatopische, areallinguistische Verteilung bestimmter dialektaler Varianten syntaktischer Variablen: Es geht also z.B. um die Frage, in welchen verschiedenen hessischen Dialekten (bzw. Orten) einerseits das Präteritum oder andererseits das Perfekt zum Ausdruck der Vergangenheit verwendet wird und in welchen linguistischen Kontexten die jeweilige Form verwendet wird (vgl. Fischer 2017) (siehe Abb. 1, o:768026).[3]

Die verwendeten Methoden sind zum einen indirekte Erhebungen durch Fragebögen, zum anderen eine direkte Erhebung (vgl. Kuhmichel/Herwig 2017). Die Fragebögen wurden an 163 Orten in Hessen sowie an 12 Orten außerhalb Hessens erhoben (vgl. Fleischer/Kasper/Lenz 2012: 5). Die Interviews wurden an 141 Orten durchgeführt. Die im Projekt gesammelten Daten sowie die Ergebnisse werden zudem einem möglichst breiten Publikum zur Verfügung gestellt und sollen über die Fragestellungen des Projekt hinaus genutzt werden können, wodurch der Open-Data-Ansatz (vgl. OECD 2015: 7) und zumindest zum Teil der Open-Science-Ansatz (vgl. European

Commission 2016: 33) umgesetzt wird. Abb. 2 (o:768115) zeigt den Erhebungsraum und die in ihn fallenden auf phonologischen Merkmalen beruhenden Dialekträume (vgl. Wiesinger 1983: 846–855; Fleischer/Lenz/Weiß 2015: 261f.).

Die gesammelten Daten ergeben sich aus insgesamt 3572 Fragebögen. Die erhobenen Aufgaben stellen größtenteils geschlossene Aufgabenstellungen, wie z.B. Bewertungsaufgaben, dar.[4] Abb. 3 (o:768243) zeigt eine solche Bewertungsaufgabe. Für die Konzeption eines Datenbanksystems ist es wichtig zu wissen, welche Art von Daten gesammelt werden (s.u.). Einerseits aus Gründen der Datenbankmodellierung (für das Back-End), andererseits für die benutzerfreundliche Erstellung von Eingabemasken (für das Front-End). Betrachtet man beispielsweise Abb. 3 (Bewertungsaufgabe) im Vergleich zu Abb. 4 (o:768332, Bildbeschreibungsaufgabe), wird schnell ersichtlich, dass es sich nicht nur um linguistisch unterschiedliche Datentypen mit unterschiedlichen Anforderungen an die Datenbank handelt, insbesondere ist dies der Unterschied zwischen der Auswahl vorhandener Antwortmöglichkeiten und der Eingabe völlig freier Antworten.

Platzbedingt dient dies nur als kleines Beispiel dafür, inwiefern die empirische Konzeption direkten Einfluss hat auf die (DH-)technische Implementierung; und gleichzeitig folgendes Plädoyer unterstreichen: Für eine effiziente, benutzerfreundliche sowie technisch konsistente Entwicklung einer Datenbank für (insbesondere auch) geisteswissenschaftliche Projekte ist das Einbinden von Überlegungen zur Datenmodellierung zu einem frühen Projektzeitpunkt empfehlenswert. Zurück zu SyHD: Die entwickelte Forschungsplattform ermöglicht folgende für das Projekt essenzielle Aspekte: mehrfacher und gleichzeitiger (Bearbeitungs-)Zugriff auf die Projektdaten, leichte und intuitive Verwendbarkeit auch für nicht technisch geschultes Personal, Datenbearbeitung und -analyse v.a. in Bezug auf areallinguistische und syntaktische Fragestellung bei gleichzeitiger Offenheit gegenüber anderer Perspektiven.[5]

Deutsch in Österreich (DiÖ). Variation – Kontakt – Perzeption

Der FWF Spezialforschungsbereich „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (F 60), oder kurz: DiÖ, ist wesentlich komplexer aufgebaut als SyHD. Das Projekt besteht aus fünf verschiedenen Taskclustern mit insgesamt neun Teilprojekten, die an den Universitäten Wien, Salzburg und Graz sowie an der Österreichischen Akademie der Wissenschaften angesiedelt sind. Folglich

potenziert sich nicht nur die Menge der MitarbeiterInnen sowie die Anzahl der Standorte, sondern auch die potenziellen Fragestellungen, die in den verschiedenen Teilprojekten untersucht werden. Der Untersuchungsgegenstand von DiÖ ist die Vielfalt, der Wandel und die Wahrnehmung der deutschen Sprachen und ihrer Varietäten in Österreich. Der SFB umfasst dabei die drei thematischen Säulen „Sprachvariation und -wandel in Stadt und Land" (Taskcluster B), „Sprachkontakt" (Taskcluster C) sowie „Spracheinstellung und -perzeption" (Taskcluster D). Die Taskcluster A und E rahmen es ein und umfassen die organisatorische Planung und Koordination (A) sowie die technische Umsetzung (E). Aus Platzgründen wird hier auf eine detaillierte Beschreibung der Cluster verzichtet, diese ist auf der Projekthomepage einzusehen.[6]

Für den vorliegenden Beitrag sind v.a. die verschiedenen Erhebungsmethoden bzw. Datentypen, die SFB-übergreifend gesammelt werden, wichtig: Zur horizontalen Variation (Dialekte) und zur soziosituativen vertikalen Variation (Standard, Dialekt und alles dazwischen) sowie zum Sprachwandel im ruralen wie urbane Raum (Cluster B) sollen insgesamt 3200 Teilnehmende befragt werden. Die Komplexität des Untersuchungsgegenstandes wie die Größe des Untersuchungsraum bedingt ein vielfältiges Erhebungsmethoden-Setting, wie aus Abb. 5 (o:768559) ersichtlich: Zum einen werden freie Gesprächsdaten anhand von Interviews und Freundesgesprächen erhoben, zum anderen kontrollierte Sprachdaten wie Vorlese- und Übersetzungsaufgaben, Sprachproduktionsexperimente und Fragebögen. Wie auch bei SyHD haben geschlossene Aufgabenstellungen den Vorteil, dass einerseits gezielt nach Phänomenen gesucht werden kann, die in freien Gesprächsdaten vielleicht seltener auftreten, bzw. diese in gezielten (sprachlichen) Kontexten abgefragt werden können (siehe auch Lenz et al. [i.Vorb.]). Darüber hinaus sind insbesondere Fragebögen eine effiziente Methode, eine hohe Quantität an Daten zu erhalten. Die freien Gesprächsdaten dagegen sind die – im linguistischen Sinne – „natürlicheren" Sprachdaten. Hinzu kommen Spracheinstellungs- und Sprachperzeptionsdaten (Cluster D), die innerhalb des oben beschriebenen Methodensets sowie mithilfe von Hörerurteilstests (Bewertung von Sprachproben) und Gruppendiskussionen erhoben werden. Cluster C ergänzt die Daten dagegen um bereits vorhandene schriftliche Quellen wie Lehrwerke, Medien, Gesetzestexte und Zensusdaten, um Fragen des Sprachkontakts diachron wie synchron (auch im Vergleich zu den Daten aus Cluster B) zu untersuchen.

Taskcluster E beschäftigt sich mit den Digital-Humanities-Aspekten von DiÖ und liefert einerseits den GeisteswissenschaftlerInnen der anderen Teilprojekte die nötigen Tools (Programme, technische Infrastruktur, …), ist für die (strukturierte, technische) Datenpflege zuständig und entwickelt andererseits die für die Öffentlichkeit zugängliche Plattform, über die alle Interessierten auf die Daten und Ergebnisse des Projekts zugreifen können. Alle Tools, die im Zuge des Projekt entwickelt werden, stehen auf *GitHub*[7] zur Verfügung und sind somit der wissenschaftlichen Community quelloffen zugänglich.

Datenbanken

Relationale Datenbanken

In beiden oben beschriebenen Projekten werden (relationale) Datenbanken eingesetzt, um die Daten strukturiert zu speichern und zugänglich zu machen. Wahrscheinlich kennen die meisten Menschen Anwendungen oder Tools im weitesten Sinne, die auf eine Datenbank zurückgreifen: ob die Kontaktliste im Smartphone, Lexika, Bibliotheken, Zettelkästen oder auch U-Bahn-Pläne – eigentlich kann dies alles als Datenbank verstanden werden.

Unter einer weiten Definition von Datenbank, in die obige Beispiele fallen können, versteht man zunächst eine strukturierte Datensicherung: Sie repräsentiert eine „Miniwelt" (vgl. Elmasri/Navathe 2011: 4), einen Ausschnitt aus der Realität (vgl. Elmasri/Navathe 2011: 27), in dem Informationen gesammelt und strukturiert werden. Diese stehen folglich in einem logischen, selbst definierten Zusammenhang, der z.B. thematisch (z.B. Zettelkästen) oder alphabetisch (z.B. Lexikon) sein kann (vgl. Elmasri/Navathe 2011: 4f.). Sie hat den Nutzen, diese Daten unter Beachtung bestimmter Anforderungen und Ziele zu systematisieren und weiterverwendbar darstellen zu können, obgleich ihre Erstellung mit einem Informationsverlust einhergeht (vgl. Dimitriadis 2009: 14f.), der – technisch oder forschungspragmatisch – begründbar ist.

Eine digitale Datenbank (Datenbank im engeren Sinne), insbesondere im Sinne einer relationalen Datenbank, baut meist auf einem Datenbankmanagementsystem (DBMS) auf, das eine generalisierte Softwarekomponente darstellt. Das DBMS wird zur Implementierung und Pflege der (digitalen) Datenbank verwendet (vgl. Elmasri/Navathe 2005: 5). Der vorliegende Beitrag konzentriert sich auf zwei Datenbanktypen: relationale und hierarchische Datenbanken

im XML-Format.[8] „In einer relationalen Datenbank stehen Datenfelder und Tabellen miteinander in Beziehung und können mithilfe von Anweisungen ausgewertet werden. Das Modell der relationalen Datenbank bietet flexible Datenstrukturen und eine hohe Verarbeitungsgeschwindigkeit" (Klinke 2017: 111). XML (eXtensible Markup Language) ist „ein Verfahren, um Texte auszuzeichnen und Informationen zu kodieren. Es ist entwickelt worden, um die Strukturen eines Dokuments kenntlich und damit für den Computer verarbeitbar zu machen, indem Kodierungen (‚Auszeichnungen') in einen laufenden Text eingefügt werden." (Vogeler/Sahle 2017: 128).

Grundsätzlich werden zwei Zugriffsebenen bzw. Ansichten einer Datenbank unterschieden. Das Back-End besteht aus Tabellen und Software, die meist EntwicklerInnen mit vollen Zugriffsrechten verwenden. Eingaben werden besser im Front-End vorgenommen, das für AnwenderInnen des Datenbankmanagementsystems (meist mit Zugriffsbegrenzungen) zugänglich und für die Eingaben von Daten optimiert ist. Hier erscheinen z.B. Auswahlfelder, die die Dateneingabe erleichtern.

Semantisches Modell

Bevor mit einer Datenbank gearbeitet werden kann, muss sie zunächst modelliert werden. In der Konzeptualisierung eines semantischen Datenmodells werden für den intendierten Verwendungszweck relevante Entitäten sowie deren Attribute und Relationen identifiziert. Eine Entität ist eine Gruppe von Objekten aus der Wirklichkeit, die eindeutig bestimmbar ist, die allerdings auch abstrakter Natur sein kann. Attribute sind Eigenschaften der Entitäten und beschreiben sie somit genauer. Dadurch können Einzelerscheinungen der Entitäten individualisiert werden: Sie unterscheiden sich durch verschiedene Werte der Attribute. Relationen sind Beziehungen zwischen zwei Entitäten (vgl. Jannidis 2017b: 102f.), zum Beispiel zwischen „Personen" und „Orten". Auch Relationen können bestimmte Attribute aufweisen, also kann z.B. die Entität „Orte" in Relation zur Entität „Personen" ein Wohnort oder ein Geburtsort sein. Dabei können sowohl die Gestaltung von Entitäten, derer Attribute, aber auch die Relationen je nach Zweck der entsprechenden Datenbank sehr unterschiedlich sein. So sind z.B. für soziolinguistische Projekte an der Entität „Personen" sicherlich Attribute wie Alter und Beruf wichtig, für dialektologische Projekte eine vielfache Relation zur Entität „Orte" („geboren in", „gelebt in" etc.) entscheidend.

In der technischen Umsetzung als relationale Datenbank sind

dabei Entitäten als Tabellen, Attribute als Spalten und Relationen als Tabellen (n:m, s.u.) oder Spalten (1:1, 1:n, s.u.) abgebildet. Abb. 6 (o:768560) zeigt dies an einem konstruierten, relationalen Beispiel: Jede Entität bekommt einen eindeutigen Identifikator (ID) zugewiesen, um eindeutig auf einen Datensatz (d. h. die Einzelerscheinung) verweisen zu können. Diese ID wird in anderen Tabellen aufgenommen. Die Beziehungen zwischen den Tabellen kann verschiedene Grade haben. Die einfachste Form ist die 1:1-Beziehung, bei der genau eine Einzelerscheinung einer Entität genau einer Einzelerscheinung einer anderen Entität zugeordnet ist. Ein klassisches Beispiel hierfür wäre die Sozialversicherungsnummer: Eine Person sollte (zumindest innerhalb eines Staats) genau eine haben, diese wiederum sollte keiner weiteren Person zugeschrieben sein.[9]

Die nächstkomplexere Form ist die der 1:n-Beziehung, bei der eine Einzelerscheinung einer Entität beliebig vielen Einzelerscheinungen einer anderen Entitäten zugewiesen sein kann. Jede Person hat (zum heutigen Stand der Technik) z.B. genau eine biologische Mutter, eine biologische Mutter kann jedoch mehrere Kinder haben. Die komplexeste Form ist die n:m-Beziehung, bei der die Anzahl der Verknüpfung von Einzelerscheinungen zweier Entitäten auf beiden Seiten beliebig groß sein kann. So kann eine Person in mehreren Orten wohnen/gewohnt haben und in jedem Ort können mehrere Personen wohnen/gewohnt haben (genauer s. Klinke 2017: 114f.).

XML vs. relationale Datenbank

Die bisherigen Ausführungen beziehen sich vorwiegend auf relationale Datenbanken, wenngleich sich die semantische Datenmodellierung auch für hierarchische Datenbanken eignet, allerdings müssen die Attribute und Beziehungen hier eben verschachtelt aufgebaut werden. In der technische Realisierung unterscheiden sich die beiden Modelle sehr grundsätzlich voneinander: XML[10]-Dokumente bzw. -Datenbanken kann man sich als strukturierte Klammerstrukturen oder Baumstrukturen vorstellen (vgl. Vogeler/Sahle 2017: 128). Attribute von Entitäten sowie Relationen ergeben sich hier durch die Einbettung in ihre Bezugsobjekte. Die hierarchische Struktur ist eine Abhängigkeitsstruktur, quasi eine Dependenzsyntax. Hier wird mit sogenannten Parent- und Child-Elementen gearbeitet (vgl. Vogeler/Sahle 2017: 130f.). Damit erscheint das hierarchische XML erst einmal sehr unflexibel, bei der Abbildung statischer Objekte bzw. (physischer) Dokumente und ihres Aufbaus, z.B. bei Fragebögen, Büchern, Manuskripten etc., ist es jedoch gut geeignet. Die Struktur

des Fragebogens kann fast eins zu eins im XML-Dokument technisch widergespiegelt werden (vgl. Klinke 2017: 127).

In der Gestaltung oder Beschreibung der Entitäten, Attribute etc. (z.B. durch die gebotenen Klammerangaben) ist man grundsätzlich sehr frei, da es sich bei XML um ein freies Format (mit freien Vokabular) handelt (eigentlich um eine Metasprache) (vgl. Vogeler/Sahle 2017: 135). Da man allerdings über verschiedene Projekte vergleichbar arbeiten möchte, greift man auf Standards oder Best Practices zurück. Ein solcher Standard, der insbesondere in den Digital Humanities der sprach- bzw. textbasierenden Wissenschaften Usus ist, wäre z.B. das TEI-Format[11], das wie eine Art Lexikon für XML ist (vgl. Vogeler/Sahle 2017: 133). Im eigentlichen, reinen XML sind keine Querverbindungen möglich. Es können jedoch Stand-off-Annotationen (Zuweisungen außerhalb der hierarchischen Struktur) genutzt werden, um solche Verbindungen doch herzustellen. Ein großer Vorteil ist die Flexibilität der Datenkonsistenz, da nicht alle Erscheinungen einer Entität dieselben Attribute oder Relationen aufweisen müssen – der Nachteil, dass daraus uneinheitliche Datensätze entstehen können.

In einer relationalen Datenbank muss für ein neues Attribut in einer Tabelle das Attribut auf alle Objekte in der Tabelle angewendet werden. Eine solche Datenverarbeitung bietet sich für komplex verknüpfte Objekte an.

Welches der beiden Modelle in einem Projekt Verwendung findet, ist abhängig von der Projektstruktur und von den technischen Möglichkeiten innerhalb des Projekts (vgl. Dimitriadis/Musgrave 2009: 21). Auch Kombinationen sind technisch möglich.

Computational Thinking

Zurück von der technischen Beschreibung zur Geisteswissenschaft: Die Datenmodellierung stellt u.E. einen geisteswissenschaftlichen Prozess dar. Sie gleicht – zumindest im linguistischen Sinne – einem strukturalistischen Segmentations- und Klassifikationsverfahren. Es handelt sich um die Interpretation realer Objekte und deren Relationen, bestimmte Informationen werden als relevant identifiziert und verdichtet. Die Modellierung ist ein iterativer Denkprozess und kann auch deswegen dem Computational Thinking (vgl. Jannidis 2017a: 89–92) nahegestellt werden. Ein vorhandenes Problem wird abstrahiert und möglichst effektiv zerlegt, sodass die Teilprobleme zu bewältigen sind. Daraufhin werden passende Lösungen identifiziert und die Probleme generalisiert. In der Wissenschaft besteht

die Problemabstraktion in den Fragestellungen zur Welt, den Hypothesen. Diese werden operationalisiert, damit sie durch empirische Überprüfung verifiziert oder falsifiziert werden können. Ein iterativer Prozess führt wieder zu den anfänglichen Hypothesen zurück und modifiziert diese entsprechend der vorherigen Ergebnisse bzw. identifiziert und formuliert neue Teilprobleme. Ganz ähnlich lässt sich der Prozess der Modellierung von Datenbanken – oder generell der Programmierung – beschreiben.

Transkription und Annotation

Transkription

Um gesprochene Sprachdaten überhaupt technisch sammeln und in Form einer Datenbank aufbereiten und später bearbeiten und analysieren zu können, steht im ersten Schritt eine Transkription der Aufnahmen. Darunter versteht man die Verschriftlichung gesprochener Sprachdaten, genauer:

> Der Terminus „Transkription" bezieht sich auf die Wiedergabe eines gesprochenen Diskurses in einem situativen Kontext mit Hilfe alphabetischer Schriftsätze und anderer, auf kommunikatives Verhalten verweisender Symbole. Die schriftliche Wiedergabe soll nicht nur „ungefähr" oder annährungsweise authentisch, sondern eine reale Kommunikationssituation möglichst genau abbildende Verschriftlichung sein. (Dittmar 2004: 50f.)

Eine linguistische Transkription soll möglichst viele Informationen beinhalten trotz inhärenten Informationsverlusts (vgl. Sauer/Lüdeling 2016: 421). Je nach Forschungsinteresse wird die Transkriptionsart und somit unterschiedliche Grade des Informationsverlusts bestimmt (s.u.).

Einer der Vorteile, ein gesprochensprachliches Korpus zu transkribieren, liegt darin, dass es visuell schneller und leichter zu rezipieren ist als auditiv. Visuell sind Untersuchende nicht mehr an einen linearen Zugriff gebunden, können quasi querlesen. Gesprochene Sprache ist zeitlich bedingt und auf verschiedene Teile von Sprache kann auditiv schwer zeitgleich zugegriffen werden.

Ein Zugang ohne viele Vorkenntnisse besonderer Verschriftlichungsformen erfolgt durch eine orthografische Umschrift oder auch normalisierte Transkription. Diese Art der Transkription kann angewendet werden, wenn v.a. Gesprächsinhalte untersucht werden. Bei der Verschriftlichung von Nichtstandarddeutsch werden phonetische Merkmale an orthografische Schreibweisen angepasst. Der Vorteil ist, dass die Transkription fast jede Person lesen kann (die

entsprechendes Schrift- und Sprachsystem erworben/erlernt hat). Auch für variationslinguistische, insbesondere (morpho-)syntaktische Fragestellungen kann dieses Transkriptionssystem hilfreich sein, solange nur auf Wortebene bzw. nur lautlich normalisiert wird, d.h. z.B. keine Anpassungen bei der Wortstellung vorgenommen werden.

Eine nah am gesprochenen Wort orientierte Transkription wäre die lautgetreue Umschrift (so auch bei SyHD bzw. teilweise bei DiÖ), die auch literarische oder diplomatische Transkription genannt wird. Dabei werden vorwiegend mit orthografischen Graphemen lautliche Besonderheiten abgebildet: Ein „heast" wird dann eben so geschrieben, und nicht als z.B. „hörst" transkribiert. Problematisch ist dabei jedoch die Vereinheitlichung. Um diesem Problem – insbesondere bei großen Projekte – entgegenzuwirken, sind exakte Transkriptionskonventionen unentbehrlich. Komplette Transkriptionssysteme ermöglichen das Erfassen über die verbale Ebene hinaus, d. h. auch der Gesprächssituation. Für z.B. gesprächs- oder diskursanalytische Fragestellungen bieten sich die Transkriptionssysteme GAT(2) (Selting et al. 2009) oder HIAT(2) (Ehlich/Rehbein 1979) an, die u.a. Sprecherwechsel oder Pausen kodieren.

Für Fragestellungen, bei denen die genaue lautliche Realisierung der Sprache relevant ist, muss phonetisch (z.B. nach IPA) transkribiert werden; auch hierbei ist die Transkription von der Transkribierendeninterpretation abhängig (vgl. Sauer/Lüdeling 2016: 421).

Wie bereits mehrfach angedeutet, handelt es sich bei der Transkription um eine Interpretation, bei der bestimmte sprachliche und auditive Signale einem technischen und visuellen Zeichen zugeordnet werden. Ferner werden häufig einzelne Segmente (Äußerungseinheiten, Sätze o.Ä.) bei der Transkription einer Zeiteinheit des Audiomaterials zugeordnet, dies nennt man „Time-Alignment". So verstanden werden die Sprachdaten sowohl durch die Transkription als auch durch das „Time-Alignment" annotiert (s.u.). Durch diese Annotation stellt die Transkription auch einen geisteswissenschaftlichen Zugang zu auditiven Sprachdaten dar.

Die SyHD-Daten aus der direkten Erhebung liegen literarisch und orthografisch transkribiert vor (vgl. Abb. 7, o:768586) – allerdings nur in Form der für die Untersuchung relevanten Einzelsätze, d.h. ohne Abbildung des stattgefundenen Erhebungsgesprächs, was dem oben beschriebenen geschlossenen Methoden entspricht.

In DiÖ dagegen sind die Transkriptionsverfahren wesentlich komplexer, entsprechend der oben formulierten Ansprüche an das

Datenmaterial, aber auch wegen der Natur der gesammelten Daten (freie Gesprächsdaten und kontrollierte Sprachdaten). Als projektübergreifende Projektionsebene der Transkripte wird eine orthografische Umschrift (auf Wortebene) herangezogen; dies ist einerseits nötig, damit die verschiedenen Daten vergleichbar sind, andererseits für die technische Verarbeitung: Durch die vereinheitlichte Verschriftlichung werden die Daten leicht durchsuchbar. Zudem sind automatische Tagger auf diese Ebene trainiert. Ansonsten werden verschiedene Transkriptionsverfahren abhängig vom Untersuchungsgegenstand und Datenmaterial gewählt: Es kommen sowohl phonetische als auch literarische Umschriften zum Einsatz (vgl. Abb. 7), sowohl in Form von Einzelsätzen als auch von Abbildungen ganzer Gespräche (vgl. Abb. 8, o:768657).

Annotation

Transkriptionen wurden oben als eine Art von Annotation beschrieben, doch wird mit Annotationen meist eine abstraktere Kategorisierung von Daten gemeint. Annotationen sind allgemein Markierungen, Kategorisierungen und Interpretation von (Sprach-)Daten. Sie dienen zur „Abstrahierung von Einzelvorkommen und deren Ordnung und Zusammenfassung in übergreifenden Klassen […] und [sind] damit sowohl Teil des wissenschaftlichen Forschungsprozesses als auch dessen Ergebnis" (Rapp 2017: 254). Mithilfe der Annotationen kann nach Mustern und Phänomenen in den Daten gesucht werden. Dabei können nach Rapp (2017) Annotationen ganz unterschiedliche Funktionen erfüllen:

– Organisation und Visualisierung von Informationen;
– Ergebnis inhaltlicher Analysen;
– Ausgangspunkt von Auswertungen → Zwischenschritt der Analyse;
– Teil des iterativen Prozesses der Wissensgenerierung und -weitergabe → Teil des *research life cycle* (der den Forschungsprozess abbildet: vgl. Van den Eynden et al. 2009).

Ein Element, das für eine Annotation benutzt wird, bezeichnet man als „Tag" (vgl. Rapp 2017: 255). Dies könnte zum Beispiel eine Variante im linguistischen Sinne sein (vgl. Lüdeling 2017: 132). Eine Auswahl von Tags, die für einen bestimmten Annotationsschritt verwendet werden oder die zu einem vergleichbaren Analysezweck herangezogen werden, wird als „Tagset" bezeichnet. Eines der bekanntesten ist das Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999) für das Part-of-Speech-Tagging (PoS-Tagging).

Unter Tagset kann allerdings auch eine individuelle Sammlung von Tags z.B. für die Untersuchung einer bestimmten Variable verstanden werden. Die Benennung der Tags an sich ist arbiträr, aber es wird versucht, Standards zu entwickeln, um getaggte Korpora miteinander vergleichen zu können

Einer der Zwecke der Annotation ist, nach den markierten Phänomenen suchen zu können und sie somit zu quantifizieren – dies auch in Kombination verschiedener Annotationsebenen. Z.B. können dadurch Fragestellungen wie „Wie oft wurde ein „täte"-Konjunktiv in einem Nebensatz von einer bestimmten Personengruppe realisiert" beantwortet werden. Annotationen können manuell, semiautomatisch oder voll automatisch geschehen, wofür systemebenenabhängig automatische Tagger[12] vorliegen. Im Bereich des Dialekts gibt es hier jedoch wenige zuverlässige automatische Tagger (vgl. Lüdeling 2017: 137).

Annotationen sind u.E. in der heutigen Umsetzung, d.h. gespeichert in Datenbanken, ein Kernbereich digitaler Geisteswissenschaft: Bei der Annotation handelt es sich einerseits um eine digitale Kategorisierung und damit um die Befähigung des Computers, geisteswissenschaftliche Fragstellungen überhaupt beantworten zu können. Andererseits stellen Annotation eine technische, digitale Abbildung der geisteswissenschaftlichen Analyse dar. Die Modellierung der Tagsets basiert auf theoretischen Konzepten aus den entsprechenden (Geistes-)Wissenschaften. Sie folgt einem bewährten iterativen Verfahren, wie Rapp (2017) ausführlich beschreibt. Aus diesem iterativen Prozess, der in großen Projekten stets auch von einer Person betreut werden sollte, ergibt sich ein für die Analyse erforderliches Tagset.

In SyHD wurden stets phänomenbezogene Analysen durchgeführt, was sich auch aus den eher geschlossenen Methoden ergibt. Die Daten sind v.a. mit syntaktischen Tags annotiert und es liegen nur teilweise morphologische oder phonologische Informationen als Teil von phänomenbezogenen Tags vor. Die Annotation kann als eindimensional-relational betrachtet werden, da eine Antwort (außer in wenigen Ausnahmen) genau einem Tag zugeordnet wird. Ein einzelner Tag kann dabei sehr viele Informationen erhalten oder schlicht entsprechend der gesuchten syntaktischen Variante, d.h. nach objektsprachlich-orientierten Kategorien, gestaltet sein.

Im SFB DiÖ wird phänomenbezogen auf mehreren linguistischen Systemebenen (Phonetik/Phonologie, Morphologie/Lexik, Syntax, …) getaggt. Hinzu kommen qualitativ-inhaltliche sowie ggf. gesprächsanalytische Tags. Je nach getaggtem Phänomen sind

z.B. auch Einflussfaktoren auf die Wahl bestimmter Varianten sowie sprachliche Kontexte für die Analyse relevant und werden annotiert. Daher wurde ein eigenes Tagsystem entwickelt, welches diesem mehrdimensionalen Annotationsanspruch genügt. Zudem soll ein möglichst einheitliches Annotationssystem aufgebaut werden, das, wenn möglich, auf existierende Standards zurückgreift. Das Besondere am DiÖ-Tagsystem ist, dass es in seiner Abbildung hierarchisch ist, auch wenn es technisch linear auf mehreren Tagebenen[13] gespeichert wird, welche von der Untersuchungsebene abhängig sind und einen Mehrbenutzerzugriff vereinfachen.

Es gibt also mehrere, hierarchisch voneinander abhängige Generationen innerhalb eines Tagsets, wie in Abb. 9.1 (o:768658) an einem Beispiel gezeigt wird: Auf 0. Generation wird das zu untersuchende Phänomen spezifiziert (z.B. Analysen zum Dativpassiv). In der nächsten Generation werden Kategorien, die für die Analyse von den verschiedenen Forschenden gefordert werden, markiert (z.B. Genus Verbi). Ab der zweiten Generation werden nun die Ausprägungen (Features) – unterhalb ihrer entsprechenden Kategorie – markiert, welche im Sprachmaterial feststellbar sind (z.B. Dativpassiv). Dieses Feature kann in der 3. Generation weiter konkretisiert werden (z.B. „kriegen"-Passiv).[14] Die Front-End-Umsetzung anhand eines Beispiels zur Possession wird in Abb. 9.2 (o:768659) ersichtlich.

Wie erwähnt ergibt sich dadurch eine hierarchische Darstellung (im Front-End), im Back-End dagegen werden die Kategorien linear gespeichert (siehe Abb. 9.3, o:768660). Der Vorteil dieses Annotationsverfahrens ist, dass spätere (und erwartbare) Ergänzungen leicht möglich sind, da durch die hierarchische Struktur Kategorien und entsprechende Features einfach ergänzt werden können – was durch die lineare Speicherung wiederum datentechnisch leicht ermöglicht wird. Die hierarchische Abbildung vereinfacht zudem den Tagging-Vorgang selbst, da jeweils nur die nächsten möglichen Tags angezeigt werden und es daher nicht zu „falschen" Reihenfolgen der Tags kommen kann oder Tags, die benötigt werden, vergessen werden. Die Anlage des hierarchischen Tags erfordert außerdem, dass sie wohlstrukturiert aufgebaut und allein durch die Struktur schon gut beschrieben sind. Für DiÖ wird außerdem eine Explikation der Tags gefordert. Nicht nur für projektinterne Zwecke der Zusammenarbeit verschiedener BenutzerInnen, sondern auch für das projektexterne Verständnis der Tags ist ein solches Verfahren wichtig.[15]

Verarbeitung/Visualisierung

Zu guter Letzt soll auf einen weiteren wichtigen Aspekt der Erstellung von Forschungsplattformen resp. Sprachdatenbanken eingegangen werden, der häufig vernachlässigt wird: die Visualisierung. Auch im vorliegenden Beitrag kann nur kurz auf diesen Aspekt eingegangen werden, was nicht weniger ein Plädoyer für eine eingehende Beschäftigung mit dem Thema darstellen soll. Wenn in der Sprachwissenschaft Visualisierung diskutiert wird, bezieht man sich v.a. auf Diagramme, in der Areallinguistik zusätzlich auf Karten. U.E. umfasst die Visualisierung aber nicht nur den Content, sondern eben auch die Datenverarbeitung. Es zählt hierzu auch die Darstellung von Datenbanken selbst, also v.a. die Umsetzung der Eingabemasken (Front-End). Gerade in den digitalen Geisteswissenschaften ist dies ein wichtiger Aspekt, der für die Arbeit mit Personen mit unterschiedlicher technischer Erfahrung entscheidend ist. Mit einer gut designten Arbeitsumgebung werden auch die BenutzerInnen effektiver in ihrer Arbeit und können sich auf Wesentliches ihrer Forschungstätigkeit konzentrieren. Wichtig ist dabei, dass Form und Funktion verbunden werden. Intuitiv verwendbare und ähnlich der Vorgabe visualisierte Eingabemasken senken die Fehlerquote, intuitiv verständliche Karten erleichtern bzw. erhöhen den Wissenstransfer (vgl. Rehbein 2017: 336).

Bei SyHD sind die Eingabemasken an das Ausgangsmaterial angepasst, was für Fragebögen natürlich leicht ersichtlich ist, da sie eine konkrete physisch-visuelle Form aufweisen. Abb. 10 (o:768661) zeigt, wie die Gestaltung des Fragebogens auf die Gestaltung der Eingabemaske übertragen wurde, damit für Eingebende a) die Wiedererkennung der gerade einzugebenden Frage gewährleistet ist, wie auch b) die ungefähre Lokalisierung der konkreten Antwort erleichtert wird. Solche Gestaltungsprinzipien senken die Fehlerquote, da eine visuelle Rückkopplung von der Vorlage des ausgefüllten Fragebogens zur Eingabe in der Datenbank ermöglicht wird. Schwieriger ist dies für Sprachaufnahmen, die eben keine visuelle Form vorweisen. Hierbei wurde versucht die mögliche Antwortstruktur (Einzelsätze auf Einzelfragen, spontane Antworten und Nachfragen) abzubilden.

Die (ausgewerteten) Daten werden in einer umfassenden Online-Publikation zur Verfügung gestellt. Die Tools von SyHD.info wurden möglichst intuitiv gestaltet, so können mit wenigen Klicks User mit verschiedensten Wissensstands interpretierte Ergebnisse zu den Daten erhalten[16] – einerseits in Form von Phänomenbeschreibungen, andererseits in Form von Diagrammen oder interaktiven

Sprachkarten. Ein niederschwelliger Wissenstransfer – insbesondere bei komplexen Daten – ist natürlich nicht nur für Nicht-WissenschaftlerInnen wünschenswert. Das Online-Tool SyHD-atlas geht über die reine Online-Publikation hinaus: Atlaskarten können in einer Art Warenkorb Fragen übergreifend gesammelt und visuell nebeneinandergestellt werden – dadurch wird das genutzte Medium wirklich genutzt, anstatt ein Print-Medium online zu stellen (wie das bei PDFs häufig der Fall ist). Bei SyHD-atlas können die Karten interaktiv komplett selbst gestaltet werden.

Um den Rahmen des vorliegenden Beitrags nicht noch weiter auszureizen, bleibt schließlich SyHD als Visualisierungsbeispiel sowie das Plädoyer zu einer eingehenden Beschäftigung mit der Visualisierung nicht nur auf Ebene Datenveröffentlichung, sondern auch der Datenverarbeitung. Visualisierung ist ein wichtiger Teil der DH-Entwicklung, aber in der Content-Visualisierung schließlich auch des geisteswissenschaftlichen Erkenntnisgewinns: Die Visualisierung der geografischen Verteilung bestimmter linguistischer Phoneme ist auf einer Karte besser analysierbar als nur anhand von Datentabellen. Die Visualisierung ist folglich eine Möglichkeit für die geisteswissenschaftliche Analyse, in digitaler Form eine neue und exhaustiv verwendbare Analysemöglichkeit.

Resümee

Zusammenfassend wird hier festgehalten, dass die Arbeit mit Datenbanken in der Linguistik – aber darüber hinaus in allen geisteswissenschaftlichen Disziplinen – nicht nur hilfreich für die Beantwortung von Fragestellungen an sehr vielen Daten ist, sondern eine prinzipielle Arbeitsweise darstellt oder darstellen sollte. Datenmodellierung, Transkription, Annotation und Visualisierung stellen in unserem Verständnis grundsätzliche Aspekte der digitalen Geisteswissenschaften dar.

Anmerkungen

1 Besonders beachtlich hierzu die Webseite whatisdigitalhumanities.com, die bei jedem Aufruf eine neue Definition liefert.
2 Zu finden ist die Plattform unter URL: www.syhd.info.
3 Die Abbildungen dieses Beitrags finden sich unter dem angegebenen Link langfristig im *Phaidra*-Repositorium der Universität Wien gespeichert. Sie können aufgerufen werden unter URL: phaidra.univie.ac.at, mit direkt anschließender o-Nummer (hier URL: phaidra.univie.ac.at/o:768026).
4 Zu den verschiedenen Aufgabetypen der indirekten Erhebungen in SyHD und deren unterschiedliche linguistische Qualität siehe Fleischer/Kasper/Lenz 2012.
5 Die Öffnung für alle erdenklichen Fragestellungen, die in einer Anforderungsanalyse nicht schon vorweg erhoben werden können, ergibt sich am leichtesten durch die Offenlegung der zugrundeliegenden Daten, am einfachsten über den Zugriff auf Daten in nicht-proprietären Formaten wie z.B. CSV-Tabellen.
6 Siehe URL: dioe.at, Abschnitt ‚Projekte'.
7 Erreichbar unter URL: github.com/german-in-austria.
8 XML-Daten werden meist dokumentbasierend erstellt, sind aber auch mithilfe eines DBMS organisierbar.
9 Häufig ist es sinnvoll, eine solche Beziehung schlicht als Attribut festzuhalten. Es kann aber Fälle geben, in denen solche Attribute als eigene Entität modelliert werden: Meistens wenn sie a) als eigene Entität im Sinne eines abgrenzbaren Objekts der Wirklichkeit beschreibbar ist und b) zumindest einige eigene Attribute besitzt. Ein Beispiel hierfür wäre ein Bibliotheksausweis einer definierbaren Bibliothek: JedeR BesitzerIn einer Bibliothekskarte (dieser Bibliothek) hat genau eine davon (mit eigener Identifikationsnummer), eine Bibliothekskarte zählt nur für eine Person. Die Karte wiederum hat weitere Attribute wie Ablaufdatum, Entlehnrechte etc.
10 Davon abgesehen gibt es natürlich auch andere hierarchische Daten(bank)formate wie z.B. JSON, auf die hier aus Platzgründen nicht näher eingegangen werden kann.
11 TEI = Text Encoding Initiative. Weitere Informationen siehe URL: www.tei-c.org/index.xml.
12 Tagger sind Programme, die automatische Annotationen durchführen. Ihre Vorgehensweisen sind entweder regelbasiert oder statistisch, sie unterscheiden sich dabei je nach der Domäne, die sie bearbeiten. Ihr Resultat hängt von der Datenmenge, auf der sie trainiert werden, (Trainingskorpus) ab (vgl. Zinsmeister 2015: 88f.).
13 Zum Tagging auf mehreren Ebenen siehe auch Sauer/Lüdeling 2016: 423.
14 Zu ersten Analysen zum Dativpassiv in den Sprachproduktionsexperimente des SFB DiÖ siehe Lenz et al. (i.Vorb.).
15 Eine gute Beschreibung ist wesentlich für die Auswertung und Interpretation der Annotation. „Wenn [eine] Variable nicht definiert ist, weiß man nicht, zu welcher Variable eine Kategorie gehört." (Lüdeling 2017: 136)
16 Siehe im Folgenden die entsprechenden Abschnitte unter URL: www.syhd.info.

Literatur

Berlin Map Task Corpus (BeMaTaC) (2007): About, URL: u.hu-berlin.de/bematac [22.5.2018].

Chambers, Jack/Trudgill, Peter (1998): Dialectology. Cambridge: Cambridge University Press.

Dimitriadis, Alexis/Musgrave, Simon (2009): Designing linguistic databases: A primer for linguists. In: Everaert, Martin/Musgrave, Simon/Dimitriadis, Alexis (Hg.): The Use of Databases in Cross-Linguistic Studies. Berlin: de Gruyter Mouton, S. 13–75.

Dittmar, Norbert (2004): Transkription. Ein Leitfaden mit Aufgaben für Studenten, Forscher und Laien. Wiesbaden: Springer.

Deutsch in Österreich (DiÖ) (2017): Überblick. In: DiÖ-Online, URL: dioe.at/details/ [19.5.2018].

Ehlich, Konrad/Rehbein, Jochen (1979): Erweiterte halbinterpretative Arbeitstranskriptionen (HIAT2): Intonation. In: Linguistische Berichte 59, S. 51–75.

Elmasri, Ramez/Navathe, Shamkant B. (2011): Fundamentals of Database Systems. Pearson Studium, URL: iips.icci.edu.iq/images/exam/databases-ramaz.pdf [17.5.2018].

European Commission (2016): Open innovation, Open Science, open to the world. A vision for Europe. Brussels: European Commission, Directorate-General for Research and Innovation, DOI: dx.doi.org/10.2777/061652 [17.5.2018].

Fischer, Hanna (2017): Präteritum/Perfekt-Distribution. In: Fleischer, Jürg/Lenz, Alexandra N./Weiß, Helmut: SyHD-atlas. Konzipiert von Ludwig M. Breuer. Unter Mitarbeit von Katrin Kuhmichel, Stephanie Leser-Cronau, Johanna Schwalm und Thomas Strobel. Marburg/Wien/Frankfurt a. M., S. 25–45, DOI: dx.doi.org/10.17192/es2017.0003 [17.5.2018].

Fleischer, Jürg/Kasper, Simon/Lenz, Alexandra N. (2012): Die Erhebung syntaktischer Phänomene durch die indirekte Methode: Ergebnisse und Erfahrungen aus dem Forschungsprojekt „Syntax hessischer Dialekte“ (SyHD). In: Zeitschrift für Dialektologie und Linguistik 79, H. 1, S. 2–42.

Fleischer, Jürg/Alexandra N. Lenz/Helmut Weiß (2015): Das Forschungsprojekt „Syntax hessischer Dialekte (SyHD)“. In: Kehrein, Roland/Alfred Lameli/Stefan Rabanus (Hg.): Areale Variation des Deutschen. Projekte und Perspektiven. Berlin/Boston: de Gruyter, S. 261–287.

Fleischer, Jürg/Lenz, Alexandra N./Weiß, Helmut (2017): SyHD-atlas. Konzipiert von Ludwig M. Breuer. Unter Mitarbeit von Katrin Kuhmichel, Stephanie Leser-Cronau, Johanna Schwalm und Thomas Strobel. Marburg/Wien/Frankfurt. DOI: dx.doi.org/10.17192/es2017.0003 [17.5.2018].

Jannidis, Fotis (2017a): Grundbegriffe des Programmierens. In: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, S. 68–95.

Jannidis, Fotis (2017b): Grundlagen der Datenmodellierung. In Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, S. 99–108.

Klinke, Harald (2017): Datenbanken. In: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, S. 109–127.

Kuhmichel, Katrin/Herwig, Katja (2017): Beispiel Durchführung, URL: www.syhd.info/ueber-das-projekt/beispiel-durchfuehrung/ [Zugriff: 19.5.2018].

Lenz, Alexandra/Breuer, Ludwig M./Fingerhuth, Matthias/Wittibschlager, Anja/Seltmann, Melanie (i.Vorb.): Exploring syntactic variation by means of „Language Production Experiments. Methods from and analyses on German in Austria“. Unveröffentlichtes Manuskript.

Lüdeling, Anke (2017): Variationistische Korpusstudien. In: Konopka, Marek/Wöllstein, Angelika (Hg.): Grammatische Variation. Empirische Zugänge und theoretische Modellierung. IDS Jahrbuch 2016. Berlin: de Gruyter, S. 129–144.

OECD (2015), „Making Open Science a Reality“. In: OECD Science, Technology and Industry Policy Papers 25, DOI: dx.doi.org/10.1787/5jrs2f963zs1-en [21.5.2018].

Rapp, Andrea (2017): Manuelle und automatische Annotation. In: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, S. 253–267.

Rehbein, Malte (2017): Informationsvisualisierung. In: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, S. 328–342.

Sauer, Simon/Lüdeling, Anke (2016): Flexible multi-layer spoken dialogue corpora. In: International Journal of Corpus Linguistics 21 (3), S. 419–438, DOI: dx.doi.org/10.1075/ijcl.21.3.06sau [20.5.2018].

Schiller, Anne/Teufel, Simone/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textkorpora mit STTS (Kleines und großes Tagset). Universität Stuttgart/Universität Tübingen, URL: www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [5.5.2018].

Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10, S. 353–402.

Van den Eynden, Verlee/Corti, Louise/Woollard, Matthew/Bishop, Libby/Horton, Laurence (2009): Managing and Sharing Data: A Best Practice Guide for Researchers, URL: www.data-archive.ac.uk/media/2894/managingsharing.pdf [2.6.2010].

Vogeler, Georg/Sahle, Patrick (2017): XML. In: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, S. 128–146.

Wiesinger, Peter (1983): Die Einteilung der deutschen Dialekte. In: Besch, Werner/Knoop, Ulrich/Putschke, Wolfgang/Wiegand, Herbert Ernst (Hg.): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. 2. Halbband. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 1.2), S. 807–900.

Zinsmeister, Heike (2015): Chancen und Grenzen von automatischer Annotation. In: Themenheft zu Automatisierte Textanalyse für Sozial- und Kulturwissenschaften. Zeitschrift für Germanistische Linguistik 43, H. 1, S. 84–110.